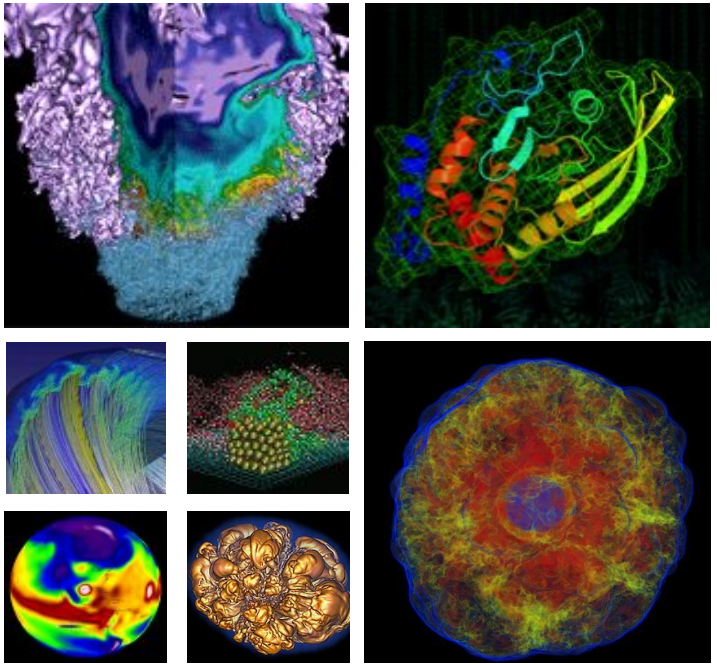
User-Friendly Data Management for Scientific Computing Users



Kirill Lozinskiy

Lisa Gerhardt

Annette Greiner et al.

CUG 2019 / May 9, 2019

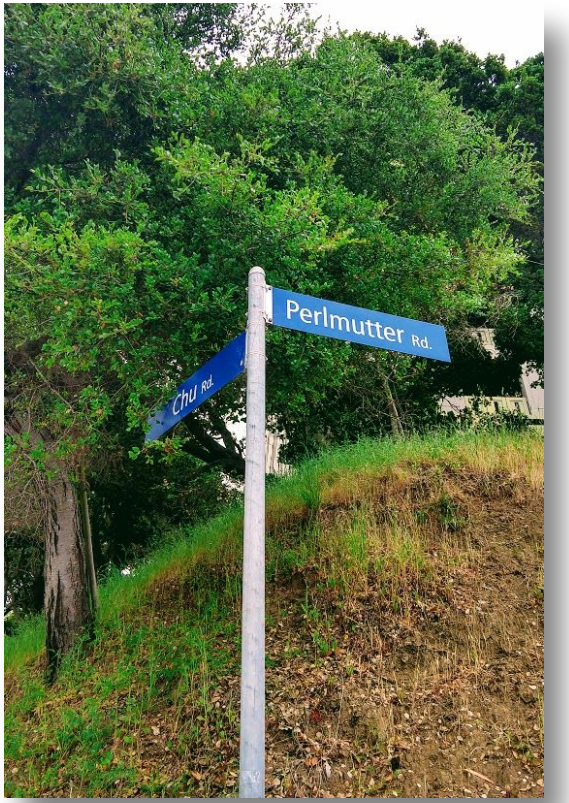
NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

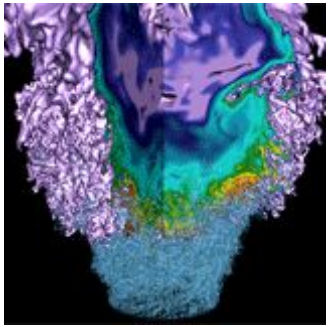
BERKELEY LAB

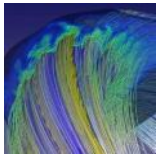
Agenda

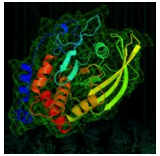
- Introduction and Motivation
  - NERSC
  - User Facilities
  - Data Environment
  - NERSC-8 / NERSC-9
  - Data Management Problems
- Data Dashboard
  - User-Centered Design Process
  - Implementation
  - UI
  - Permissions Wrangler
- Future Plans
- Conclusion

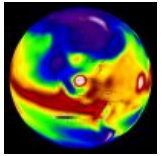


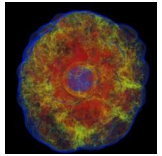
NERSC Overview

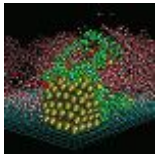








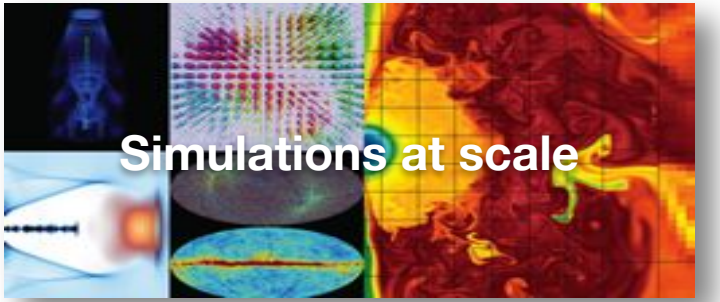


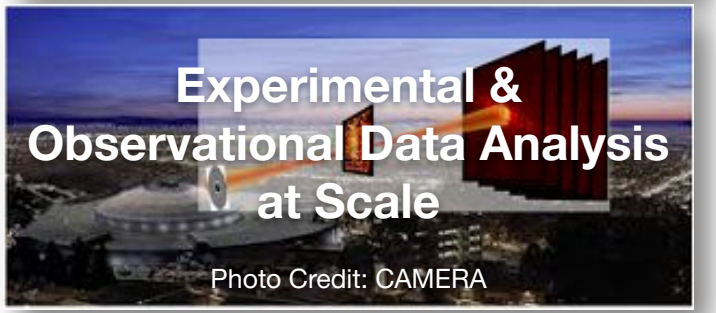


NERSC

NERSC @ Berkeley Lab (LBNL)

