NERSC (storage) experiences - WLCG/HSF 2020

Thanks to Glenn Lockwood, Debbie Bard, Annette Greiner, Bjoern Enders and Lisa Gerhardt for slides

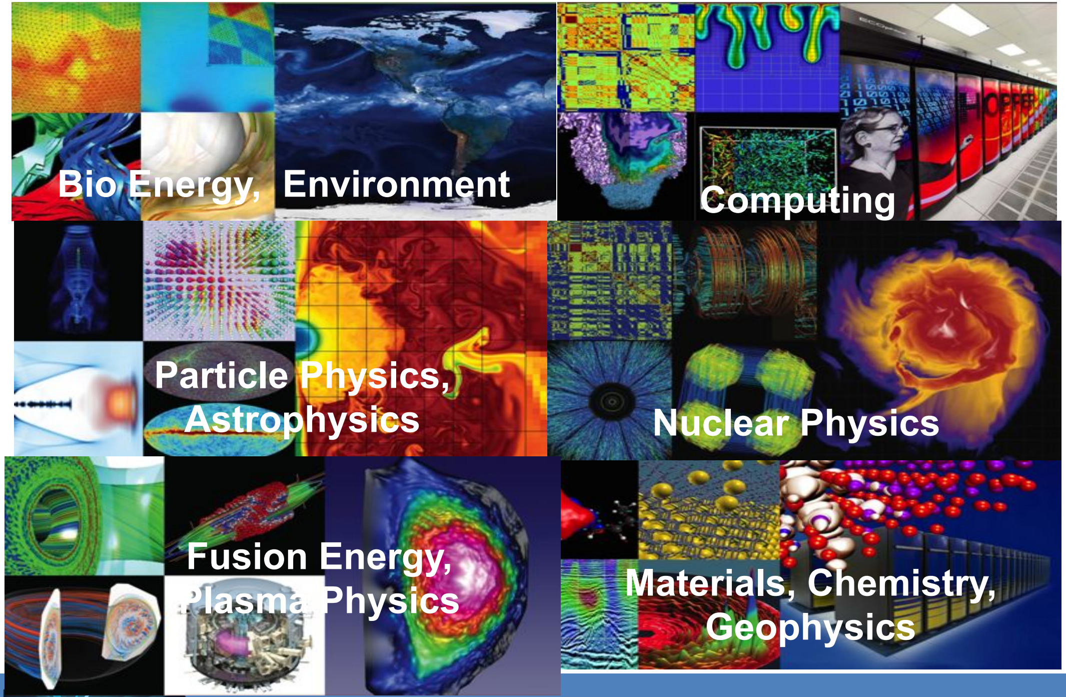
Wahid Bhimji

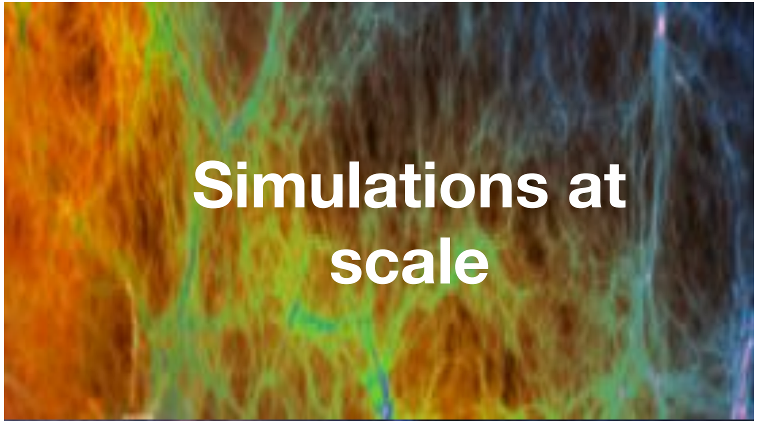
Data and Analytics Group, NERSC, Berkeley Lab

November 24, 2020

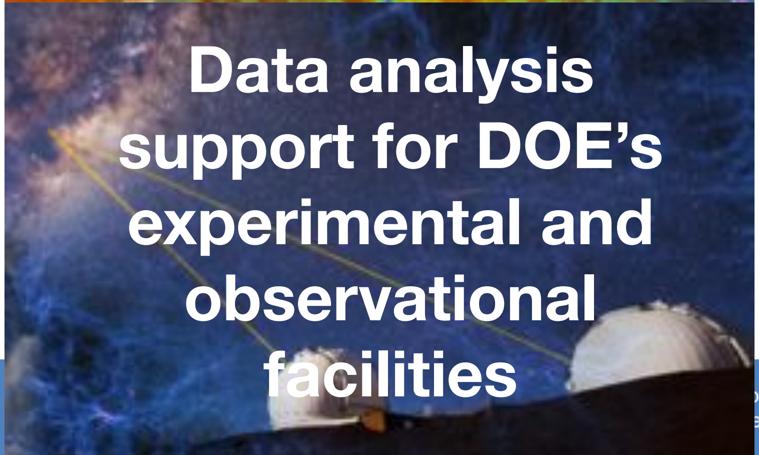
NERSC is the **mission** HPC and Data facility for the Dept Of Energy Office for Science

7,000+ Users, 800+ Projects *2000+ citations /yr*


Bio Energy, Environment
Computing
Particle Physics, Astrophysics
Nuclear Physics
Fusion Energy, Plasma Physics
Materials, Chemistry, Geophysics

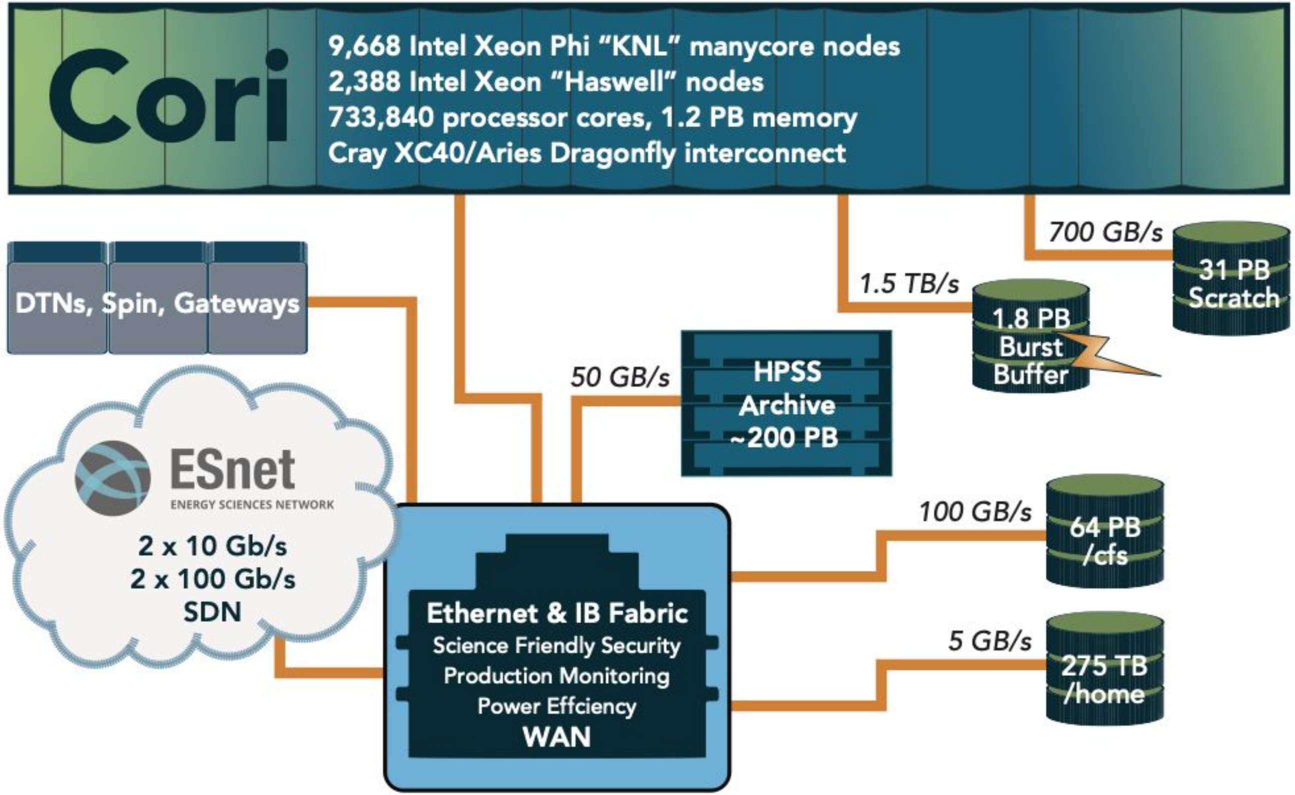


Simulations at scale



Data analysis support for DOE's experimental and observational facilities

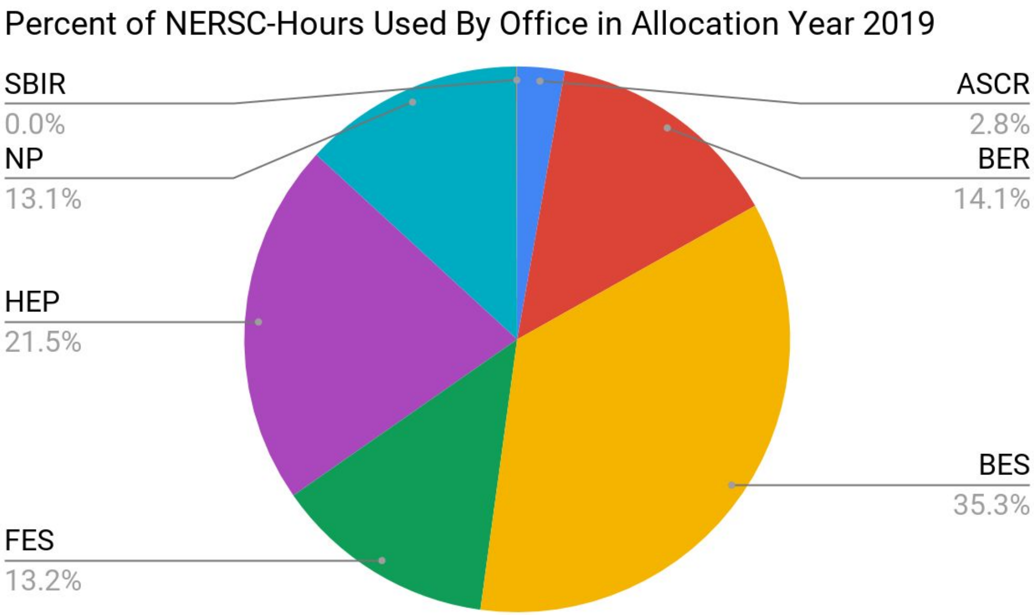
NERSC Centre 2020



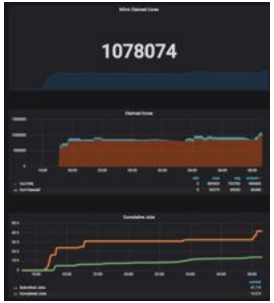
Cori

9,668 Intel Xeon Phi "KNL" manycore nodes
2,388 Intel Xeon "Haswell" nodes
733,840 processor cores, 1.2 PB memory
Cray XC40/Aries Dragonfly interconnect

