Contribution ID: **8** Type: **Long Talk**

The HPC+AI Cloud: flexible and performant infrastructure for HPC and AI workloads

Tuesday, 1 October 2024 17:05 (20 minutes)

In recent years, in particular with the rise of AI, the diversity of workloads that need to be supported by research infrastructures has exploded. Many of these workloads take advantage of new technologies, such as Kubernetes, that need to be run alongside the traditional workhorse of the large batch cluster. Some require access to specialist hardware, such as GPUs or network accelerators. Others, such as Trusted Research Environments, have to be executed in a secure sandbox.

Here, we show how a flexible and dynamic research computing cloud infrastructure can be achieved, without sacrificing performance, using OpenStack. By having OpenStack manage the hardware, we get access to APIs for reconfiguring that hardware, allowing the deployment of platforms to be automated with full control over the levels of isolation. Optimisations like CPU-pinning, PCI passthrough and SR-IOV allow us to take advantage of the efficiency gains from virtualisation without sacrificing performance where it matters.

The HPC+AI Cloud becomes even more powerful when combined with Azimuth, an open-source self-service portal for HPC and AI workloads. Using the Azimuth interface, users can self-service from a curated set of optimised platforms from web desktops through to Kubernetes apps such as Jupyter notebooks. Those applications are accessed securely, with SSO, via the open-source Zenith application proxy. Self-service platforms provisioned via Azimuth can co-exist with large bare-metal batch clusters on the same OpenStack cloud, allowing users to pi the environments and tools that best suit their workflow.

Topic

Needs and solutions in scientific computing: Platforms and gateway

Primary author: PRYOR, Matt (StackHPC)

Co-author: Mr GARBUTT, John (StackHPC)

Presenter: PRYOR, Matt (StackHPC)

Session Classification: Bridging the Gap: Integrating the HPC Ecosystem