Modes of Operation: perspectives from NERSC

NERSC

Debbie Bard
Acting Group Lead
Data Science Engagement Group
NERSC, LBNL

NITRD workshop on the convergence of HPC, data and ML
Oct 29th, 2018

Office of Science Laboratories

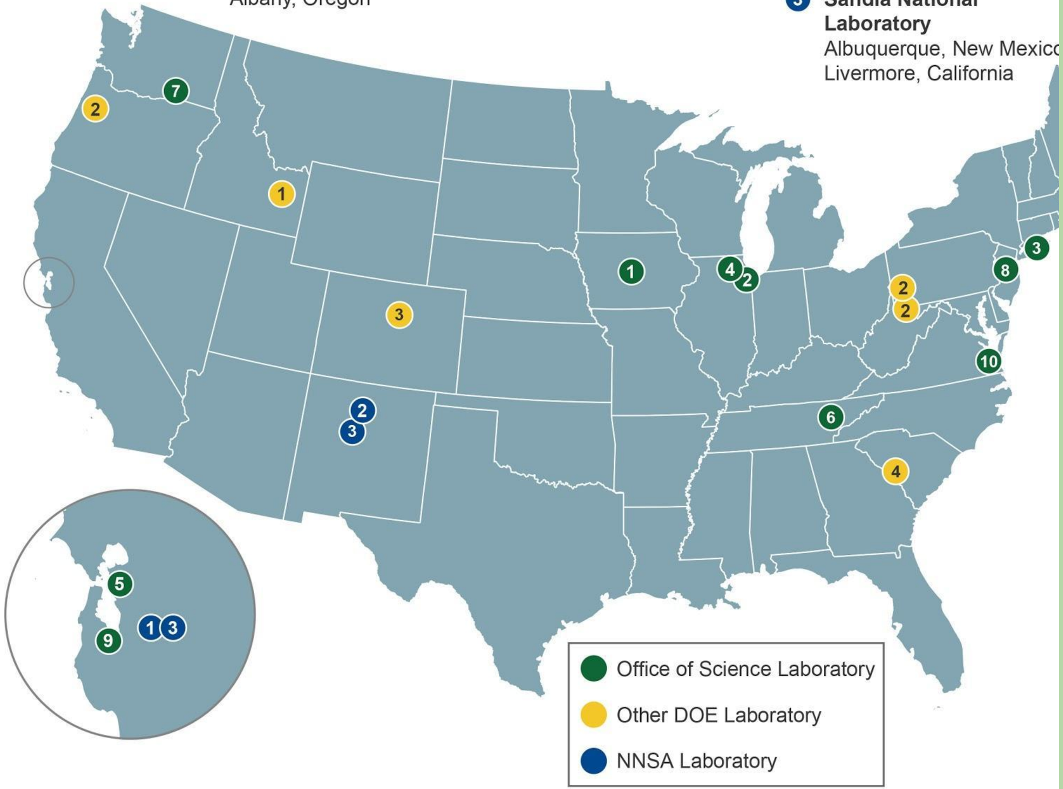
1. **Ames Laboratory**
   Ames, Iowa
2. **Argonne National Laboratory**
   Argonne, Illinois
3. **Brookhaven National Laboratory**
   Upton, New York
4. **Fermi National Accelerator Laboratory**
   Batavia, Illinois
5. **Lawrence Berkeley National Laboratory**
   Berkeley, California
6. **Oak Ridge National Laboratory**
   Oak Ridge, Tennessee
7. **Pacific Northwest National Laboratory**
   Richland, Washington
8. **Princeton Plasma Physics Laboratory**
   Princeton, New Jersey
9. **SLAC National Accelerator Laboratory**
   Menlo Park, California
10. **Thomas Jefferson National Accelerator Facility**
    Newport News, Virginia

Other DOE Laboratories

1. **Idaho National Laboratory**
   Idaho Falls, Idaho
2. **National Energy Technology Laboratory**
   Morgantown, West Virginia
   Pittsburgh, Pennsylvania
   Albany, Oregon
3. **National Renewable Energy Laboratory**
   Golden, Colorado
4. **Savannah River National Laboratory**
   Aiken, South Carolina

NNSA Laboratories

1. **Lawrence Livermore National Laboratory**
   Livermore, California
2. **Los Alamos National Laboratory**
   Los Alamos, New Mexico
3. **Sandia National Laboratory**
   Albuquerque, New Mexico
   Livermore, California



