Contribution ID: **15** Type: **not specified**

Reproducibility of parallel workflows for Digital Twins

Monday, 28 October 2024 11:00 (30 minutes)

The today's computational capabilities and the availability of large data volumes is allowing to develop Digital Twins able to provide unrivaled precision. Geophysics is one field that benefits from the ability to simulate the evolution of multi-physics natural system across a wide spatio-temporal scale range. This is also possible thanks to the access to HPC systems and programming frameworks that can provide paradigms to combine the massive data streams generated by observational systems with large-scale numerical simulations in HPC or cloud environments.

The DT-GEO initiative is implementing interdisciplinary and interrelated DT Components (DTCs) for geophysical hazards from earthquakes (natural or anthropogenically induced), volcanoes, and tsunamis. These DTCs are designed as self-contained and containerized entities embedding flagship codes, AI layers, data streams in workflows.
PyCOMPSs is used for the development and execution of the DTCs as parallel workflows on top of the FENIX Infrastructure that includes supercomputers as Leonardo at CINECA and MareNostrum at BSC. PyCOMPSs is a task-based programming model that simplifies the development of applications for distributed infrastructures, such as HPCs, clouds, and other managed clusters. Its integration enables efficient parallelization of tasks within DTCs workflows, optimizing computational resources and enhancing overall performance. It also provides a lightweight interface to implement dynamic HPC+AI workflows which can change their behaviour at execution time due to exceptions or faults. A parallel Machine Learning library built on top of PyCOMPSs, dislib, is also available to DT-GEO users.

Reproducibility of experiments has become very important in order to guarantee the validation of results in the publication of research papers. The recording of metadata in the form of provenance is one of the most effective ways to achieve reproducibility of workflow experiments.

This contribution provide insights on the programming interfaces to define the DTCs workflows in DT-GEO and explains how the DTCs can be packaged in order to reduce the effort required to deploy them by the underlying services on the computing infrastructure and how can be integrated and reused in different workflows using PyCOMPSs as orchestrator. We will also show how Workflow Provenance is recorded in PyCOMPSs, following the RO-Crate metadata specification, and how this metadata can be used to achieve FAIR workflows through their publication in WorkflowHub.

Primary authors: LEZZI, Daniele (Barcelona Supercomputing Center); Dr SIRVENT, Raul (Barcelona Supercomputing Center)

Presenters: LEZZI, Daniele (Barcelona Supercomputing Center); Dr SIRVENT, Raul (Barcelona Supercomputing Center)

Session Classification: IBERGRID

Track Classification: Design and implementation of Digital Twins