EGU23-6857, updated on 20 Apr 2024
https://doi.org/10.5194/egusphere-egu23-6857
EGU General Assembly 2023

Convergence of HPC, Big Data and Machine Learning for Earth System workflows

Donatello Elia[1], Sonia Scardigno[1], Alessandro D'Anca[1], Gabriele Accarino[1,2], Jorge Ejarque[3], Francesco Immorlano[1,2], Daniele Peano[1], Enrico Scoccimarro[1], Rosa M. Badia[3], and Giovanni Aloisio[1,2]

[1]Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC), Lecce, Italy
[2]Università del Salento, Dept. of Engineering for Innovation, Lecce, Italy
[3]Barcelona Supercomputing Center (BSC), Barcelona, Spain

Typical end-to-end Earth System Modelling (ESM) workflows rely on different steps including data pre-processing, numerical simulation, output post-processing, as well as data analytics and visualization. The approaches currently available for implementing scientific workflows in the climate context do not properly integrate the entire set of components into a single workflow and in a transparent manner. The increasing usage of High Performance Data Analytics (HPDA) and Machine Learning (ML) in climate applications further exacerbate the issues. A more integrated approach would allow to support next-generation ESM and improve the workflow in terms of execution and energy consumption.

Moreover, a seamless integration of components for HPDA and ML into the ESM workflow will open the floor to novel applications and support larger scale pre- and post-processing. However, these components typically have different deployment requirements spanning from HPC (for ESM simulation) to Cloud computing (for HPDA and ML). It is paramount to provide scientists with solutions capable of hiding the technical details of the underlying infrastructure and improving workflow portability.

In the context of the eFlows4HPC project, we are exploring the use of innovative workflow solutions integrating approaches from HPC, HPDA and ML for supporting end-to-end ESM simulations and post-processing, with a focus on extreme events analysis (e.g., heat waves and tropical cyclones). In particular, the envisioned solution exploits PyCOMPSs for the management of parallel pipelines, task orchestration and synchronization, as well as PyOphidia for climate data analytics and ML frameworks (i.e., TensorFlow) for data-driven event detection models. This contribution presents the approaches being explored in the frame of the project to address the convergence of HPC, Big Data and ML into a single end-to-end ESM workflows.