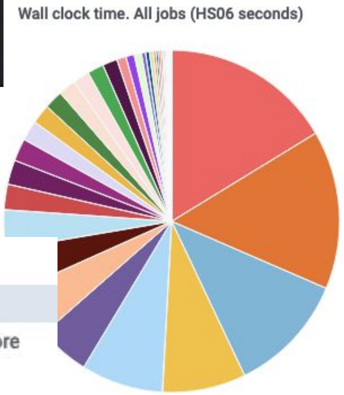
NERSC mostly non-HEP/NP - but valuable HEP resource


Percent of NERSC-Hours Used By Office in Allocation Year 2019

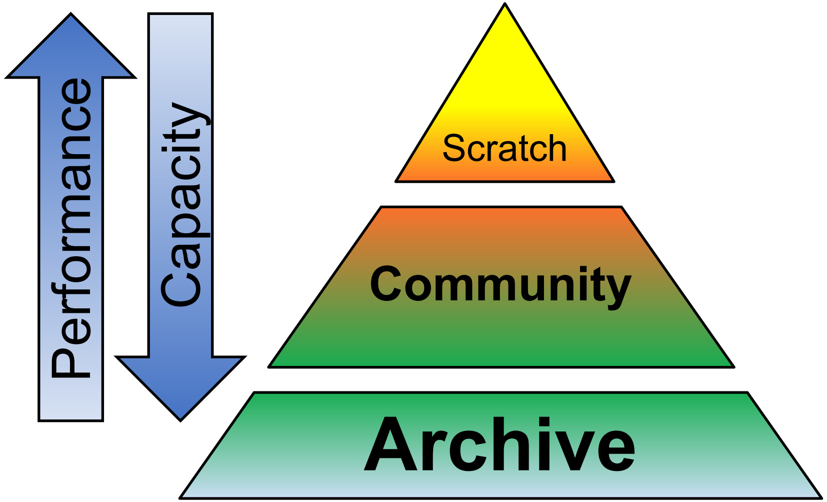
E.g. NoVA '1m cores' news article



US-ATLAS
(2020 to Aug):



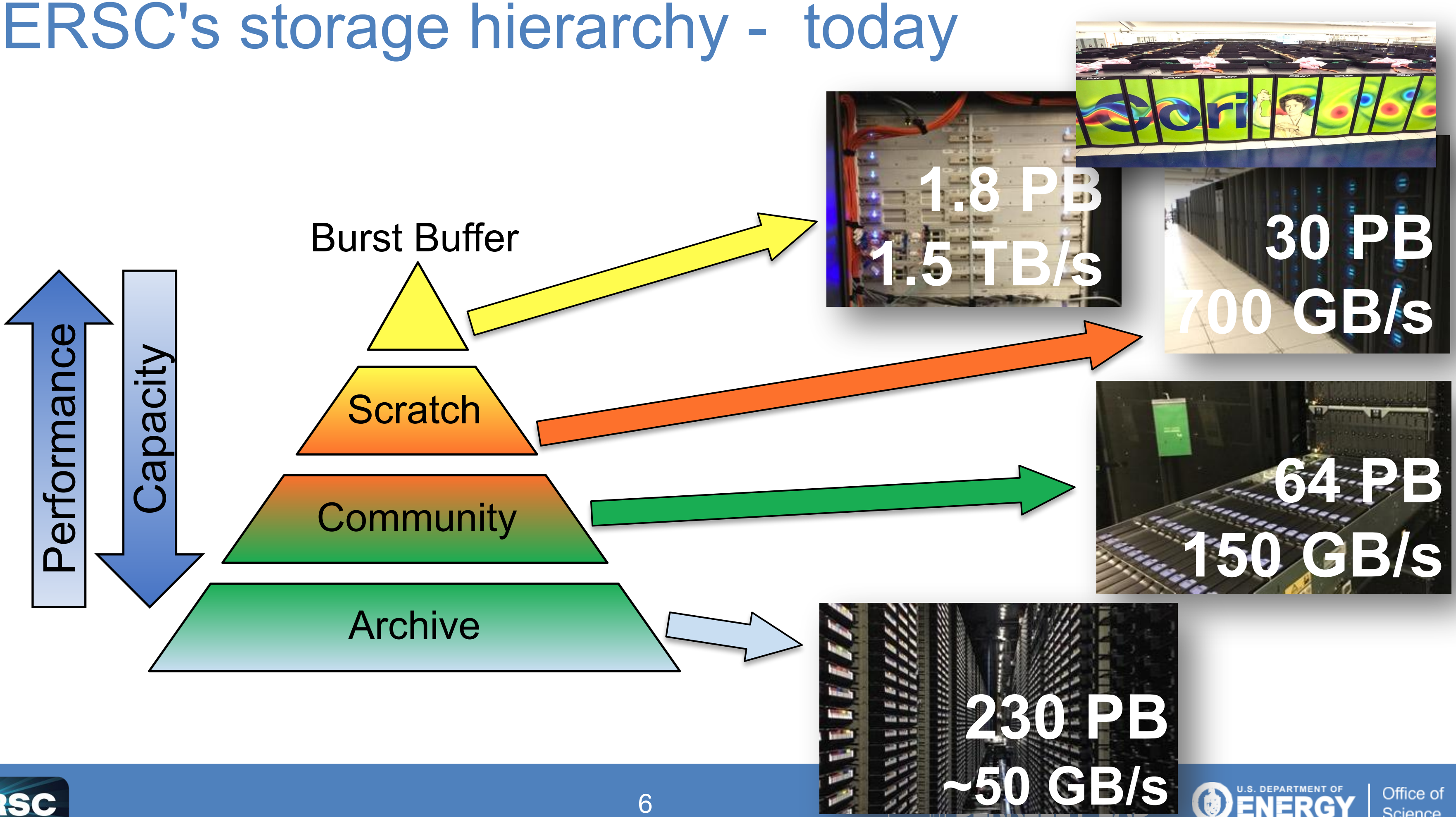
NERSC perspective on storage hierarchy



- **Scratch (weeks – months)**
  - Mounted on only one HPC system
  - User data purged after 4-12 weeks
- **Community (months – years)**
  - Mounted center-wide (HPCs, web, k8s)
  - Quotas
  - User data archived at project end
- **Archive (years – decades)**
  - Not "mounted" anywhere (object-like)
  - No effective quota

More info: G. K. Lockwood *et al.*, “Storage 2020: A Vision for the Future of HPC Storage,” Berkeley, CA, 2017.
Report: https://doi.org/10.2172/1632124
Powerpoint: https://hps.vi4io.org/_media/events/2018/hpc-iodc-18-futurestorage.pdf

NERSC's storage hierarchy - today



Burst Buffer: 1.8 PB, 1.5 TB/s

Scratch: 30 PB, 700 GB/s

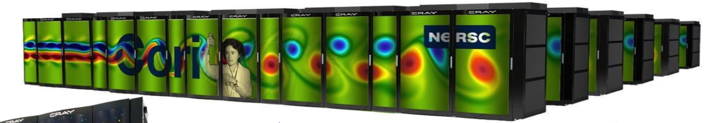
Community: 64 PB, 150 GB/s

Archive: 230 PB, ~50 GB/s

NERSC Systems Roadmap

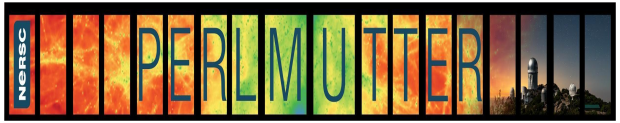
2013

NERSC-7:
Edison
Multicore
CPU



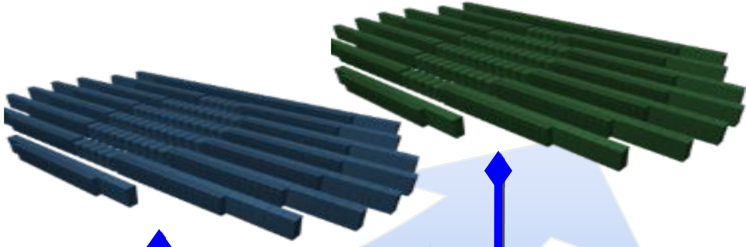
2016

NERSC-8: Cori
Manycore CPU
NESAP Launched:
transition applications to
advanced architectures



2020

NERSC-9: Perlmutter
CPU and GPU nodes
Continued transition of
applications and support for
complex workflows



