Contribution ID: **3** Type: **Demonstration**

Serverless computing across the Cloud continuum for Deep Learning Inference with OSCAR

Thursday, 21 October 2021 13:15 (10 minutes)

OSCAR is an open-source platform to support serverless computing for compute-intensive data-processing applications. OSCAR runs on dynamically provisioned auto-scaled Kubernetes clusters deployed through the Infrastructure Manager (IM), an open-source Infrastructure as Code (IaC) tool. These Kubernetes clusters include support for MinIO, an open-source object storage, which fires events in response to file uploads in order to trigger the file processing as Kubernetes jobs. The clusters can grow and shrink thanks to CLUES, an open-source elasticity manager that deploys additional nodes through the IM and terminates them whenever they are no longer needed.

OSCAR has evolved in the last years to tighten the integration with SCAR, an open-source tool to execute containers in AWS Lambda with automatic delegation into AWS Batch. The use of a Functions Definition Language (FDL) allows to define data-processing serverless workflows that can perform some processing in an on-premises Cloud while delegating the most computationally-intensive part in a public Cloud such as Amazon Web Services (AWS).

OSCAR is integrated with the IM Dashboard to facilitate the deployment of OSCAR clusters across a myriad of Cloud providers including major public clouds, widely used Cloud Management Platforms and federated infrastructure such as the EGI Federated Cloud. Thanks to the integration with EGI Check-In, users can seamlessly access the IM dashboard to self-provision these clusters.

In order to expand the variety of use cases supported, two additional features have been recently implemented. First, synchronous support with scale-to-zero capabilities. This allows synchronous data-processing functions packaged as Docker containers that can be invoked through both a REST API and the CLI. Second, OSCAR can run on minified Kubernetes distributions (such as K3s) in order to execute on low-powered devices such as Raspberry PIs. This is required for use cases that require executions in the edge, for lightweight processing, such as inference of previously trained Deep Learning (DL) models.

Together, the integration of OSCAR / SCAR provides an open-source platform that supports serverless computing across the Cloud continuum, where execution can take place in low-powered devices, in on-premises Clouds and in federated or public Clouds. In this contribution we will demonstrate how a user can seamlessly provision an OSCAR cluster using the IM Dashboard in the EGI Federated Cloud in order to create a serverless workflow for mask detection from a trained DL model in public crowds across different infrastructures and using EGI DataHub as one of the storage back-ends.

Speaker bio:

Sebastián Risco received a BSc degree in Computer Engineering from the Universitat Politècnica de València (UPV), Spain, in 2017. In 2017 he started his MSc degree in Information Management. He joined the Grid and High Performance Computing research group (GRyCAP) in 2018, while he worked on his Master's Thesis. His research interests are focused in Serverless Computing, Cloud Computing and Container Orchestration Systems.

Most suitable track

Delivering services and solutions

By submitting my abstract, I agree that my personal data is being stored in accordance to conference Privacy Policy

Primary author: RISCO, Sebastián (UPVLC)

Co-authors: Dr NARANJO, Diana M. (UPVLC); Dr CABALLER, Miguel (UPVLC); Dr MOLTO, German (UPVLC)

Presenter: RISCO, Sebastián (UPVLC)

Session Classification: Demonstration