Contribution ID: **393** Type: **Oral**

Advancing physics simulation and analysis workflows from customized local clusters to Cori - the HPC optimized sub-million cores system at NERSC

Thursday, 7 November 2019 12:00 (15 minutes)

Abstract: Over the last few years, many physics experiments migrated their computations from customized locally managed computing clusters to orders of magnitude larger multi-tenant HPC systems often optimized for highly parallelizable long-runtime computations. Historically, physics simulations and analysis workflows were designed for a single core CPUs with abundant RAM, plenty of local storage, direct control of the software stack and job scheduler, exclusive access to physically localized hardware, and predictable steady throughput. We will discuss what changes needed to happen in terms of the data pipeline organization, software, and user habits when computations are executed at scale on Cori, where none of those assumptions are true anymore.

STAR experiment at BNL took on the challenge as one of the first. We will discuss the efficient solutions for sustainable processing of experimental data at HPC system 5000 miles away from an experiment, with 2-way just-in-time data transfer. Due to limited administrative privileges at HPC machines, Docker/Shifter become one of the main vehicles to transport custom vetted code to the HPC environment, supplemented by CVMFS software delivery system mounted via DVS servers providing local cache. DayaBay, LZ, ATLAS, and Majorana experiments followed this journey of transformations, by developing schemes for injecting short leaving single core tasks to multi-node, 1000s-core, long run-time jobs scheduled on Cori - the most efficient way to compete with other tenants for CPU cycles. The high variability of scheduling required assembly of 'convoy'jobs composed of many nodes dedicated to the execution of tasks and designating one node to carry a read-only clone of the database. The computing capability of one 'convoy'can be compared to the whole PDSF and one user can schedule 10s of such jobs to run concurrently on Cori. The fine-tuning of tasks concurrency per node to maximize the output for per-node charge-hour given a rather low RAM/CPU ratio and benefiting from 2- or 4-threading capability will be also discussed.

Consider for promotion

No

Primary author: BALEWSKI, Jan (Lawrence Berkeley National Lab. (US))

Co-authors: Mr KRAMER, Matthew; MUSTAFA, Mustafa (Lawrence Berkeley National Laboratory); LEE, Rei; JEFF, Porter; TSULAIA, Vakho (Lawrence Berkeley National Lab. (US))

Presenter: BALEWSKI, Jan (Lawrence Berkeley National Lab. (US))

Session Classification: Track 9 –Exascale Science

Track Classification: Track 9 –Exascale Science