Superfacility and Gateways for Experimental and Observational Data

Debbie Bard
Lead, Superfacility Project
Lead, Data Science Engagement Group

Cory Snavely
Deputy, Superfacility Project
Lead, Infrastructure Services Group

August 17, 2020
Superfacility: an ecosystem of connected facilities, software and expertise to enable new modes of discovery

Superfacility@ LBNL: NERSC, ESnet and CRD working together

- A model to integrate experimental, computational and networking facilities for reproducible science
- Enabling new discoveries by coupling experimental science with large scale data analysis and simulations
The Superfacility concept is a key part of LBNL strategy to support computing for experimental science.

- User Engagement
- Data Lifecycle
- Automated Resource Allocation
- Computing at the Edge

Computing Sciences Strategic Initiatives
- Learning
- Beyond Moore
- Superfacility
NERSC supports many users and projects from DOE SC’s experimental and observational facilities.
NERSC supports many users and projects from DOE SC’s experimental and observational facilities. 

~35% of NERSC projects in 2018 said the primary role of the project is to work with experimental data.
Compute needs from experimental and observational facilities continues to increase

Needs go beyond compute hours:

- High data volumes (today use ~19% of computing hours, but store 78% of data.)
- Real-time (or near) turnaround and interactive access for running experiments
- Resilient workflows to run across multiple compute sites
- Ecosystem of persistent edge services, including workflow managers, visualization, databases, web services…

Taken from Exascale Requirements Reviews
Compute needs from experimental and observational facilities continues to increase.

Needs go beyond compute hours:

- High data volumes (today use ~19% of computing hours, but store 78% of data.)
- Real-time (or near) turnaround and interactive access for running experiments
- Resilient workflows to run across multiple compute sites
- Ecosystem of persistent edge services, including workflow managers, visualization, databases, web services...

You will hear much more about this in the next breakout for the NUGX SIG for Experimental Science Users!

Taken from Exascale Requirements Reviews.
Timing is critical

- Experiments may need HPC feedback: real-time scheduling

- Workflow may run continuously and automatically: API access, dedicated workflow nodes

First experiment of LCLS-II: studying protease for SARS-Cov-2 and inhibitors
Data management is critical

• Experiments move & manage data across sites and collaborators

• Scientists need to search, collate and reuse data across sites and experiments
Access is critical

- Experiments have their own user communities and policies: Federated ID

- Scientists need access beyond the command line: Jupyter, API…
The CS Area Superfacility ‘project’ coordinates and tracks this work

Project Goal:
By the end of CY 2021, 3 (or more) of our 7 science application engagements will demonstrate automated pipelines that analyze data from remote facilities at large scale, without routine human intervention, using these capabilities:

• Real-time computing support
• Dynamic, high-performance networking
• Data management and movement tools
• API-driven automation
• Authentication using Federated Identity
We’ve developed and deployed many new tools and capabilities this year...

### Automation to reduce human effort in complex workflows
- Released [programmable API](#) to query NERSC status, reserve compute, move data etc
- Upgraded Spin: Container-based platform to support workflow & edge services
- Designed federated ID management across facilities

### Enabled *time-sensitive workloads*
- Added appropriate scheduling policies, including real-time queues
- Slurm NRE for job pre-emption, advance reservations and dynamic partitions
- Workload introspection to identify spaces for opportunistic scheduling

### Supported HPC-scale *Jupyter* usage by experiments
- Scaled out Jupyter notebooks to run on 1000s of nodes
- Developed real-time visualization and interactive widgets
- Curated notebooks, forking & reproducible workflows

### Deployed *data management tools* for large geographically-distributed collaborations
- Introduced [Globus sharing](#) for collaboration accounts
- Deployed prototype [GHI (GPFS-HPSS interface)](#) for easier archiving
- PI dashboard for collaboration management
Superfacility Annual Meeting Demo series

In May/June we held a series of virtual demonstrations of tools and utilities that have been developed to support the needs of experimental scientists at ESnet and NERSC.