U.S. DEPARTMENT OF ENERGY

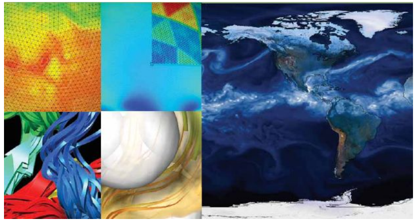
- 17 National Labs
- 29 User Facilities, including:
  - 5 x-ray light sources
  - 5 nanoscale science research centers
  - 4 compute facilities
- Connected by ESNet

NERSC is the Production HPC & Data Facility for DOE Office of Science

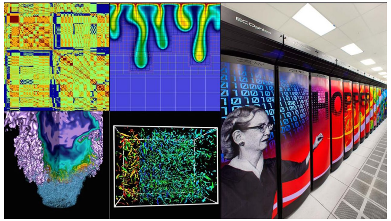
NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

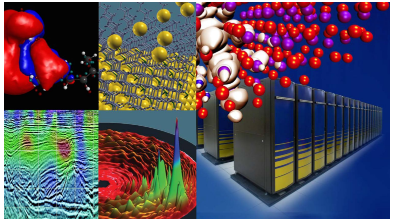
Largest funder of physical science research in U.S.



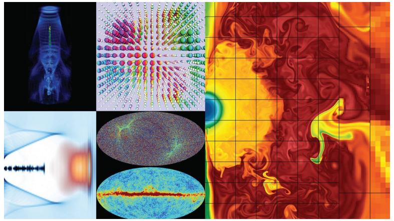
Biology, Energy, Environment



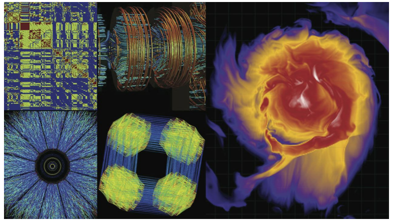
Computing



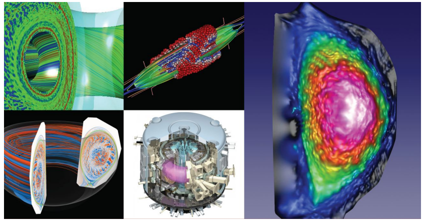
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

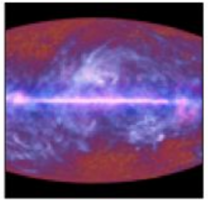
U.S. DEPARTMENT OF ENERGY

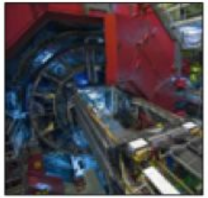
7000+ users, 750+ groups, 2000+ papers/year

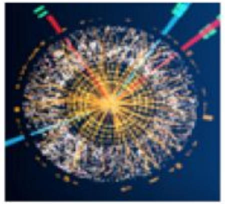
BERKELEY LAB

NERSC has a long history of working with experimental and observational data facilities

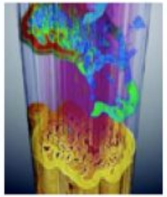
Palomar Transient Factory Supernova


Planck Satellite Cosmic Microwave Background Radiation


Alice Large Hadron Collider

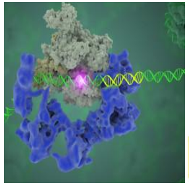

Atlas Large Hadron Collider

Dayabay Neutrinos

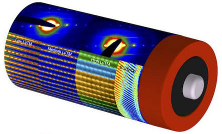

ALS Light Source

LCLS Light Source

Joint Genome Institute Bioinformatics


Cryo-EM

Office of Science


NCEM

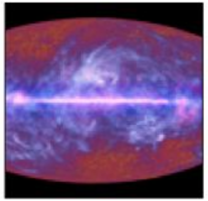
DESI

In 2017, **~35%** of projects self identified as confirming the primary role of the project is to:

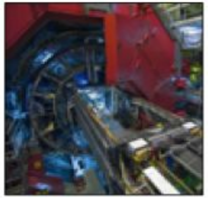
1. analyze experimental data
2. create tools for experimental data analysis
3. combine experimental data with simulations and modeling

NERSC has a long history of working with experimental and observational data facilities