2024

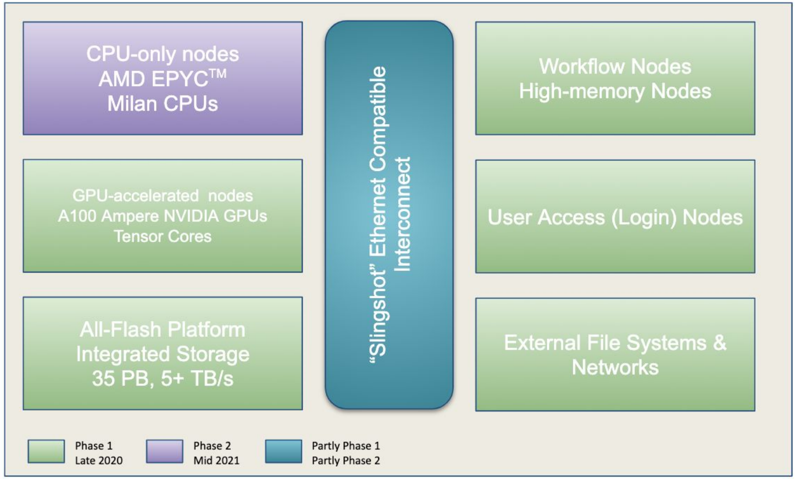
NERSC-10:
Exa system

2028

NERSC-11:
Beyond
Moore

Perlmutter

- Cray Shasta System providing 3-4x capability of Cori
- GPU-accelerated and CPU-only nodes
  - Large CPU-only partition providing capability similar to Cori
  - Services for complex workflows
  - Optimized data software enabling analytics and ML at scale
- GPU nodes: 4 NVIDIA A100 “Ampere” GPUs each w/Tensor Cores, NVLink-3 and High-BW memory + 1 AMD “Milan” CPU
  - Over 6000 GPUs
  - Unified Virtual Memory support improves programmability
- Cray “Slingshot” - High-performance, scalable, low-latency Ethernet- compatible network
  - Capable of Terabit connections to/from the system
- Single-tier All-Flash Lustre based HPC file system
  - 6x Cori’s bandwidth
  - Cray’s ClusterStor E1000 system



CPU-only nodes
AMD EPYC™
Milan CPUs

GPU-accelerated nodes
A100 Ampere NVIDIA GPUs
Tensor Cores

All-Flash Platform
Integrated Storage
35 PB, 5+ TB/s

“Slingshot” Ethernet Compatible Interconnect

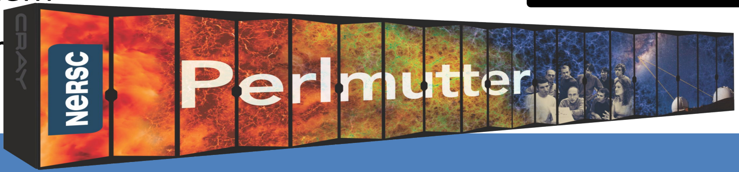
Workflow Nodes
High-memory Nodes

User Access (Login) Nodes

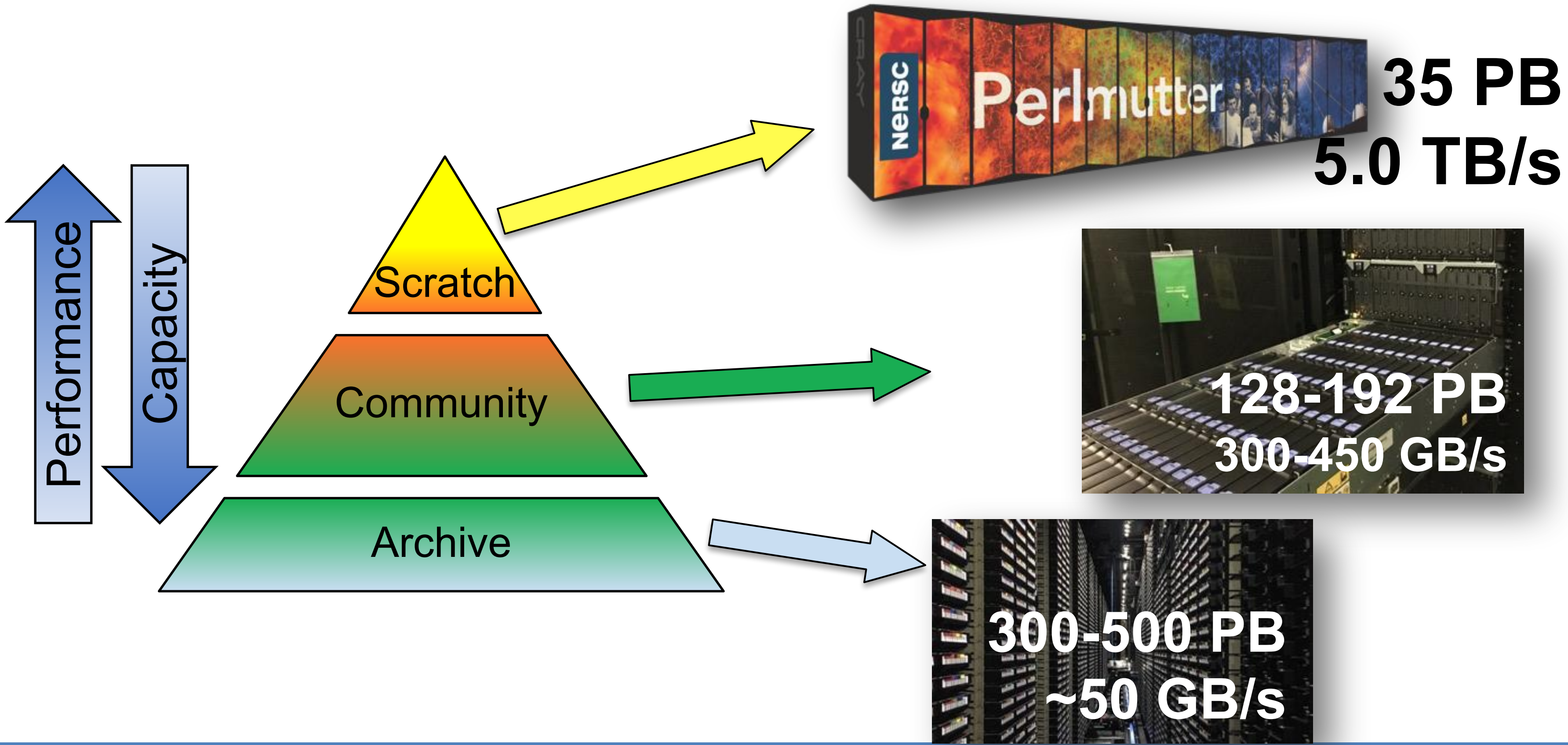
External File Systems & Networks

🟩 Phase 1 Late 2020
🟪 Phase 2 Mid 2021
🟦 Partly Phase 1 Partly Phase 2

Phased delivery
1st phase: Early 2021
2nd phase: Summer 2021



NERSC's storage hierarchy - soon

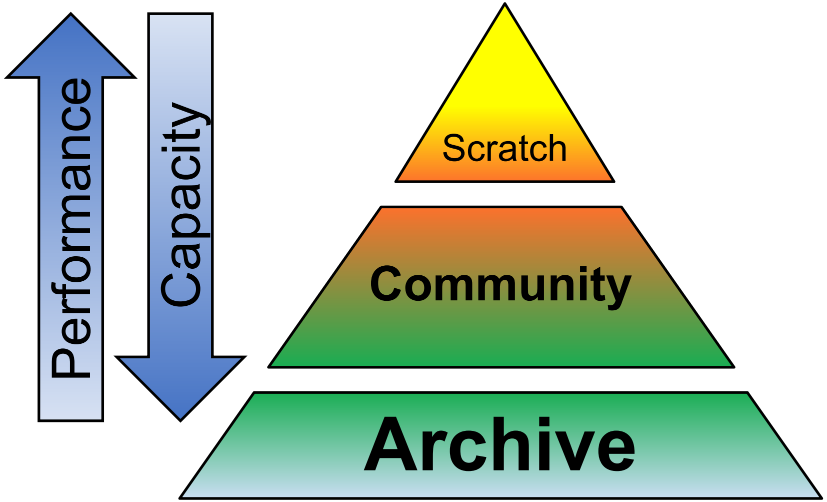


35 PB
5.0 TB/s

128-192 PB
300-450 GB/s

300-500 PB
~50 GB/s

Data Management at NERSC

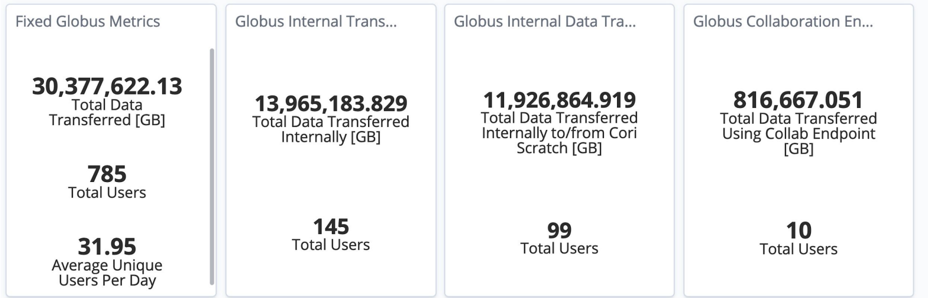


Community File System: Intended for sharing scientific data between groups of NERSC users. 27 PB and 3 billion inodes used

Very useful, but has some issues

- Group quotas that fill up
- Permissions drift
- Data migration between the tiers

30% of Globus data movement is purely internal



Fixed Globus Metrics

30,377,622.13
Total Data Transferred [GB]

785
Total Users

31.95
Average Unique Users Per Day

Globus Internal Trans...

13,965,183.829
Total Data Transferred Internally [GB]

145
Total Users

Globus Internal Data Tra...