- Recordings available here: https://www.nersc.gov/research-and-development/superfacility/
  - SENSE: Intelligent Network Services for Science Workflows (Xi Yang and the SENSE team)
  - New Data Management Tools and Capabilities (Lisa Gerhardt and Annette Greiner)
  - Superfacility API: Automation for Complex Workflows at Scale (Gabor Torok, Cory Snavely, Bjoern Enders)
  - Docker Containers and Dark Matter: An Overview Of the Spin Container Platform with Highlights from the LZ Experiment (Cory Snavely, Quentin Riffard, Tyler Anderson)
  - Jupyter, Matthew Henderson (w. Shreyas Cholia and Rollin Thomas)

- Planning a second demo series in the Fall as we roll out next round of capabilities
Priorities for 2020

1. Continue to deploy and integrate new tools, with a focus on the top “asks” from our partner facilities
   o API, Data management tools, Federated ID
2. Resiliency in the PSPS era
   o Working with NERSC facilities team to motivate center resilience
   o Working with experiments to help build more robust workflows
     • eg cross-site data analysis for LZ, DESI, ZTF, LCLS: using ALCC award and LDRD funding
3. Perlmutter prep
   o Key target: at least 4 superfacility science teams can use Perlmutter successfully in the Early Science period
Perlmutter was designed to include features that are good for Superfacility

- The Supernova Cosmology Project, lead by Perlmutter, was a pioneer in using NERSC supercomputers combine large scale simulations with experimental data analysis
  - Advocate for and proponent of “team science”
• **Slingshot is Ethernet compatible**
  – Blurs the line between the inside/outside machine
  – Allow for seamless external communication
  – Direct interface to storage

4D-STEM microscope at NCEM will directly benefit from this
• Currently has to use **SDN** and direct connection to NERSC network to stream data to Cori compute nodes
  – uses a buffer into the data flow to send data to Cori via TCP, avoiding packet loss
All-Flash scratch Filesystem

• Fast across many dimensions
  – 4 TB/s sustained bandwidth
  – 7,000,000 IOPS
  – 3,200,000 file creates/sec

• Optimized for NERSC data workloads
  – NEW small-file I/O improvements
  – NEW features for high IOPS, non-sequential I/O

Astronomy (and many other) data analysis workloads will directly benefit from this
• IO-limited pipelines need random reads from large files and databases
Demo: a Science Gateway in 5 Minutes
Motivation for Spin

“How can I run services alongside HPC that can...

... access file systems
... access HPC networks
... scale up or out
... use custom software

... outlive jobs (persistence)
... schedule jobs / workflows
... stay up when HPC is down
... be available on the web

and are managed by my project team?”
Many Projects Need More Than HPC

Spin answers this need.

Users can deploy their own science gateways, workflow managers, databases, and other network services with Docker containers.

- Use public or custom software images
- Access HPC file systems and networks
- Orchestrate complex workflows
- ...on a secure, scalable, managed platform
Spin Embraces the Docker Methodology

**Build**
images on your laptop with your custom software, and when they run reliably, …

**Ship**
them to a registry for version control and safekeeping
- DockerHub: share with the public
- NERSC: keep private to your project

**Run**
your workloads

Spin
Use a UI, Dockerfile, YAML Declarations...

my-project.yml

```yaml
baseType: workload
containers:
  name: app
  image: flask-app:v2
  imagePullPolicy: always
  environment:
    TZ: US/Pacific
  volumeMounts:
    - mountPath: 
      name: 
      type: 
      readOnly: false
...
```

Dockerfile

```bash
FROM ubuntu:18.04
RUN apt-get update --quiet -y && \
    apt-get install --quiet -y \
    python-flask
WORKDIR /app
ENTRYPOINT ["python"]
CMD ["app.py"]
```
...to create running services.