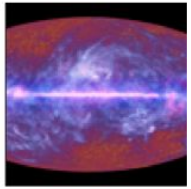
- NERSC is the mission HPC computing center for the DOE Office of Science
- HPC and data systems for the broad Office of Science community
- 7,000 Users, 870 Projects, 700 Codes
- >2,000 publications per year
- 2015 Nobel prize in physics supported by NERSC systems and data archive
- Diverse workload type and size
  - Biology, Environment, Materials, Chemistry, Geophysics, Nuclear Physics, Fusion Energy, Plasma Physics, Computing Research
- New experimental and AI-driven workloads





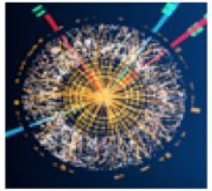
NERSC supports a large number of users and projects from DOE SC's experimental and observational facilities

Palomar Transient Factory Supernova

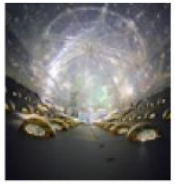


Planck Satellite Cosmic Microwave Background Radiation

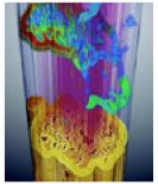
Star Particle Physics



Atlas Large Hadron Collider



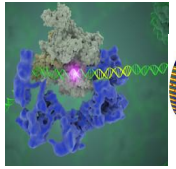
Dayabay Neutrinos



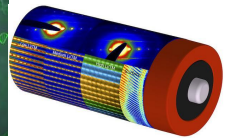
ALS Light Source

LCLS Light Source

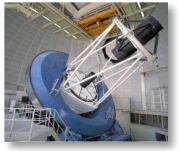
Joint Genome Institute Bioinformatics



Cryo-EM

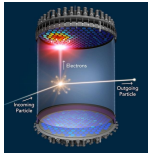


NCEM

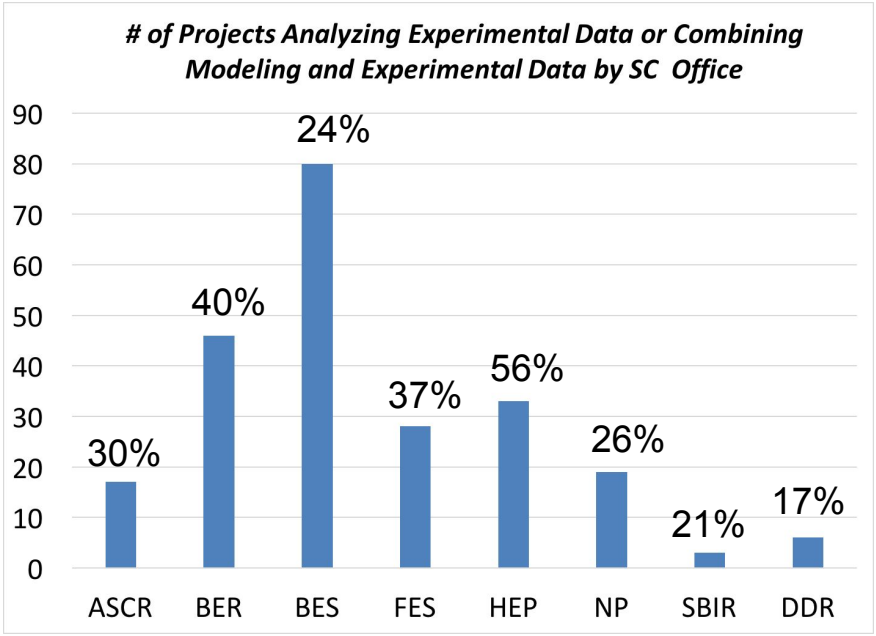


DESI

LSST-DESC



LZ



~35% (235) of ERCAP projects self identified as confirming the primary role of the project is to 1) analyze experimental data or; 2) create tools for experimental data analysis or; 3) combine experimental data with simulations and modeling

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Science Engagements

ALS

Synchrotron light source uses NERSC for real-time experimental feedback, data processing/management, and comparison to simulation

LSST DESC Dark Energy Science Collaboration

Processing streaming alerts (from NCSA) for detection of supernova and transient gravitational lensing events

LCLS

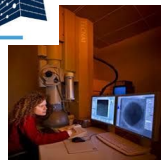
High-rate detectors use ESnet and NERSC for real-time experimental feedback and data processing

FICUS Facilities Integrating Collaborations for User Science

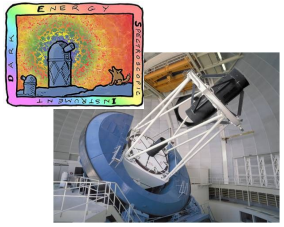
JGI-EMSL Collaborative Science Call

Complex multi-stage workflow to analyse response of soil microbes to climate change