11,926,864.919
Total Data Transferred Internally to/from Cori Scratch [GB]

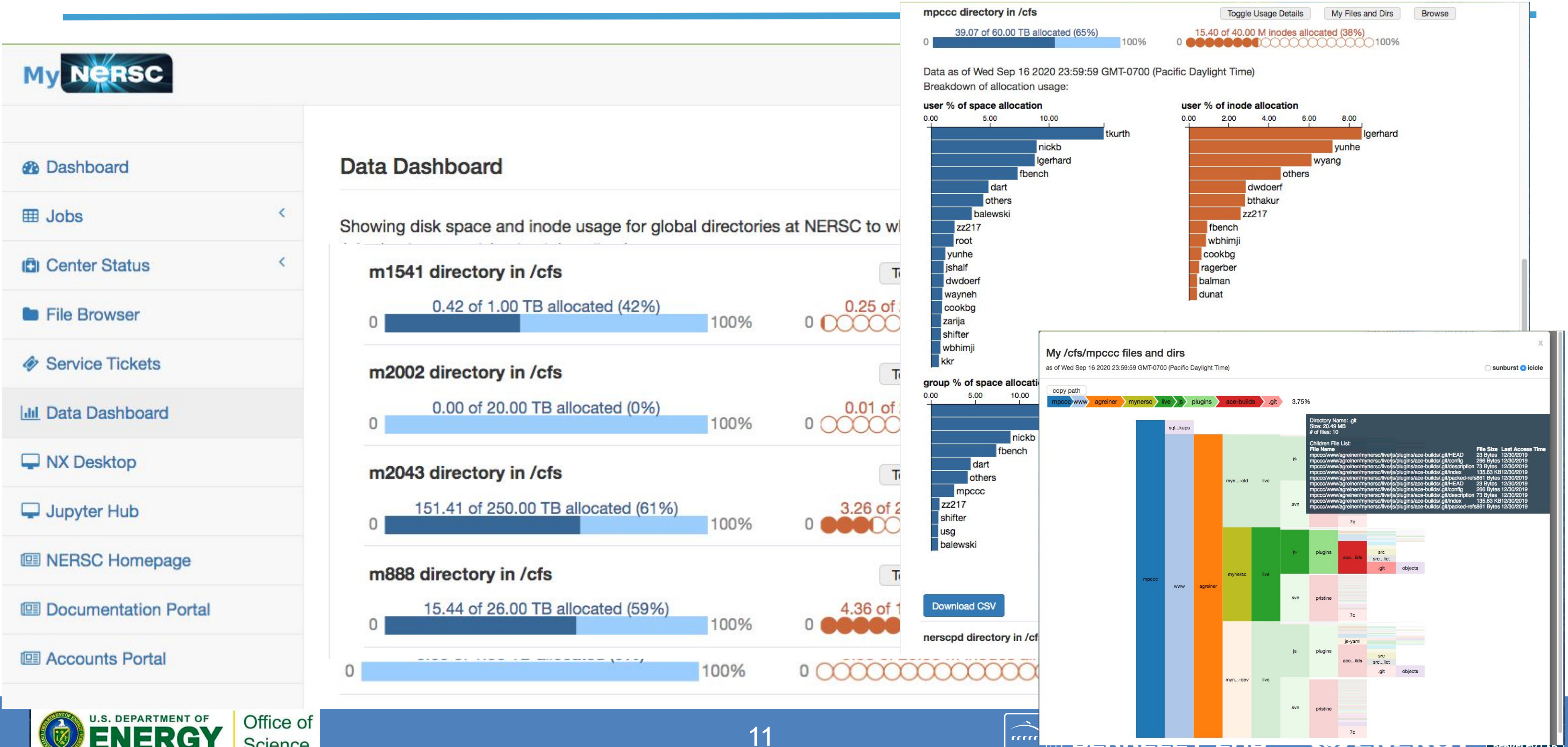
99
Total Users

Globus Collaboration En...

816,667.051
Total Data Transferred Using Collab Endpoint [GB]

10
Total Users

Data Dashboard: Helping Users Share Their Space



Data Dashboard

Showing disk space and inode usage for global directories at NERSC to wl

m1541 directory in /cfs

0 0.42 of 1.00 TB allocated (42%) 100%   0 0.25 of

m2002 directory in /cfs

0 0.00 of 20.00 TB allocated (0%) 100%   0 0.01 of

m2043 directory in /cfs

0 151.41 of 250.00 TB allocated (61%) 100%   0 3.26 of 2

m888 directory in /cfs

0 15.44 of 26.00 TB allocated (59%) 100%   0 4.36 of 1

0 100%   0

mpccc directory in /cfs

0 39.07 of 60.00 TB allocated (65%) 100%   0 15.40 of 40.00 M inodes allocated (38%) 100%

Data as of Wed Sep 16 2020 23:59:59 GMT-0700 (Pacific Daylight Time)
Breakdown of allocation usage:

nerscpd directory in /cf

My /cfs/mpccc files and dirs

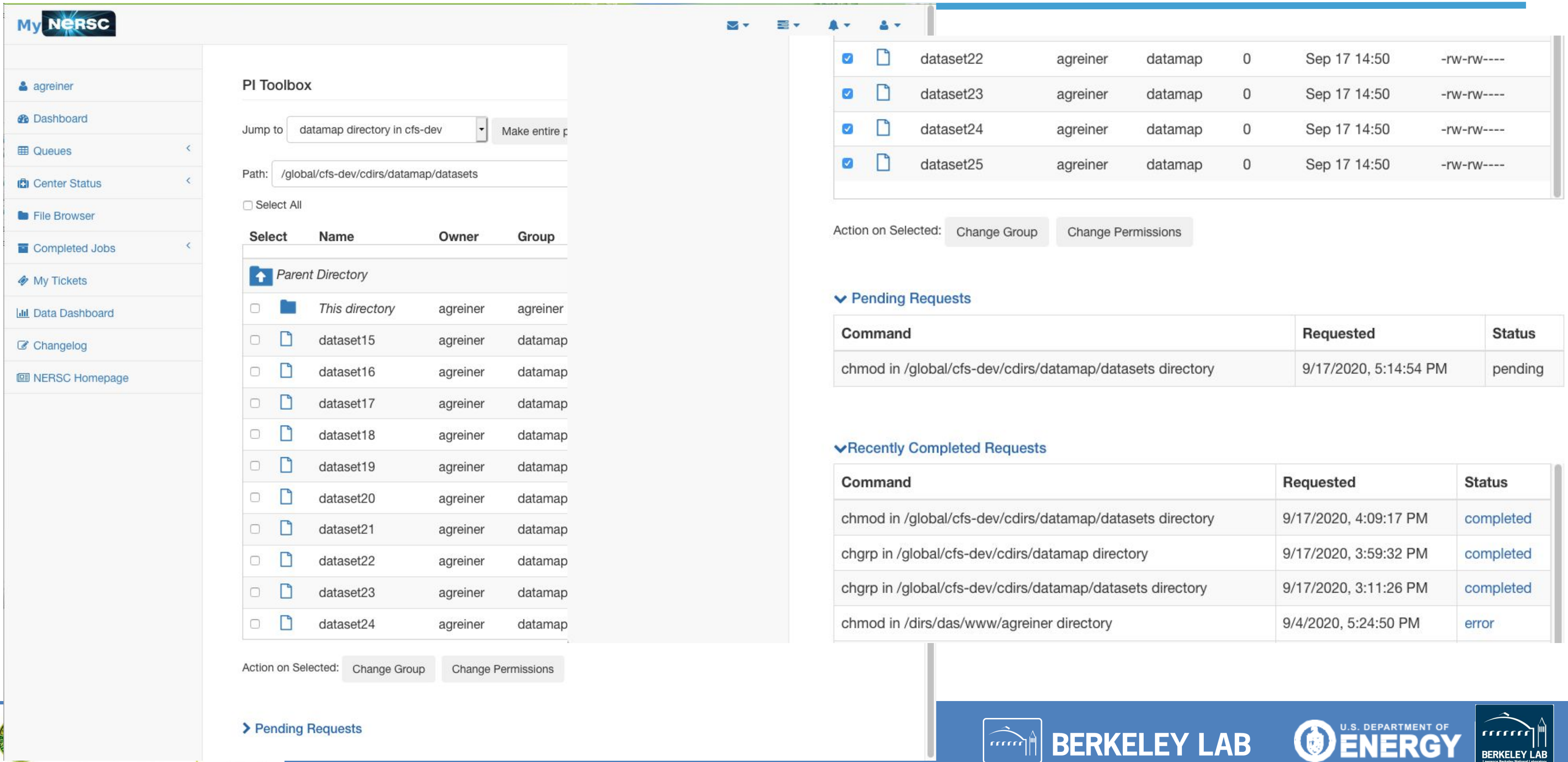
as of Wed Sep 16 2020 23:59:59 GMT-0700 (Pacific Daylight Time)

Directory Name: .git
Size: 20.49 MB
# of files: 10

Children File List:

```
File Name                                                          File Size   Last Access Time
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/HEAD        23 Bytes   12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/config      266 Bytes  12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/description 73 Bytes   12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/index       135.63 KB12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/packed-refs861 Bytes  12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/HEAD        23 Bytes   12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/config      266 Bytes  12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/description 73 Bytes   12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/index       135.63 KB12/30/2019
mpccc/www/agreiner/mynersc/live/js/plugins/ace-builds/.git/packed-refs861 Bytes  12/30/2019
```

PI Toolbox: Permission Wrangling



PI Toolbox

Jump to: datamap directory in cfs-dev | Make entire p

Path: /global/cfs-dev/cdirs/datamap/datasets

□ Select All

Select	Name	Owner	Group
	Parent Directory		
□	This directory	agreiner	agreiner
□	dataset15	agreiner	datamap
□	dataset16	agreiner	datamap
□	dataset17	agreiner	datamap
□	dataset18	agreiner	datamap
□	dataset19	agreiner	datamap
□	dataset20	agreiner	datamap
□	dataset21	agreiner	datamap
□	dataset22	agreiner	datamap
□	dataset23	agreiner	datamap
□	dataset24	agreiner	datamap

Action on Selected: Change Group | Change Permissions

Pending Requests

☑	dataset22	agreiner	datamap	0	Sep 17 14:50	-rw-rw----
☑	dataset23	agreiner	datamap	0	Sep 17 14:50	-rw-rw----
☑	dataset24	agreiner	datamap	0	Sep 17 14:50	-rw-rw----
☑	dataset25	agreiner	datamap	0	Sep 17 14:50	-rw-rw----

