



Contribution ID: 8

Type: **Presentation (15' + 5' for questions)**

Infrastructure Manager: A Deployment Service for the Computing Continuum

Wednesday, 30 October 2024 10:00 (20 minutes)

The Infrastructure Manager (IM) is an open-source production-ready (TRL 8) service used for the dynamic deployment of customized virtual infrastructures across multiple Cloud back-ends. It has evolved in the last decade through several European projects to support the needs of multiple scientific communities. It features a CLI, a REST API and a web-based graphical user interface (GUI), called IM Dashboard that provides users with a set of customizable curated templates to deploy popular software (e.g. JupyterHub on top of a Kubernetes cluster). Users can use the IM to facilitate the deployment of these templates, which follow the TOSCA standard, on whichever Cloud they have access to. This allows easier reproducibility of computational environments and rapid deployment on multiple Cloud platforms such as AWS, Azure, Google Cloud Platform, OpenStack, etc.

The IM has been used in production in the EGI Federated Cloud, one of the largest distributed Cloud infrastructures in Europe, supporting deployment of popular execution environments for scientific users, ranging from data-processing SLURM-based clusters to big data Hadoop-based clusters and customizable elastic Kubernetes clusters. It supports the deployment of Virtual Machines, containers on Kubernetes clusters and functions on AWS Lambda and OSCAR, an open-source serverless platform, to support deployment of infrastructures along the computing continuum.

This contribution summarizes how the IM is being adopted by the current active projects to showcase its functionality. For example, in AI4EOSC, it supports the automated deployment of the Nomad clusters used for training and the OSCAR clusters used for inference of pre-trained AI models. In InterTwin, a rich set of TOSCA templates have been produced to deploy the required software stacks to support the activities of developing a Digital Twin Engine. This is the case of Apache Nifi, KubeFlow, Kafka, AirFlow, MLFlow, Horovod, STAC, etc. In DT-GEO, the IM deploys elastic virtual clusters which mimic the software configuration employed in real HPC (High Performance Computing) clusters so that users get trained in the virtual cluster instead of wasting precious computation time in the actual HPC facilities. In EOSC-Beyond, IM is used as the Deployment Service of the Execution Framework a new EOSC Core service that is technically compatible with the EOSC EU Node.

This work was partially supported by the project AI4EOSC (Grant 101058593), interTwin (Grant 101058386), DT-GEO (101058129) and EOSC-Beyond (101131875). Also, Grant PID2020-113126RB-I00 funded by MICIU/AEI/10.13039/501100011033.

Primary authors: CABALLER, Miguel (Universitat Politècnica de València); CALATRAVA ARROYO, Amanda (Universitat Politècnica de València); MOLTÓ, Germán (Universitat Politècnica de València); BLANQUER ESPERT, Ignacio (Universitat Politècnica de València)

Presenter: MOLTÓ, Germán (Universitat Politècnica de València)

Session Classification: IBERGRID

Track Classification: Developments oriented to foster the Compute Continuum