Contribution ID: **6** Type: **Practical Demonstration (20')**

Serverless Scientific Computing with OSCAR

Tuesday, 26 September 2023 13:40 (20 minutes)

OSCAR is an open-source platform for serverless event-driven data-processing applications built on Kubernetes. The event-driven architecture allows for flexible and scalable applications which execute in response to events coming from different sources, such as object storage systems (e.g. MinIO or dCache).

An OSCAR cluster can be dynamically deployed by the Infrastructure Manager (IM) [2] on any major public (e.g. AWS), on-premises (e.g. OpenStack) or federated Cloud (e.g EGI Compute), both using web-based user interfaces (IM Dashboard) or programmatically. OSCAR supports ARM-based computer architectures deployed in low-powered devices such as clusters of Raspberry Pis.

An OSCAR service is created by specifying a Docker image, which can be in an image container registry (e.g. Docker Hub), certain computing requirements (e.g. vCPUs, RAM, GPUs) and a user-provided shell script that will be executed inside a dynamically created container on a horizontally scalable Kubernetes cluster which grows and shrinks depending on the workload. An OSCAR service can also support synchronous invocations to create highly-scalable HTTP endpoints via Knative. It also provides the ability to expose load-balanced services accessed via HTTP requests, a more performant approach when deploying AI models for inference, where the weights need to be pre-loaded in memory to be reused for subsequent inference requests.

OSCAR services can be chained in a Functions Definition Language (FDL) to create data-driven pipelines, even across multiple OSCAR clusters, so that the output of one service is uploaded to the input object storage of another service. By chaining these services, data-processing pipelines along the cloud-to-edge continuum can be created.

OSCAR is integrated with EGI Notebooks and Elyra to support the composition of AI inference pipelines from Jupyter Notebooks. It is also integrated with scientific object storage systems such as dCache to react upon file uploads to a certain folder This functionality, when coupled with Apache Nifi for scalable event-driven ingestion, provides the ability to support data-driven processing in an scalable Kubernetes-based platform.

In AI-SPRINT, OSCAR supports the scalable inference of pre-trained AI modes in use cases related to agriculture 4.0, personalised healthcare and maintenance and inspection. In AI4EOSC, OSCAR is also used to deploy AI models, but extending its support to create visual AI pipelines using both Node-Red and Elyra for the AI4Compose service. In InterTwin, OSCAR performs data-driven ingestion and processing via dCache and Apache Nifi.

In this contribution, we want to showcase some benefits of OSCAR as a serverless platform for scientific computing.

Acknowledgements
Project PDC2021-120844-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. Grant PID2020-113126RB-I00 funded by MCIN/AEI/10.13039/501100011033. This work was supported by the project AI-SPRINT "AI in Secure Privacy-Preserving Computing Continuum"that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant 101016577. Also, by the project AI4EOSC "Artificial Intelligence for the European Open Science Cloud"that has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant 101058593.

References

[1] OSCAR. https://oscar.grycap.net
[2] Infrastructure Manager. https://im.egi.eu

Primary authors: ALARCÓN, Caterina (Universitat Politècnica de València); LANGARITA, Sergio (Universitat Politècnica de València); CABALLER, Miguel (Universitat Politènica de València); CALATRAVA ARROYO, Amanda (Universitat Politècnica de València); Dr MOLTÓ, Germán (Universitat Politècnica de València)

Presenters: ALARCÓN, Caterina (Universitat Politècnica de València); Dr MOLTÓ, Germán (Universitat Politècnica de València)

Session Classification: IBERGRID Contributions with Demonstrations

Track Classification: Development of innovative software and services