Action on Selected: Change Group | Change Permissions

Pending Requests

Command	Requested	Status
chmod in /global/cfs-dev/cdirs/datamap/datasets directory	9/17/2020, 5:14:54 PM	pending

Recently Completed Requests

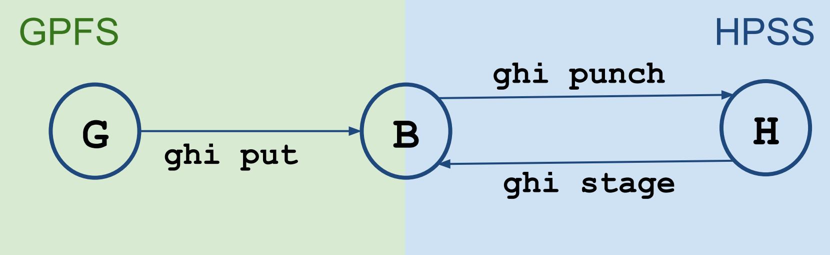
Command	Requested	Status
chmod in /global/cfs-dev/cdirs/datamap/datasets directory	9/17/2020, 4:09:17 PM	completed
chgrp in /global/cfs-dev/cdirs/datamap directory	9/17/2020, 3:59:32 PM	completed
chgrp in /global/cfs-dev/cdirs/datamap/datasets directory	9/17/2020, 3:11:26 PM	completed
chmod in /dirs/das/www/agreiner directory	9/4/2020, 5:24:50 PM	error

GPFS-HPSS Interface

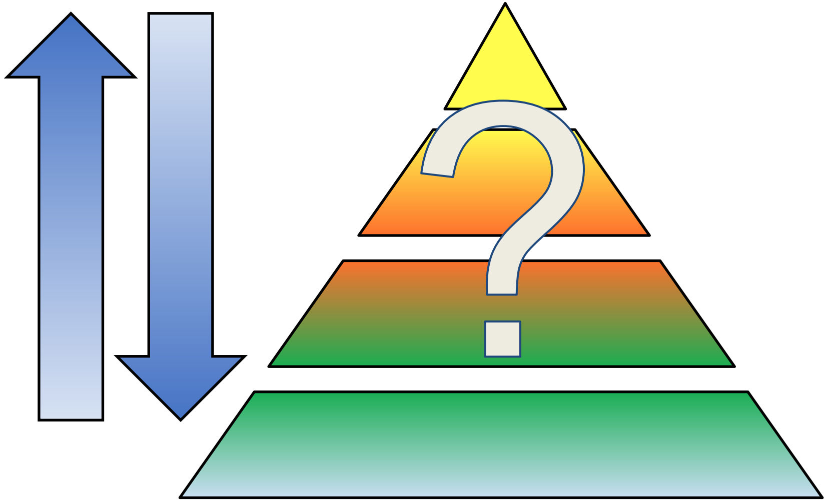
Use HPSS using the familiar file system interface

`ghi ls`	show what file system the files are currently on, files are marked ’G’ for GPFS, ’H’ for HPSS and ’B’ for both
`ghi put`	copy the files to HPSS, makes files dual residents on both files systems
`ghi stage`	move the files back from HPSS to GPFS
`ghi punch`	move the files to HPSS (leaving a stub behind on GPFS)
`ghi pin`	keep the files from being removed from GPFS
`ghi lock`	keep the files from being removed using rm or otherwise modified

- Users interact with a file system directory that’s tied to HPSS behind the scenes
- GHI puts the files optimally into HPSS on your behalf, no need to htar small files together or break things into 500GB sized chunks
- Shared namespace, so deleting a file from GPFS will remove it from HPSS