MOLECULAR FOUNDRY



4D STEM data streamed to NERSC, used to design ML algorithm for future deployment on FPGAs close to detector



Nightly processing of galaxy spectra to inform next night's telescope targets

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Example: Synchrotron light sources

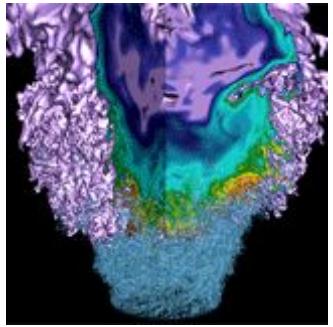
Big picture

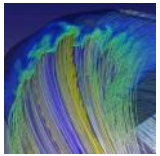
- Coherent or full-field experiments use high frame rate 2D detectors for their science.
- High data volume that needs real-time feedback and sharing with non-facility users.

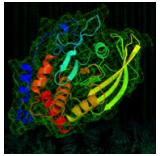
Data Lifecycle

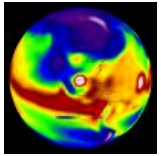
- 30 - 60 TB of data per week per detector, expected to increase x10 in 2025.
- Data is copied from experimental facility and processed, stored in the archive, and shared **only with the specific beamline scientists.**
- Data movement and processing is triggered by the beamline maintainer and scientist only consume finished products.

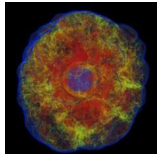
NERSC Data Environment

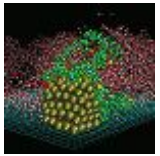






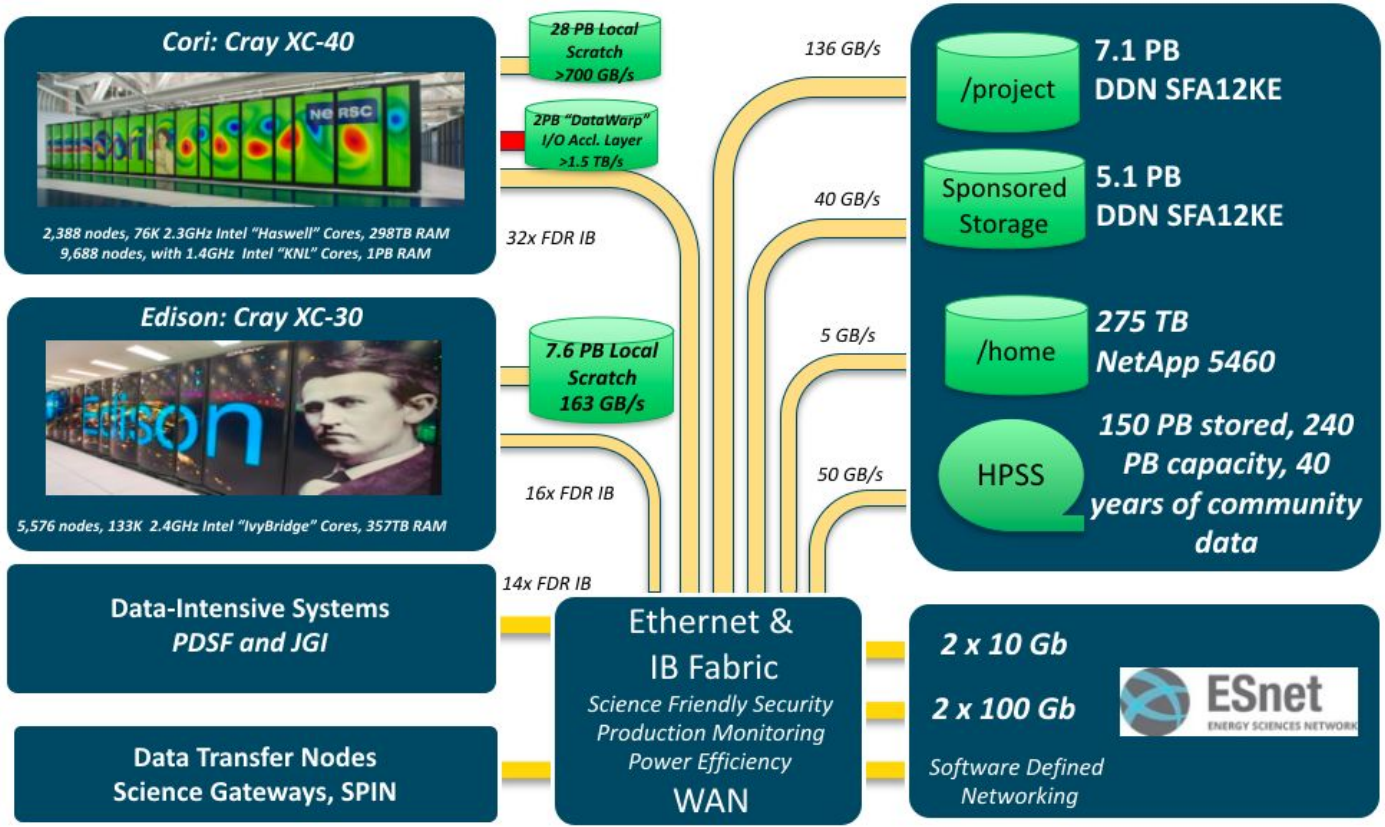






NERSC

HPC and Storage at NERSC



NERSC-8 aka Cori (Cray XC-40)