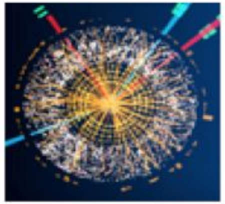
Palomar Transient Factory Supernova



Planck Satellite Cosmic Microwave Background Radiation

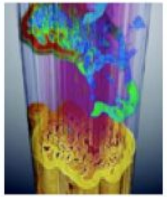


Alice Large Hadron Collider



Atlas Large Hadron Collider

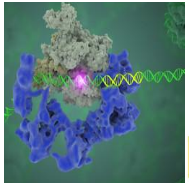
Dayabay Neutrinos



ALS Light Source

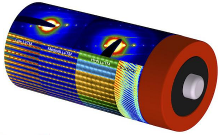
LCLS Light Source

Joint Genome Institute Bioinformatics



Cryo-EM

Office of Science



NCEM

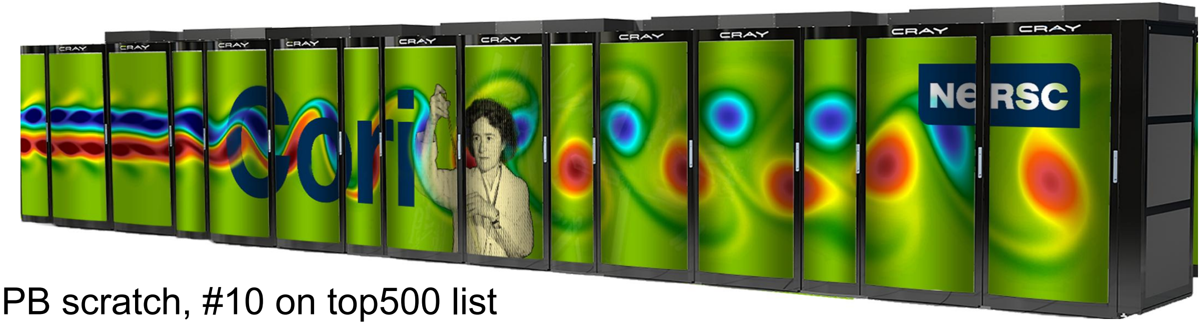
DESI

What's changing?

- ***Experiments are producing larger datasets***
- Scientists are integrating simulation and data analysis at large scales
- Scientific workflows have become more complex
- New advances in machine learning and tools available to non-experts
- New memory, storage, processor and accelerator technologies are available

NERSC systems are designed with both simulation and data users in mind

- High bandwidth external connectivity to experimental facilities from compute nodes (Software Defined Networking).
- NVRAM Flash Burst Buffer as I/O accelerator.
- More login nodes for managing advanced workflows.
- Support for real time and high-throughput queues with Slurm.
- Virtualization capabilities with Shifter (docker containers).
- Optimised and scalable analytics software stack (python, Julia, machine learning).



Cori: 27 PFlops, 28PB scratch, #10 on top500 list

Requirements reviews and users from experimental facilities describe numerous pain points

- **Workflows** require manual intervention and custom implementations
- Difficult to surge experimental codes at HPC facility in ‘**real-time**’
- I/O performance, storage space and access methods for **large datasets** remain a challenge
- Searching, publishing and sharing **data** are difficult
- **Analysis codes** need to be adapted to advanced architectures
- Lack of **scalable analytics software**

Research

- **Resilience strategy** needed for fast-turnaround analysis needs
  - including: coordinating maintenances, fault tolerant pipelines, rolling upgrades, alternative compute facilities...
- No **federated identity** between experimental facilities and NERSC
- Not all scientists want command-line access.

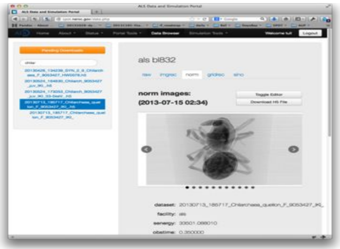
Policy

Work in progress

- **Resiliency planning**
  - *Avoiding NERSC downtimes*: rolling upgrades, automated fault detection and diagnosis...
  - **Portability between sites**: containers, VMs, agreement on software environments e.g. ECP)...
- **Surge computing requirements from experiments while supporting existing workload**
  - *Real-time queues*
    - How to handle idle nodes waiting on incoming compute jobs?
    - Automatic checkpointing of jobs? Killable jobs?]
- **Make it easier to access info and interact with NERSC: designing a "SuperFacility" API**
  - Outages, status, location of data, queue occupancy, ports available…