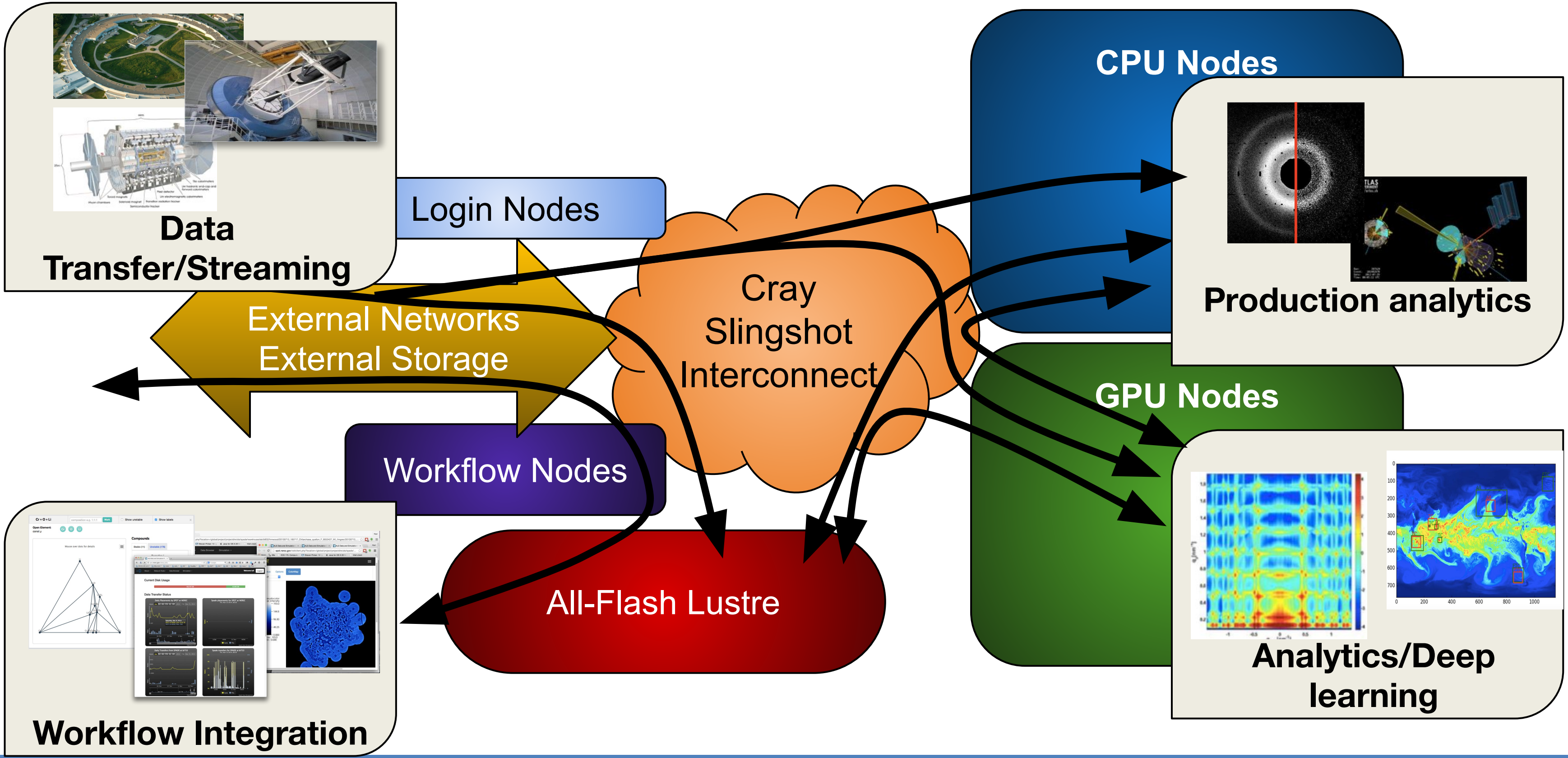


NERSC's storage future

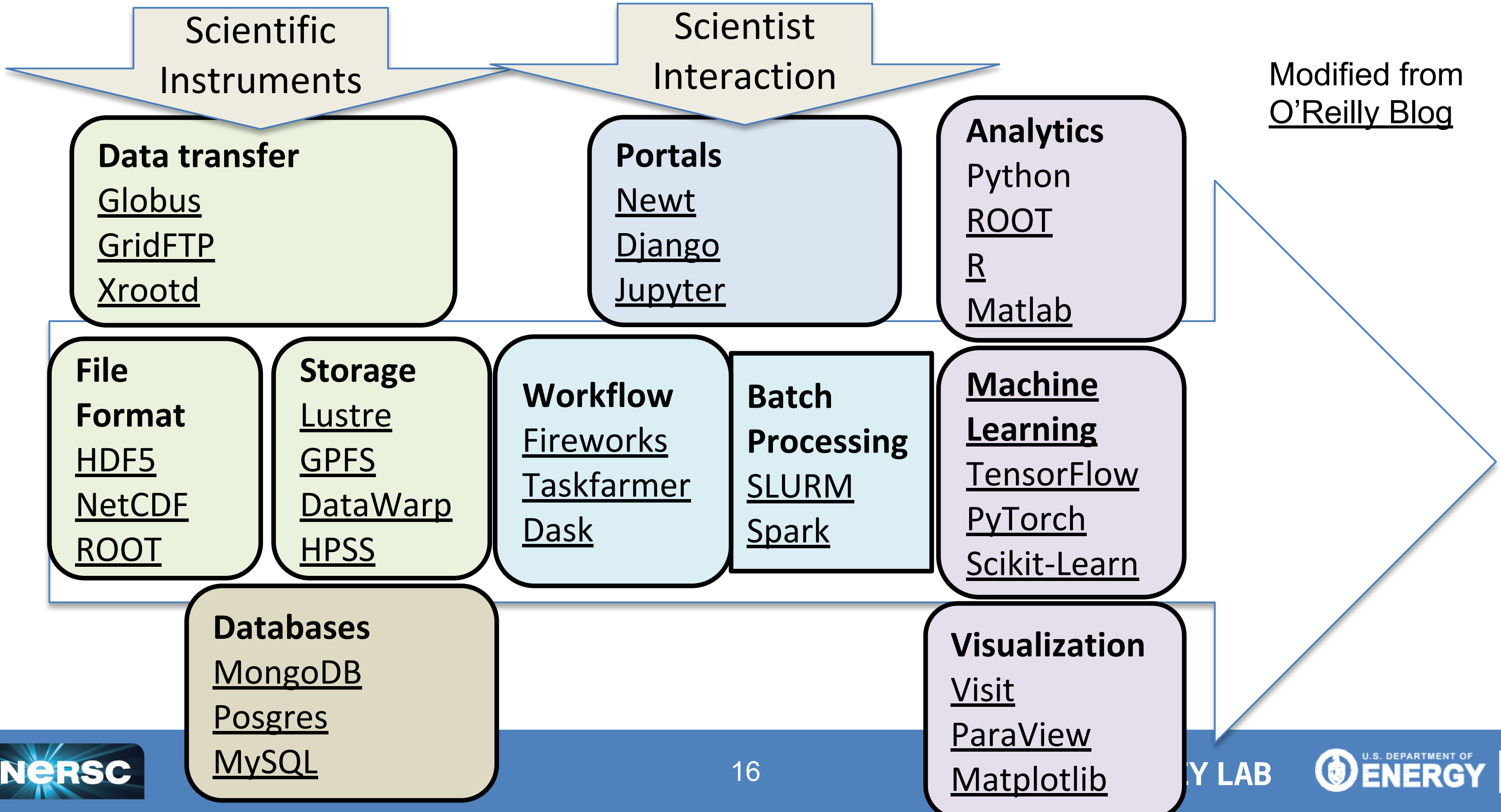


- Various demands on storage e.g.
  - Write-heavy high-bw HPC
  - Read-heavy high-iops Deep learning and analytics
  - Metadata management
  - Workflow chaining
- Various storage innovations e.g.
  - Object-store access
  - Filesystems on demand
  - Very high IOPS NVRam
- Challenges include
  - Stability and reliability
  - Multi-user HPC integration

Wider NERSC infrastructure



Wider NERSC data services



Scientific Instruments

Scientist Interaction

Data transfer
Globus
GridFTP
Xrootd

Portals
Newt
Django
Jupyter

Analytics
Python
ROOT
R
Matlab

File Format
HDF5
NetCDF
ROOT

Storage
Lustre
GPFS
DataWarp
HPSS

Workflow
Fireworks
Taskfarmer
Dask

Batch Processing
SLURM
Spark

Machine Learning
TensorFlow
PyTorch
Scikit-Learn

Databases
MongoDB
Posgres
MySQL

Visualization
Visit
ParaView
Matplotlib

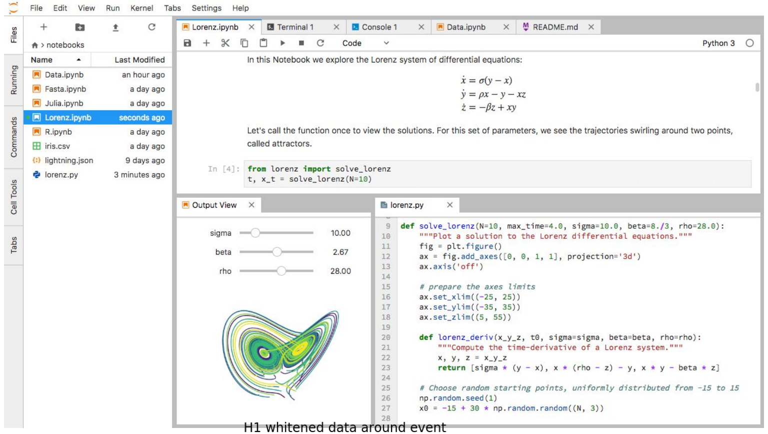
Modified from O'Reilly Blog

Jupyter Notebooks for HPC

- **Jupyter growing in popularity at NERSC and broader community**
  - **> 700 unique users @NERSC**
- **NERSC's goal - Enable exploratory data analytics, deep learning, workflows, through Jupyter on HPC**
- **New features in 2019-2020:**
  - **Access to Cori compute nodes**
  - **Access to Cori GPUs**

Go to https://jupyter.nersc.gov

[New User Training talk]()



Name	Last Modified
Data.ipynb	an hour ago
Fasta.ipynb	a day ago
Julia.ipynb	a day ago
Lorenz.ipynb	seconds ago
R.ipynb	a day ago
iris.csv	a day ago
lightning.json	9 days ago
lorenz.py	3 minutes ago

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

Let's call the function once to view the solutions. For this set of parameters, we see the trajectories swirling around two points, called attractors.

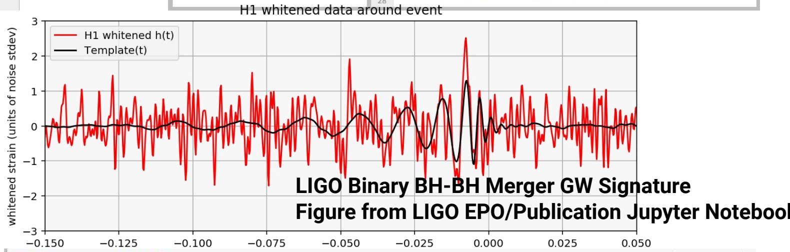
```
In [4]: from lorenz import solve_lorenz
        t, x_t = solve_lorenz(N=10)
```

sigma 10.00

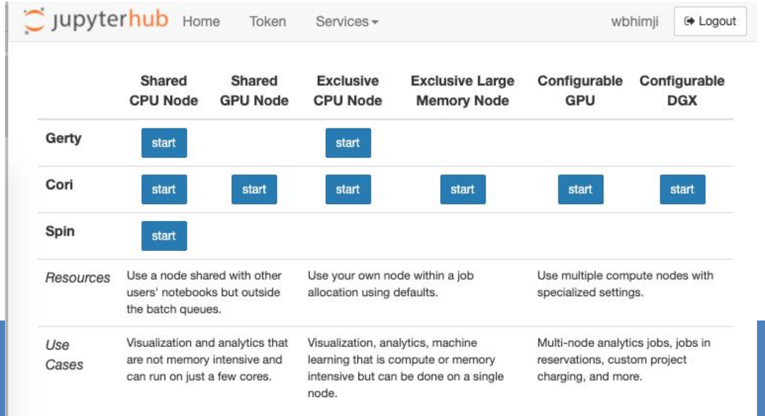
beta 2.67

rho 28.00

```
def solve_lorenz(N=10, max_time=4.0, sigma=10.0, beta=8./3, rho=28.0):
    """Plot a solution to the Lorenz differential equations."""
    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')

    # prepare the axes limits
    ax.set_xlim((-25, 25))
    ax.set_ylim((-35, 35))
    ax.set_zlim((5, 55))

    def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
        """Compute the time-derivative of a Lorenz system."""
        x, y, z = x_y_z
        return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]

    # Choose random starting points, uniformly distributed from -15 to 15
    np.random.seed(1)
    x0 = -15 + 30 * np.random.random((N, 3))

```



LIGO Binary BH-BH Merger GW Signature
Figure from LIGO EPO/Publication Jupyter Noteboo



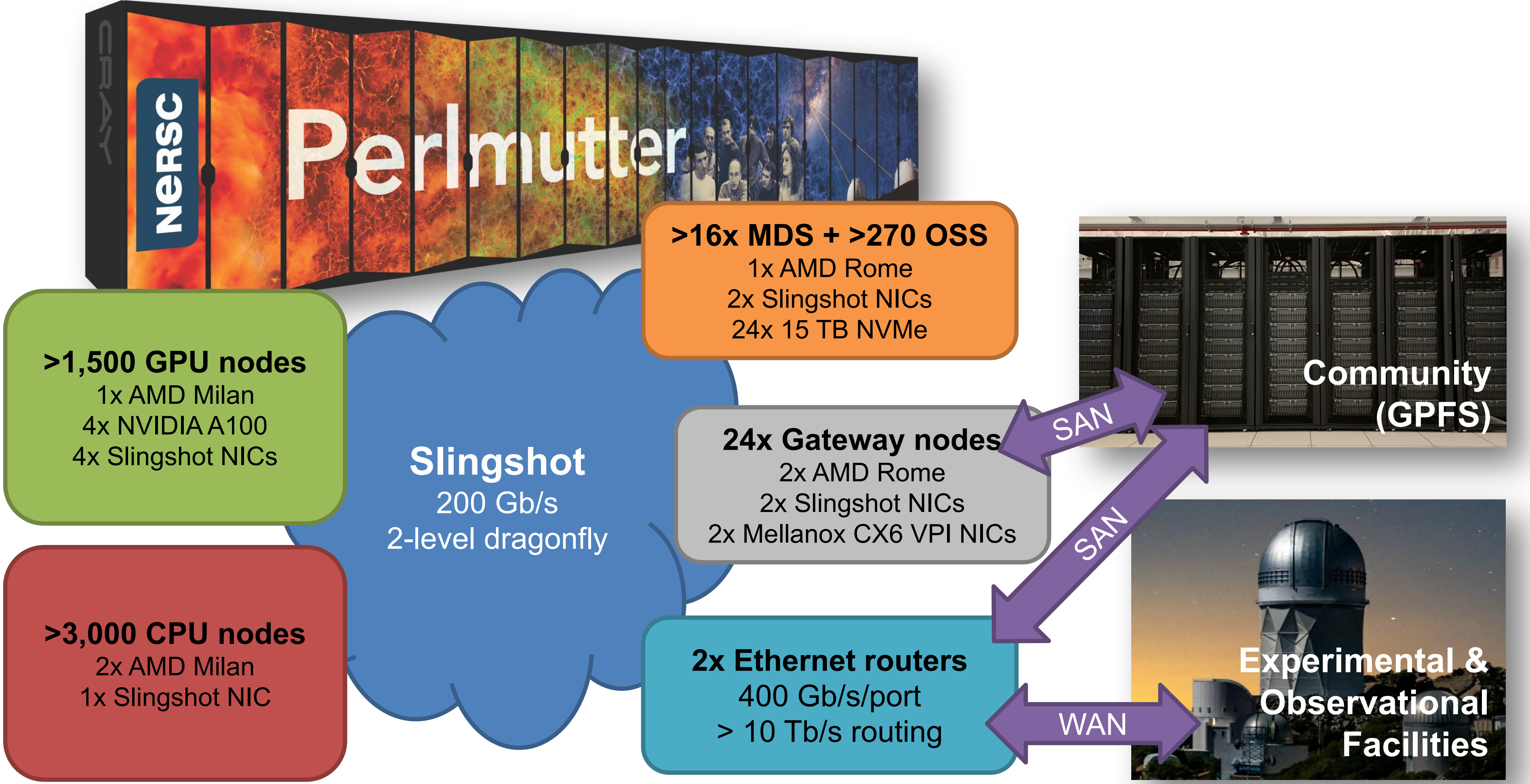
	Shared CPU Node	Shared GPU Node	Exclusive CPU Node	Exclusive Large Memory Node	Configurable GPU	Configurable DGX
Gerty	start		start			
Cori	start	start	start	start	start	start
Spin	start					
Resources	Use a node shared with other users' notebooks but outside the batch queues.		Use your own node within a job allocation using defaults.		Use multiple compute nodes with specialized settings.	
Use Cases	Visualization and analytics that are not memory intensive and can run on just a few cores.		Visualization, analytics, machine learning that is compute or memory intensive but can be done on a single node.		Multi-node analytics jobs, jobs in reservations, custom project charging, and more.	