Compute

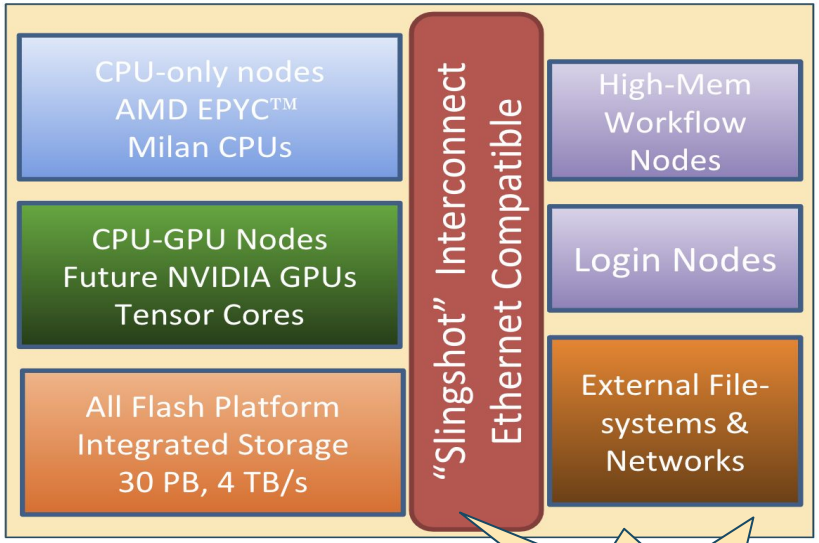
- 9,688 Intel KNL nodes
- 2,388 Intel Haswell nodes

Storage

- 30 PB, 700 GB/s scratch
  - Lustre (Cray ClusterStor)
  - 248 OSSes x 41 HDDs x 4 TB
  - 8+2 RAID6 declustered parity
- 1.8 PB, 1.5 TB/s burst buffer
  - Cray DataWarp
  - 288 BBNs x4 SSDs x 1.6 TB
  - RAID0

NERSC-9 aka Perlmutter

- Designed for both large scale simulation and data analysis from experimental facilities
- Overall 3x to 4x capability of Cori
- Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
- Slingshot Interconnect
- Single Tier, All-Flash Lustre scratch filesystem



Delivery in late-2020

Multiple Storage Tiers

Lustre “scratch” and Burst Buffer

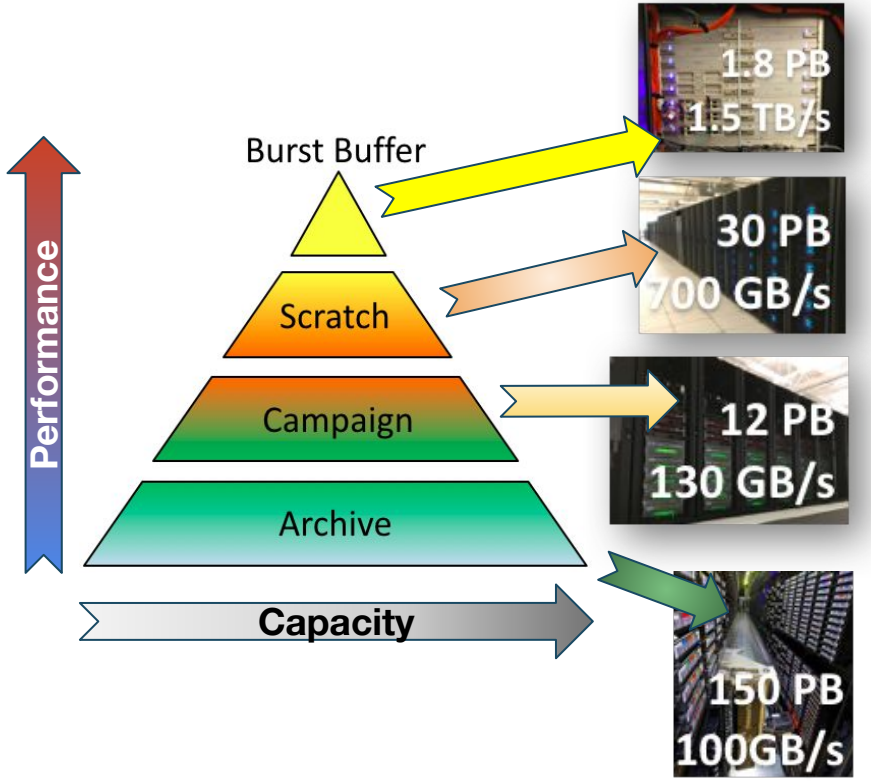
- Ephemeral storage, data purged if not accessed, user-based quotas and permissions
- Intended for high speed access to active data used for running computations

Spectrum Scale “project” file system

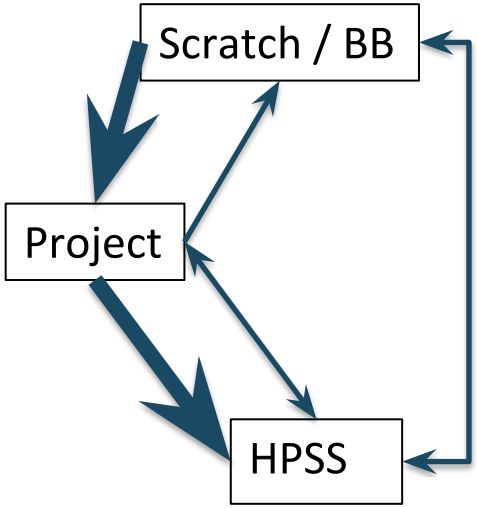
- Medium term storage, data never purged, group quota and permissions
- Intended for shared data needed by entire science group, will be used for computing in the near future