Policy +
Smart Operations

Superfacility: A network of connected facilities, software and expertise to enable new modes of discovery



als bl832

norm images:
(2013-07-15 02:34)

dataset: 20130713_185717_Chilarchaea_quellon_F_9053427_IKI_

facility: als

senergy: 33501.088010

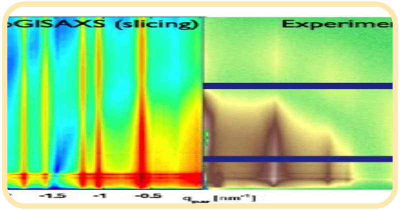
obstime: 0.350000

Real-time analysis and Data management

Experimental Facilities

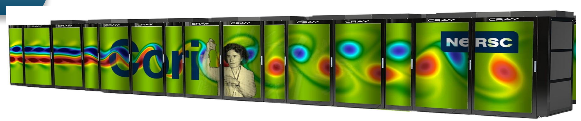
Emerging common needs across DoE SC offices and experimental facilities

New Mathematical Analyses



Fast Implementations on latest computers

Computing Facilities



ESnet
ENERGY SCIENCES NETWORK

Network for Big Data Science

Important policy questions

- **Users analyzing data from experimental facilities value fast turn around access**
  - Is utilisation still the right reportable metric?
  - Workflow turnaround time?

- **Different sets of users have varying access needs:**
  - Read-only data access
  - Limited functionality access through web interface (e.g. Jupyterhub)
  - Full shell access
  - ***How to handle identification/authorisation?***

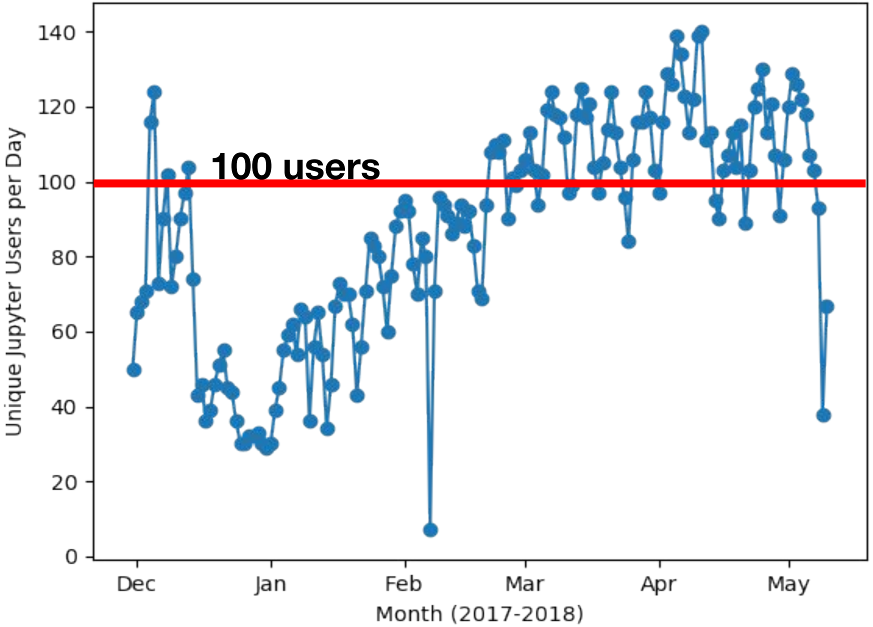
Jupyter is very popular with NERSC Users

- **Over 600 unique users of Jupyter on Cori over past 9 months**

- **> 50% of users who submitted jobs have used Jupyter on Cori**

- **We want users to have a:**

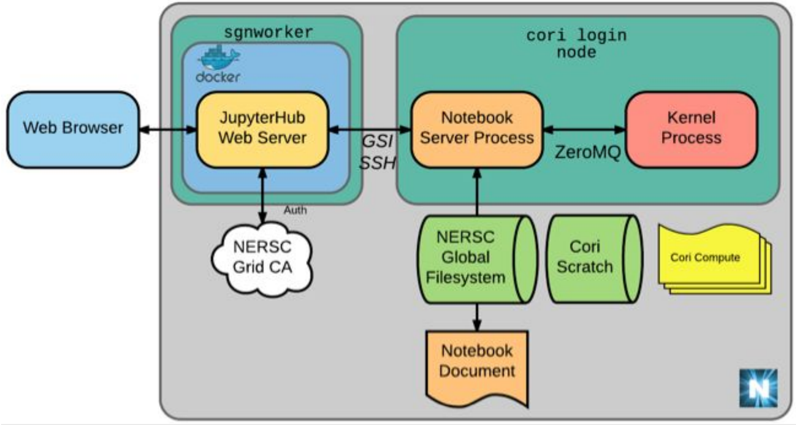
familiar Python environment
productive Python experience
performant Python software stack



"… jupyter notebooks are very important for me:
The 3 most important things in life: food, shelter and jupyter… everything else is optional." -- NERSC User

