The Superfacility project: automated pipelines for experiments and HPC

Debbie Bard*, Cory Snavely*, Lisa Gerhardt*, Jason Lee*, Becci Totzke*, Katie Antypas*, Suren Byna*, Ravi Cheema*, Shreyas Cholia*, Mark Day*, Bjoern Enders*, Aditi Gaur*, Annette Greiner*, Taylor Groves*, Mariam Kiran*, Quincey Koziol*, Kelly Rowland*, Chris Samuel*, Ashwin Selvarajan*, Alex Sim*, David Skinner*, Rollin Thomas* and Gabor Torok*
*NERSC,
Lawrence Berkeley National Lab
Berkeley, California 94720
Email: djbard@lbl.gov

***Abstract*—As data sets from DOE user facilities grow in both size and complexity HPC facilities face urgent needs for capabilities to transfer, reduce, analyze, store, search and curate the growing data in order to facilitate scientific discovery. In the past 18 months, NERSC and ESnet have expanded services and designed new capabilities in support of experimental workflows via ASCR's powerful computing, storage and networking resources. In this talk we will introduce the Superfacility project at LBNL- a framework for integrating experimental and observational research instruments with computational and data facilities at NERSC and ESnet. We will discuss the science requirements that are driving this work, and how this has translated into technical innovations in data management, workload scheduling, networking and automation. We will illustrate the impact of this work using examples of teams that are using our systems for real-time experimental data analysis, pushing our infrastructure in new ways. In particular, we will focus on how science teams are keeping up with demanding data rates, the new ways experimental scientists are accessing HPC facilities, and what the future holds for automated data analysis pipelines.**

I. Introduction

Large-scale analysis of experimental data is an increasingly important workload at supercomputing facilities. Moore's Law applies differently to detectors than computers, and some detectors and their data rates can still meet or exceed Moore's exponential curve. The National Energy Scientific Computing Center (NERSC) is the mission HPC and data center for the DOE, and has supported such experimental computing workloads since its inception as a general-purpose supercomputing facility. Today, many of the data analysis and workflow pipelines from experiments have been developed for individual, custom applications that are domain-specific and cannot be reused or shared. Efforts to help a new experimental facility transition their pipeline and analysis tools to an HPC facility like NERSC remain labor-intensive, often resulting in one-off solutions.

The aim of the Superfacility project at LBNL is to develop a more unified, seamless environment that combines hardware solutions, application software and data management tools to deliver breakthrough science. Re-use and re-purposing of previously built workflow components has a strong value proposition as workflow components can be expensive to build. To accomplish this we must enable science applications to run automated pipelines that analyze data from remote facilities at large scale, without routine human intervention, using these capabilities:

- Real-time computing support
- Data management and movement tools
- API-driven automation
- Authentication using Federated Identity
- Dynamic, high-performance networking

The success of this project hinges on close and active engagement with experimental facilities and major science experiments to build tools and optimized pipelines that are of genuine use to the experimental and observational science (EOS) community. These engagements help us identify the commonalities in the needs of these diverse experiments, which will drive the requirements in the goals of scheduling, automation, networking and data management. We have selected initial engagements with experimental teams that have a variety of science drivers and stress the NERSC infrastructure in a variety of ways.

II. Scheduling

One key concept in the support of experimental science is timeliness. Some experiments require supercomputing-scale resources in real time to analyse data from running experiments, leading to requirements for streaming data management, seamless networking across WAN and LAN, and scheduling compute resources with minimal disruption to the always-busy queues for computing resources at NERSC. Other experiments need short-turnaround computing, for example to analyse data from a night's observations at a telescope in time to decide observing plans for the next night, or to make processed data public shortly after acquisition. This also places demands on workload scheduling, data management and integration of these needs within the wider NERSC workload. To meet these needs, we have developed new tools to handle the scheduling of in-demand supercomputing capabilities. A key concern is the loss in system utilization that would result from a large number of nodes being held idle, waiting for an incoming bursty workload from a running experiment. In

order to keep nodes in use, we have designed a protocol to run a preemptible workload that can be canceled at short notice when high-priority jobs come in.

III. Automation

A second key concept is automation. Currently, the supercomputing needs of the experimental community can only be met with the effort of many people, both in advance and during the running of the experiment. Many experiments run with automated analysis pipelines, and these pipelines need a way to communicate with the HPC center without a human getting involved. NERSC is developing an API[1] that allows automation of many of the tasks that currently require a human-in-the-loop, such as advance reservations of compute power, movement of data across the different storage systems, and monitoring job submission. This API also enables important capabilities around workflow resiliency. When NERSC systems are unavailable (due to a scheduled maintenance, for example), the API can be used to query NERSC availability and redirect the workflow to another location as necessary.

NERSC users need to automate not just their compute but also their data management across large collaborations. To this end, NERSC has developed a number of innovations in data management that allow large collaborations to handle their data transfer via centrally-controlled accounts[2], and designed intuitive dashboards for PIs to view and manage their project's data usage at NERSC. We have also implemented a prototype for a new way for teams to archive their data, transferring data from our Spectrum Scale file system to our HPSS tape archive via a simple command-line interface [3].

IV. Modes of access

The third key concept is ease-of-use. Experimental teams don't necessarily need to interact with the supercomputing center via the command line - interactive tools like Jupyter[4] are increasingly being used to control and run experiments, and monitor and analyse data in real time. Jupyter is an interactive open-source web application that allows you to create and share documents called "notebooks," that contain live code, equations, visualizations, narrative text and interactive widgets.

At LBNL, we are developing capabilities to run Jupyter notebooks at HPC-scale, allowing real-time analysis of large-scale datasets, including visualisation of streaming data. We have designed various interactive tools and widgets to help enable exploratory analyses and parameterization across datasets, along with curated notebooks and reproducible workflows, that can be shared, forked, and customized.

Additionally, science teams can also build web science gateways connected with our API, to provide custom domain-centric user interfaces that can be accessed remotely over the web.

[1]https://api.nersc.gov/

[2]https://docs.nersc.gov/services/globus/#using-globus-with-collaboration-accounts

[3]https://docs-dev.nersc.gov/filesystems/ghi/

[4]https://docs.nersc.gov/services/jupyter/

Another element required to make these workflows seamless is federated identity, so a user of multiple experimental and compute facilities can use a single identity to access all the resources they need. We have therefore developed the framework and infrastructure necessary to enable federated identity management across institutions.

V. Networking

The final key concept is networking. The ability to stream data directly into the compute nodes on the supercomputer is a key feature of many pipelines, particularly those that require real-time supercomputing for data analysis. We are enhancing this data movement process by adding the ability to co-schedule network resources concurrently with the job running on the computational nodes. This model allows experimental facilities to schedule the processing and networking requirements simultaneously. We have also implemented advanced networking configurations to ensure fast bandwidth from an external location directly into a supercomputer node. For geographically separated workloads that require real-time feedback (e.g. to adjust instrument parameters for steering or re-calibrate an experiment), the Perlmutter system provides advanced congestion control to reduce the delays experienced on shared systems. These innovations greatly enhance the scientific data processing, allowing a researcher to steer a complex experiment in near real-time.

VI. conclusion

This talk describes the science needs that are driving the technical and policy work required to support experimental workflows at HPC centers. It highlights key areas of technical development in scheduling, automation, data management and ease-of-use, and gives examples of how the tools developed by this project are being used by experimental teams at NERSC today.