Contribution ID: **86** Type: **Long Talk**

Bridging Cloud and HPC for Scalable Event-driven Processing of AI Workloads

Thursday, 3 October 2024 10:00 (20 minutes)

Cloud computing has revolutionized how we store, process, and access data, offering flexibility, scalability, and cost-effectiveness. On the other hand, High Performance Computing (HPC) provides unparalleled processing power and speed, making it an essential tool for complex computational tasks. However, leveraging these two powerful technologies together has been a challenge.

In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have grown exponentially, with many software tools being developed for the Cloud. Despite this, the potential of integrating these tools with HPC resources has yet to be explored.

Containers have revolutionized application delivery due to their lightweight and versatility. They are standard in cloud-native applications, but new containerization technologies have emerged specifically for HPC environments.

Our team presents a solution for seamlessly integrating Cloud and HPC environments using two essential tools: OSCAR, a Kubernetes-based serverless event-driven platform where the user can easily create services for running jobs within a container, and interLink, a middleware that allows the offloading of tasks created in a Kubernetes cluster to an HPC cluster.

The OSCAR-interLink integration together with iTwinAI, a framework for advanced AI/ML workflows, allows for AI workloads, such as 3DGAN inference, to take advantage of the resources available in HPC, including GPU processing power.

Our results showcase a successful use case, integrating dCache, Apache NiFi, OSCAR, interLink, and iTwinAI, based on a 3DGAN for particle simulation, demonstrating the benefits of the approach by exploiting remote GPUs from an HPC facility from an OSCAR cluster running on a Cloud infrastructure.

This work was supported by the project "An interdisciplinary Digital Twin Engine for science"(interTwin) that has received funding from the European Union's Horizon Europe Programme under Grant 101058386. GM would like to thank Grant PID2020-113126RB-I00 funded by MCIU/AEI/10.13039/501100011033. GM and SL would like to thank project PDC2021-120844-I00 funded by MCIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Topic

Needs and solutions in scientific computing: Digital Twins

Primary authors: PARCERO, Estibaliz (Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València); Mr LANGARITA, Sergio (Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València); CIANGOTTINI, Diego (INFN); BUNINO, Matteo; MOLTO, German (Universitat Politècnica de València); SPIGA, Daniele

Presenter: PARCERO, Estibaliz (Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València)

Session Classification: Processing Research Data with Artificial Intelligence and Machine Learning