Jupyter at NERSC

- **Architecture to run on Cori - currently mainly runs on one server**
  - Access to file-system and batch-system
  - OK for plotting etc.- not so for e.g. distributed deep learning



- **Various models for expanding compute onto Cori compute nodes**
  - NERSC SLURM magic
  - Interactive notebook connecting to ipyparallel/dask
  - Future plans to spawn general purpose servers/kernels easily/automatically

Takeaways

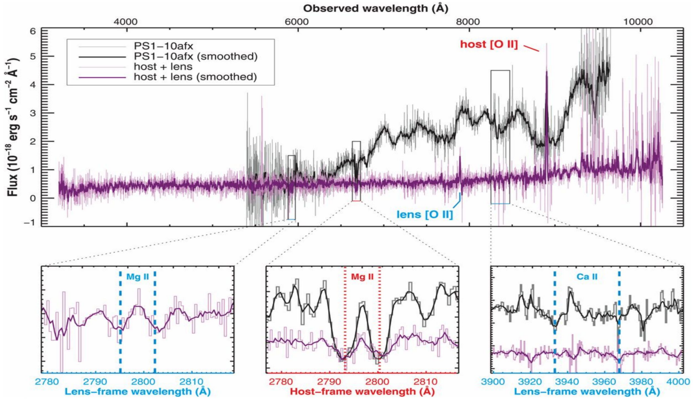
- **We talk to our users, and we listen to what they're telling us**
  - Increasing number of users self-identify as "data users"
  - Design our systems specifically with data requirements in mind
  - Experimental facilities have specific requirements to support their real-time computing needs
- **Data users need:**
  - Resiliency planning
  - Fast turnaround of compute jobs
  - New ways of accessing data and compute
  - Better and more appropriate interfaces to NERSC
  - *Policies that prioritise their needs*

Example: LSST-DESC

Data rates ~10 TB/night.

- **Supernova detection pipeline: ingest alert stream from NCSA via ESNet**
  - In-stream data analysis to detect events of interest
  - Compare to O(100)TB reference data
  - AI for fast detection of “interesting” objects
- **Regular data processing and simulation analytics need:**
  - Fast access to large external DBs
  - Next day turnaround of analysis
  - Publish and serve catalogs for general publi

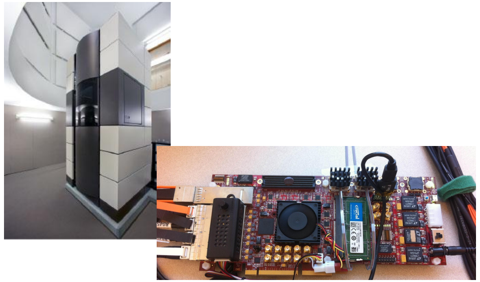
LSST - DESC



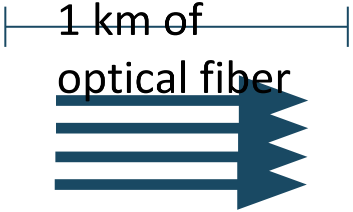
From Quimby et al. (2014), Science: Vol. 344, Issue 6182,

Example: NCEM

- **A fast framing detector with high throughput readout system**
- **High bandwidth networks for data transfer**
- **Requires computing power/memory of a supercomputer to handle data**
    - ○ Current project: Attach instrument directly to NERSC network: stream up to 400GB/s directly to SSD storage inside supercomputer
    - ○ Use data to train AI algorithm which will then be deployed on FPGAs close to instrument to down-filter data stream

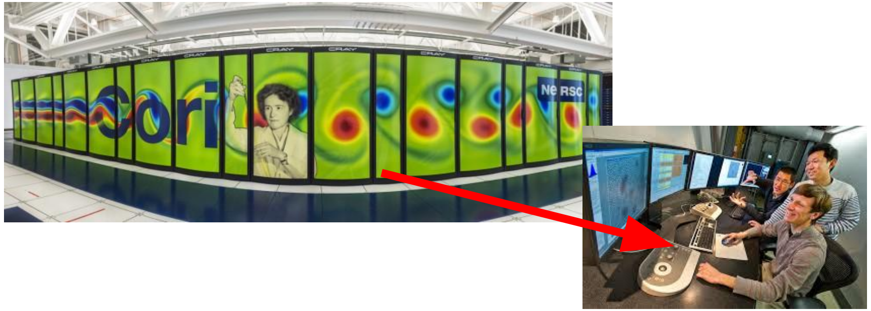


FPGA based readout system



1 km of optical fiber

4 x 100 GbE



"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov