Contribution ID: 420 Type: Poster

STAR Data Production at NERSC/Cori, an adaptable Docker container approach for HPC

Thursday, 13 October 2016 16:30 (15 minutes)

The expected growth in HPC capacity over the next decade makes such resources attractive for meeting future computing needs of HEP/NP experiments, especially as their cost is becoming comparable to traditional clusters. However, HPC facilities rely on features like specialized operating systems and hardware to enhance performance that make them difficult to be used without significant changes to production workflows. Containerized software environment running on HPC systems may very well be an ideal scalable solution to leverage those resources and a promising candidate to replace the outgrown traditional solutions employed at different computing centers.

In this talk we report on the first test of STAR real-data production utilizing Docker containers at the Cori-I supercomputer at NERSC. Our test dataset was taken by the STAR experiment at RHIC in 2014 and is estimated to require ~30M CPU hours for full production. To ensure validity and reproducibility, STAR data production is restricted to a vetted computing environment defined by system architecture, Linux OS, compiler and external libraries versions. Furthermore, each data production task requires certain STAR software tag and database timestamp. In short, STAR's data production workflow represents a typical embarrassingly parallel HEP/NP computing task. Thus, it is an ideal candidate to test the suitability of running containerized software, normalized to run on a shared HPC systems instead of its traditional dedicated off-the-shelf clusters. This direction, if successful, could very well address current and future experiments computing needs. We will report on the different opportunities and challenges of running in such an environment. We will also present the modifications needed to the workflow in order to optimize Cori resource utilization and streamline the process in this and future productions as well as performance metrics.

Tertiary Keyword (Optional)

Secondary Keyword (Optional)

High performance computing

Primary Keyword (Mandatory)

Data processing workflows and frameworks/pipelines

Primary authors: PORTER, Jeff (Lawrence Berkeley National Lab. (US)); Dr LAURET, Jerome (Brookhaven

National Laboratory); Dr MUSTAFA, Mustafa (Lawrence Berkeley National Laboratory)

Presenter: Dr MUSTAFA, Mustafa (Lawrence Berkeley National Laboratory)

Session Classification: Posters B / Break

Track Classification: Track 6: Infrastructures