HPSS Tape Archive

- Long term storage, data never purged, user-based and group quotas and permissions
- Permanent archival of scientific data



Data Lifecycle at NERSC



Lively file system

- Scratch has 16 PB of data and 1.2B inodes. Average 18M files modified and 0.4 PB of data read per day
- Project has 7PB and 1B inodes. Average fluctuation in number of inodes is 1M and 7 TB of data deleted or created per day
- HPSS increases by ~30TB / day (at times hundreds of TBs)

Moving and managing data is consuming a larger and larger fraction of scientists time

Only traditional Linux tools

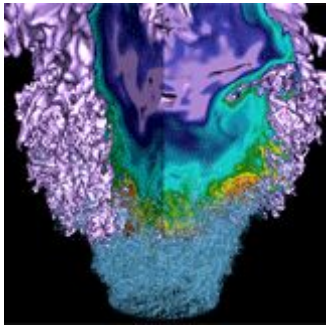
- “ls”, “find”, and “du” to see usage
- All movement between tiers is manual: “cp”, “mv”, “hsi put”, “htar”

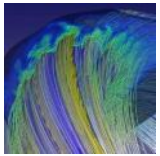
```
cori04> nohup du -h /project/projectdirs/mpccc/lgerhard/
nohup: ignoring input and appending output to 'nohup.out'
```

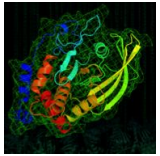
Data Management Problems

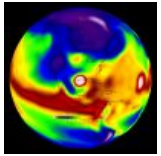
- Common issues across groups at NERSC
  - Finding files and understanding data
  - Moving data internally across tiers at the center
  - Triggering workflows
- Complexity is only going to increase with data volumes
- Ambitious project, started by dealing with the most common requests for help which are centered around our “project” file system
- Space is shared among science groups at NERSC (as large as 300 members)
  - Project managers must “du” and manually nag members to clean up to stay under quota
  - Permission drifts so that files are no longer group readable
  - Users/PIs can’t find their data

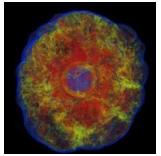
Data Dashboard

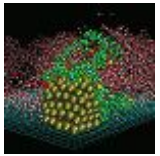












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

User-Centered Design Process

Needs Assessment:

- Interviews with PIs and proxies
- Paper prototype
- Usability testing
- Interactive prototype
- More usability testing
- Ongoing discussions with users about proposed features

Data Dashboard - Implementation

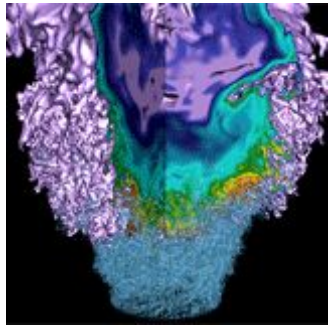
- Uses daily Spectrum Scale (ILM) scans
  - Outputs full path, ctime, mtime, atime, size, owner uid and gid for every file and directory
  - Text file ~200GB w/1 billion lines
  - Scans are facilitated by housing file system metadata on SSD by scaling the number of servers doing the scan
- Automatically detects when scan is finished and submits a Spark job to the batch system, output is JSON files and PostgreSQL database

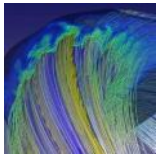
Data Dashboard - Implementation

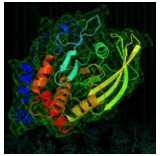
“Biggest Files and Dirs” and File Browser

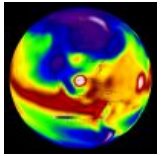
- On user request, web page within MyNERSC connects via NEWT (API), triggers PHP script
- Script gets metadata for files and directories owned by user from PostgreSQL
- Visualizations rendered from PostgreSQL data with D3 JavaScript visualization library

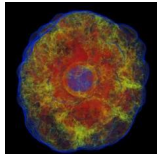
Demonstration

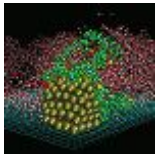










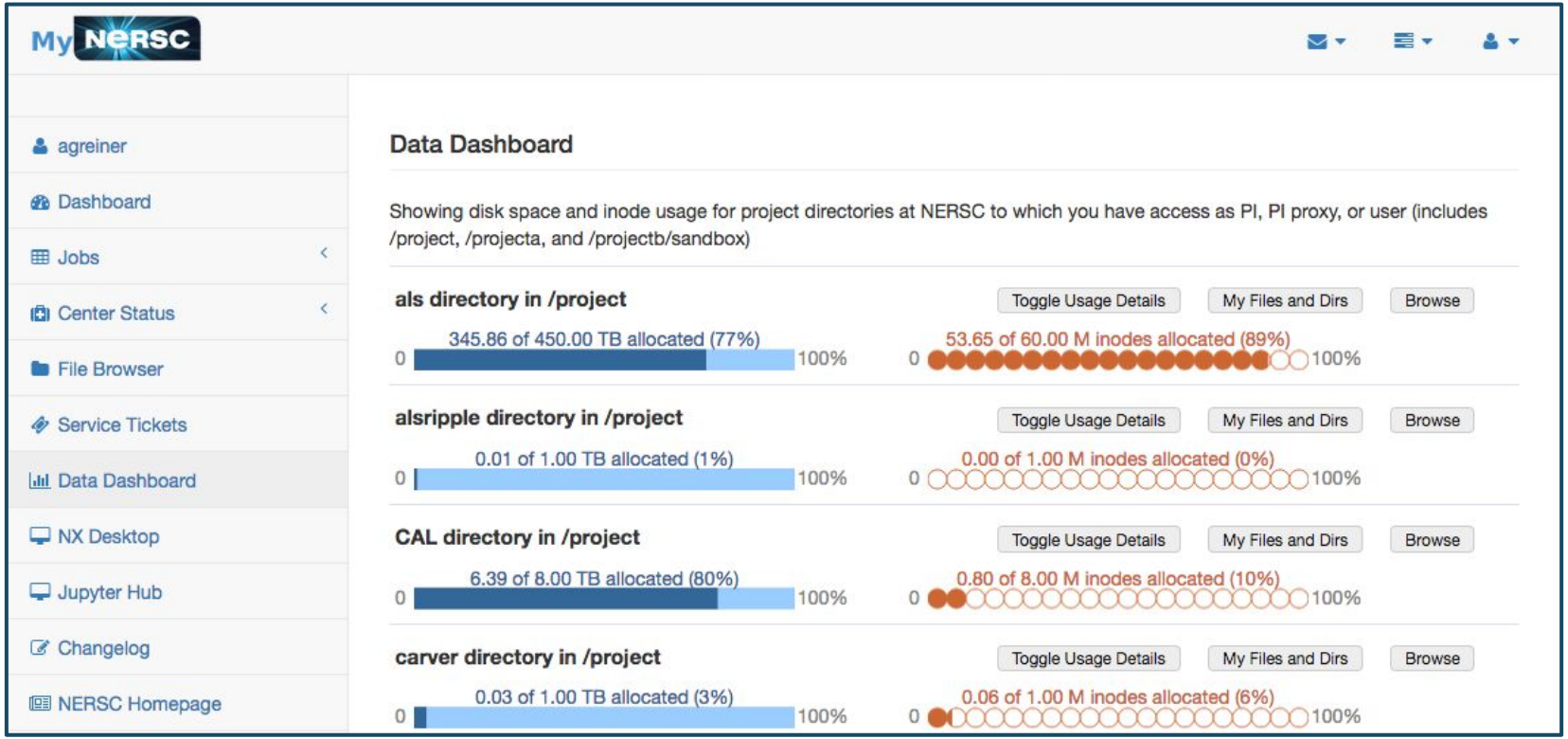


NERSC

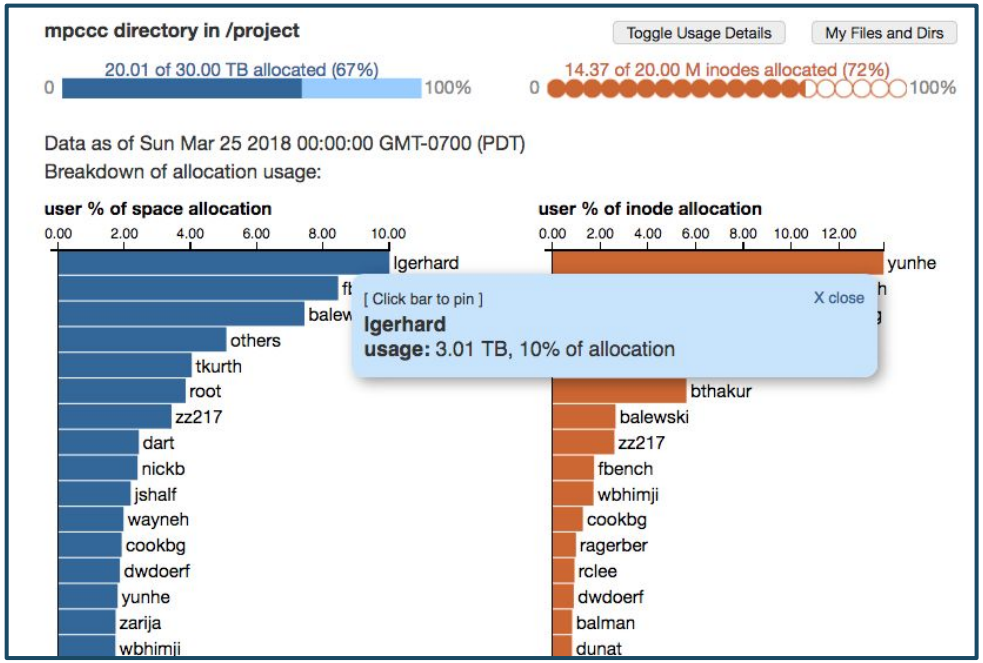
U.S. DEPARTMENT OF ENERGY | Office of Science