A typical example:

1. multiple nginx frontends
2. custom Flask backend
3. database or key-value store (dedicated, not shared)
4. private overlay network.

*Rancher starts all the containers and ensures they stay running.*
High-Level Spin Architecture

User-managed

NERSC handles the rest!

docker

management UI/CLI

security policy enforcement

image registry

web frontend 1

web frontend 2

app backend

database

key-value

node 1

node 2

... node n

CFS

CVMFS

NFS

CFS

NERSC

handles
the rest!
Demo: Creating a Service in Spin
Learn More about Spin

Attend a SpinUp Workshop to learn how you can build your own science gateways!

More info: https://www.nersc.gov/systems/spin/
New API functionality: https://api.nersc.gov/

- Workflow automation needs to interact w/ NERSC w/o a human in the loop:
  - Eg beamline at NSLS-II wants to send data for analysis
  - Requirements based on detailed survey in winter 2019
    - Ask questions like:
      - Is NERSC in maintenance?
      - When are future maintenances scheduled?
      - Is the scratch file system available?
    - Perform actions like:
      - Move my data
      - Launch a job
      - Make a reservation...
- Finalizing authentication model and implementation
  - Not yet visible to users - pending completion and security review
- Staff to contribute via Gitlab-based process
New Data Movement tools deployed

- Large collaborations (e.g., LZ, LSST-DESC) struggle to manage their data between CFS and HPSS
  - **GHI is deployed to early users**
    - Easy way to archive data from CFS using command line tools
    - Automatically bundles data to optimal HPSS size
  - Experiments often share the data management between multiple staff - we use collab accounts to enable this
    - **Collaboration accounts enabled for Globus sharing**
    - Dedicated endpoint allows specified users to transfer data in as collab user, no extra step needed to manage permissions
  - PIs of large teams often have to ask NERSC to chown/chgrp collaboration data when users leave or mess up their permissions
    - **PI Data Dashboard enables these actions via a click of a button**
Areas of Technical Work

**Advanced Scheduling; Resiliency**
Support forecasted real-time computing demands

**Software-Defined Networks; SENSE; Self-Managed Systems**
Provide on-demand connectivity, QoS, fault handling, etc

**Data Movement; Data Dashboard; HDF5**
Simplify data management tasks and optimize data production and analysis

**Spin: Containers-as-a-Service Platform**
Support “edge services” adjacent to HPC for workflows

**API and Federated Identity**
Automate it all and use modern cross-facility authentication

---

**Drivers:**
- Complex workflows
- Data-driven projects
- Real-time compute
- Streaming instrument data
LLAna: LCLS-LBNL Data Analytics Collaboration

Pilot to design and deploy a new computing environment for the next generation of free electron lasers: tools for composable workflows, data management and analysis.

HDF5 for high-performance file access and management, designed for LCLS-II needs

Workflow profiling, characterization and optimization for real-time LCLS-II analysis on HPC resources

Detector

Data Reduction Pipeline

Online Monitoring

Fast feedback storage

Fast Feedback

Offline storage

LCLS-II or BES facility generating HDF5

Jupyter for shared analysis notebooks, with HPC backend

BERKELEY

Bringing Science Solutions to the World

HPC
The NERSC-9 Project is Proceeding Well

- Scope, Cost, Schedule
- Facility Upgrade on Track
- Health & Safety Processes
- CD 2/3
- App Readiness Progress
- Staff Experience
- System Contract Award
- Risks Defined & Managed
- Well Trained CAMS

Annual Project Review Nov. 5-6, 2019

Only 1 recommendation: Continue prioritization of hiring a permanent lab project manager

12.5 MVA power upgrade and associated cooling for N9 underway
Hopper Memory Usage

• Feugiat, facilisis mauris.
• Erat arcu lorem donec sceleris
• Parturient

Caption: Memory used on Hopper by the NERSC workload in 2013.
A Table

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Alternative Section Divider