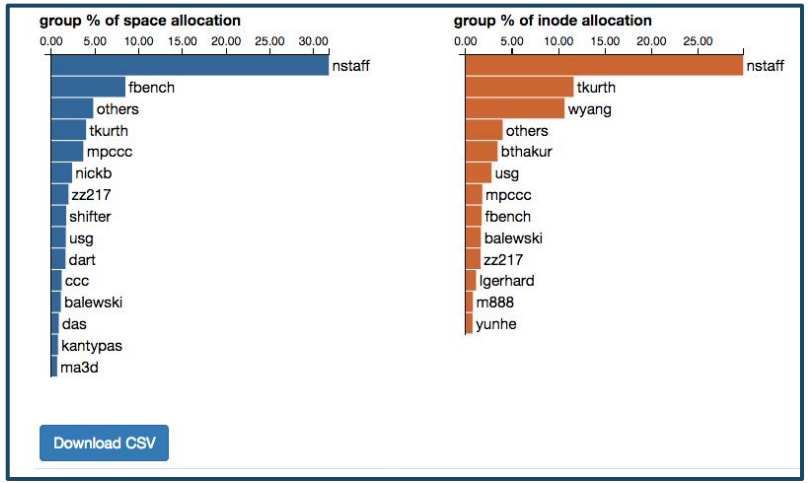
BERKELEY LAB
Lawrence Berkeley National Laboratory

Data Dashboard UI



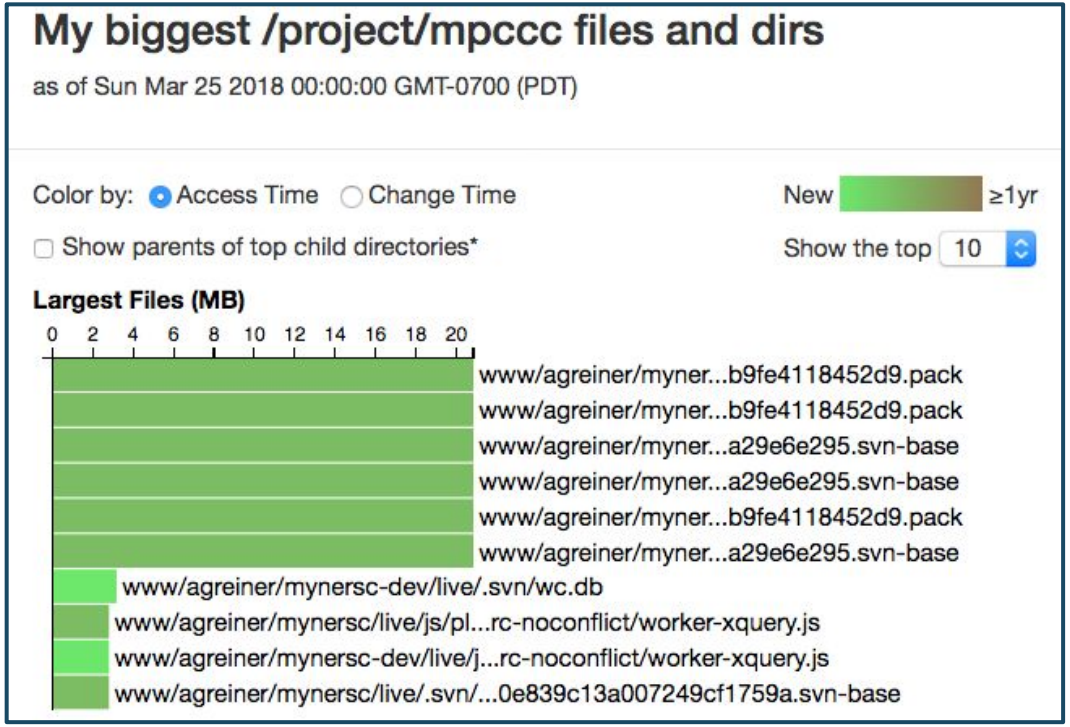
Data Dashboard UI

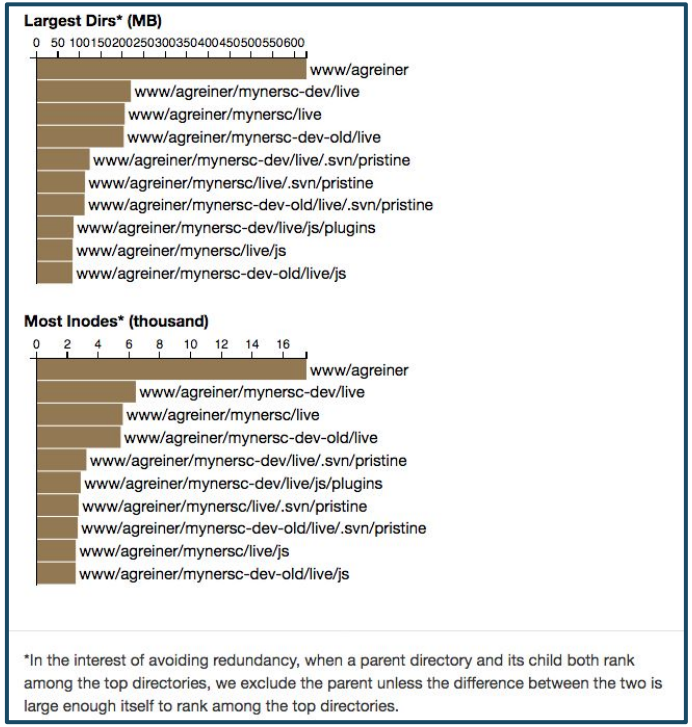




One project expanded to show individual groups' and users' usage

Biggest Files and Dirs



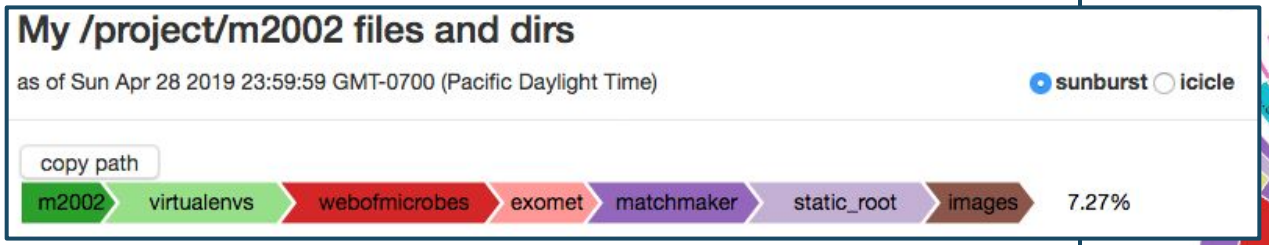


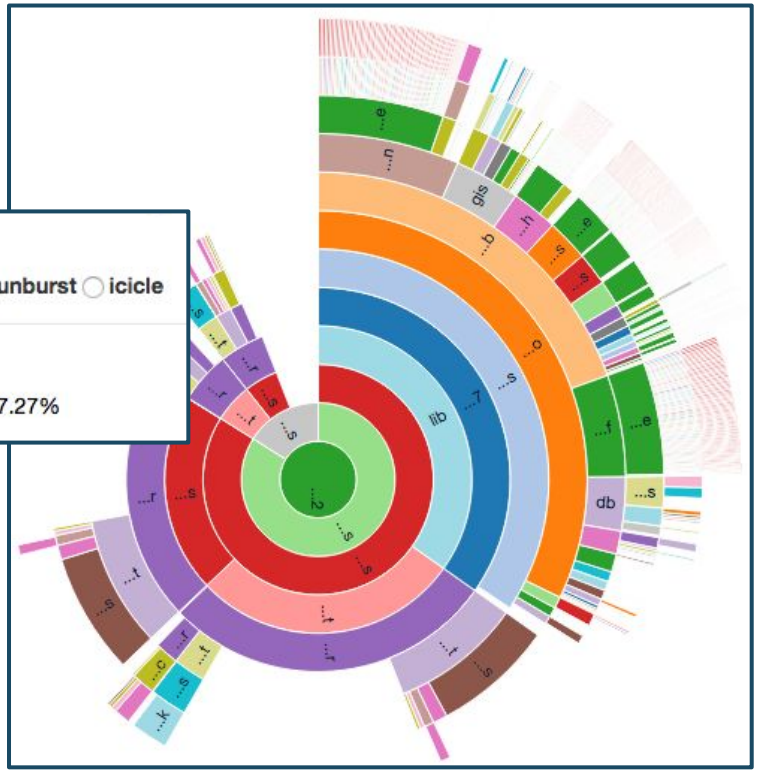
Popup showing one user's biggest files and directories, colored by age

File Browser

Two views for browsing user files and directories

- Sunburst
- Icicle

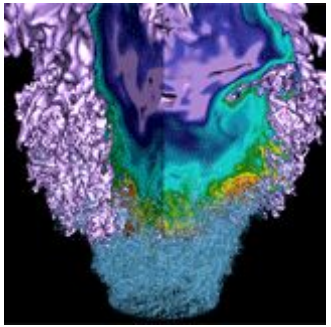


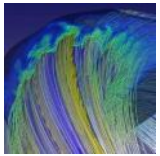


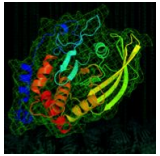
Data Dashboard - Permissions Wrangler

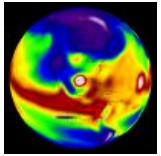
- Designed to address the issue of inadvertent permission drifting
- Deployed for a few large groups at NERSC
- Mines the scans for files and directories that have deviated from group readable permissions
- Generates a list of chmod commands that storage admins automatically pick up and run
  - Very quick, takes O(minutes) to chmod several of million of files
- Will extend to every group at NERSC in the near future (with an opt-in feature on the Data Dashboard)

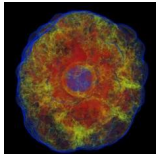
Future Plans

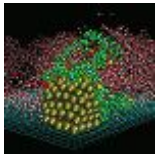












NERSC

Future Plans

- Extend to Lustre and HPSS to present a holistic view of data at NERSC
  - Current scan tools for Lustre do not keep up with the file system churn, very difficult to get timely visibility
  - HPSS not optimized for fast scanning of file contents
- Integrate with other Superfacility efforts to enable data movement across all tiers
  - Data triggered actions
  - Batch system integration
  - Automatic archiving via the Data Dashboard
- Enable predictive estimation in data trends
  - Predictive warnings
  - Estimates for users based on historical data

Conclusion

- We provide a solution to most common issues across groups at NERSC
  - Finding files and understanding data
  - Moving data internally across tiers at the center
  - Triggering workflows
- Addressing the increasing complexity of managing growing data volumes
- Many areas to address, tackled the most common burden, which is centered around our “project” file system
- Still many more exciting features to come!

Authors

National Energy Research Scientific Computing Center (NERSC) at

Lawrence Berkeley National Laboratory:

- Lisa Gerhardt
- Annette Greiner
- Ravi Cheema
- Damian Hazen
- Kristy Kallback-Rose
- Rei Lee
- Kirill Lozinskiy

Thank You