Conclusions

- NERSC supports an increasing number of data-rich projects across difference science domains

- Upcoming upgrades to center systems and storage
  - All-flash scratch tier with Perlmutter
  - Large capacity “community” filesystem

- Deploying tools to ease data management within the center and beyond



>1,500 GPU nodes
1x AMD Milan
4x NVIDIA A100
4x Slingshot NICs

>3,000 CPU nodes
2x AMD Milan
1x Slingshot NIC

Slingshot
200 Gb/s
2-level dragonfly

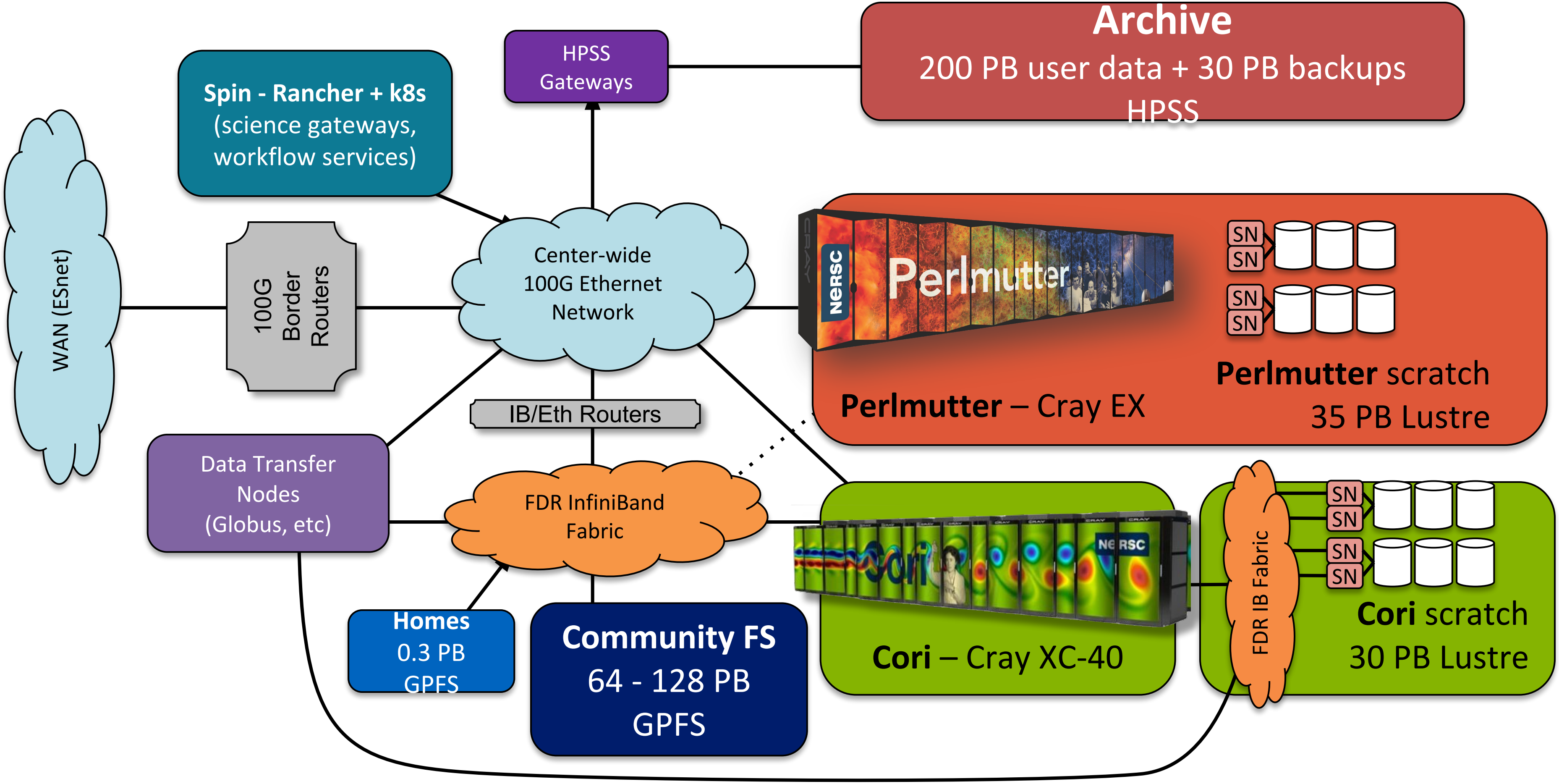
>16x MDS + >270 OSS
1x AMD Rome
2x Slingshot NICs
24x 15 TB NVMe

24x Gateway nodes
2x AMD Rome
2x Slingshot NICs
2x Mellanox CX6 VPI NICs

2x Ethernet routers
400 Gb/s/port
> 10 Tb/s routing

Community (GPFS)

Experimental & Observational Facilities



Archive
200 PB user data + 30 PB backups
HPSS

HPSS Gateways

Spin - Rancher + k8s
(science gateways, workflow services)

WAN (ESnet)

100G Border Routers

Center-wide 100G Ethernet Network

IB/Eth Routers

Data Transfer Nodes (Globus, etc)

FDR InfiniBand Fabric

Homes
0.3 PB
GPFS

Community FS
64 - 128 PB
GPFS

Perlmutter – Cray EX

Perlmutter scratch
35 PB Lustre

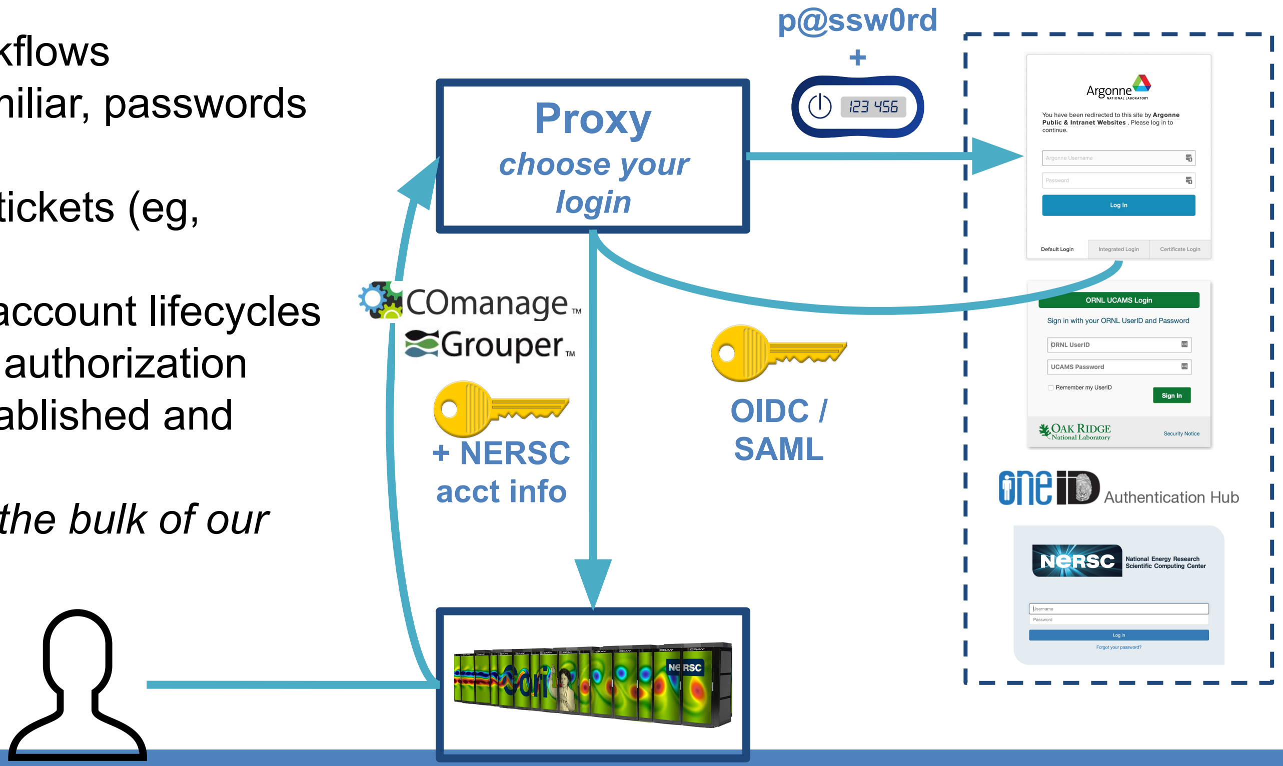
Cori – Cray XC-40

FDR IB Fabric

Cori scratch
30 PB Lustre

Federated Identity (FedID) allows a person to use a single digital identity across multiple organizations

- Simplifies cross-facility workflows
- Users have fewer, more familiar, passwords and login pages
- NERSC has fewer support tickets (eg, password resets)
- Home institution manages account lifecycles
- NERSC still manages local authorization
- Core technology is well-established and mature
- *Policy/trust decisions were the bulk of our analysis*



Argonne NATIONAL LABORATORY

You have been redirected to this site by **Argonne Public & Intranet Websites** . Please log in to continue.

Argonne Username

Password

Log In

Default Login | Integrated Login | Certificate Login

ORNL UCAMS Login

Sign in with your ORNL UserID and Password

ORNL UserID

UCAMS Password

□ Remember my UserID

Sign In

OAK RIDGE National Laboratory

Security Notice

NERSC National Energy Research Scientific Computing Center

Username

Password

Log in

Forgot your password?

Spin: Container Services for Science

Spin

Many projects need more than HPC.

Spin is a platform for services.

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

- *Access HPC file systems and networks*
- *Use public or custom software images*
- *Orchestrate complex workflows*
- *Secure, scalable, and managed*

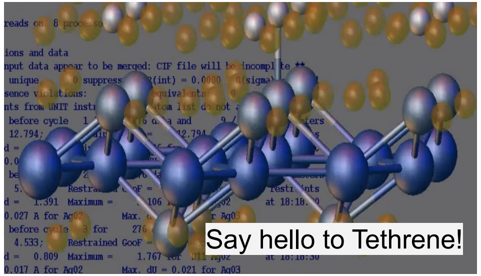
docker RANCHER kubernetes

Some projects using Spin:

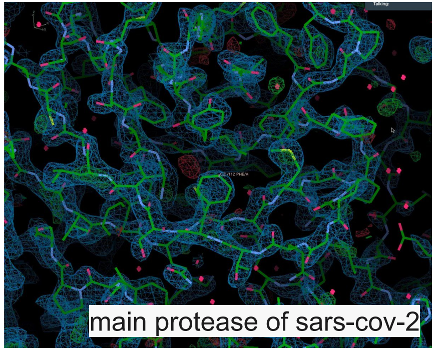
DARK ENERGY SPECTROSCOPIC INSTRUMENT	Track and compare analyses of nightly sky surveys	science gateway
ESS-DIVE	Classify and store reusable earth sciences data	data repository
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY	Manage production genomic workflows and data at scale	science gateway
LZ	Process real-time events for dark matter detection	workflow manager
The Materials Project materialsproject.org	Explore materials properties or build simulated materials	science gateway

LCLS-II is using NERSC for real-time data analysis

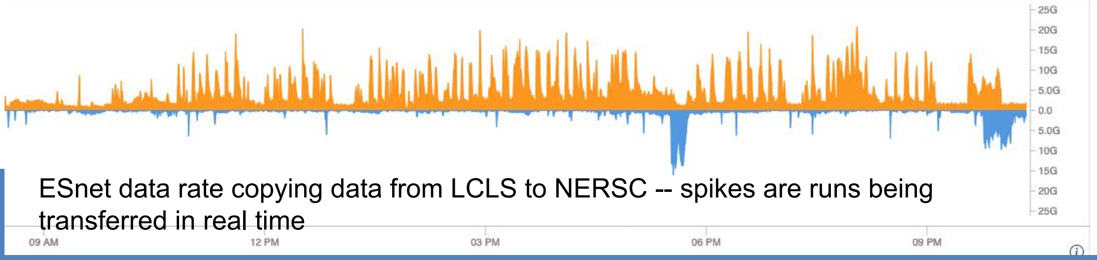
- Several experiments at the LCLS-II (x-ray free electron laser at SLAC) are now using NERSC for real-time data analysis for materials science and Covid-19 research
- Can analyze a 5 minute experiment in ~3 minutes for feedback to beamline staff, transferring 15TB/day to NERSC
  - **Real-time** data analysis using real-time queue and advanced reservations
  - Used services running on **Spin** to orchestrated jobs/parameters/results in real time between several concurrent remote users



Say hello to Tethrene!

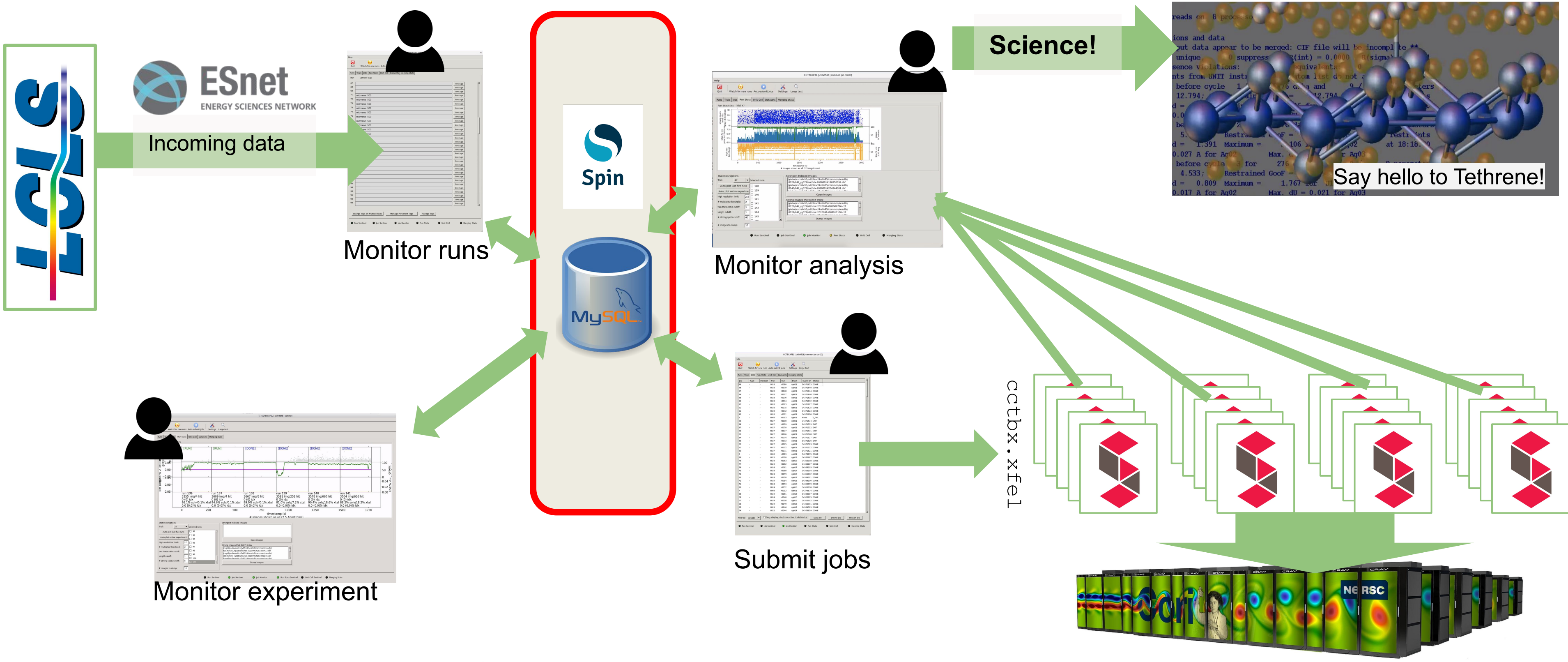


main protease of sars-cov-2



ESnet data rate copying data from LCLS to NERSC -- spikes are runs being transferred in real time

Collaborative Distributed Data Analysis with Spin



Incoming data

Monitor runs

Monitor analysis

Science!

Say hello to Tethrene!

Submit jobs

cctbx.xfel

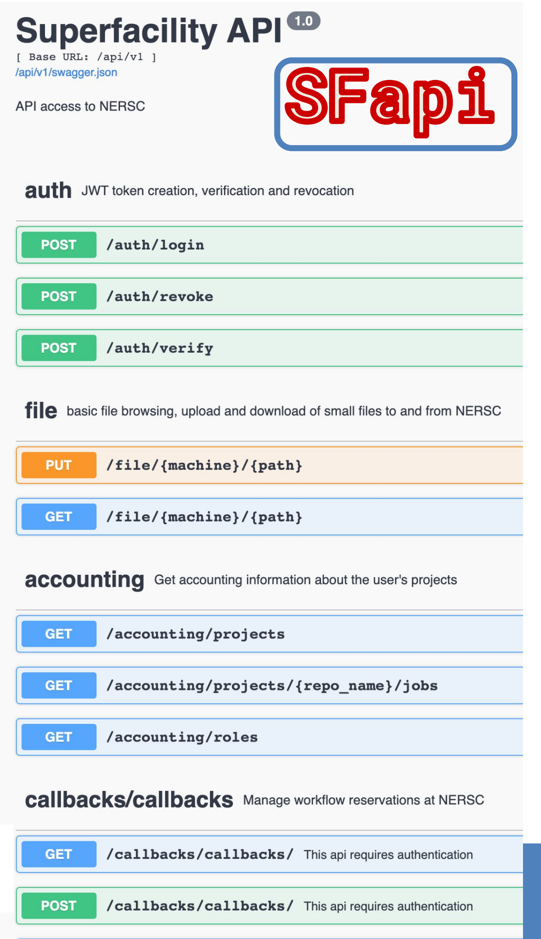
Monitor experiment

Machine-readable supercomputers: the Superfacility API

Vision: all NERSC interactions are callable; backend tools assist large or complex operations.

Endpoints currently prototyped:

`/accounting`	retrieve allocation info for a user or project
`/auth`	obtain OAuth2 authentication tokens (JWTs)
`/callbacks`	register callbacks for asynchronous/chained operations
`/file`	browse, upload, and download files
`/health`	retrieve system health status
`/jobs`	submit jobs and check job status
`/transfer`	move data with Globus or between NERSC storage tiers
`/reservations`	submit and manage future compute reservations



Superfacility API 1.0

[Base URL: /api/v1]

/api/v1/swagger.json

API access to NERSC

auth JWT token creation, verification and revocation

POST `/auth/login`

POST `/auth/revoke`

POST `/auth/verify`

file basic file browsing, upload and download of small files to and from NERSC

PUT `/file/{machine}/{path}`

GET `/file/{machine}/{path}`

accounting Get accounting information about the user's projects

GET `/accounting/projects`

GET `/accounting/projects/{repo_name}/jobs`

GET `/accounting/roles`

callbacks/callbacks Manage workflow reservations at NERSC

GET `/callbacks/callbacks/` This api requires authentication

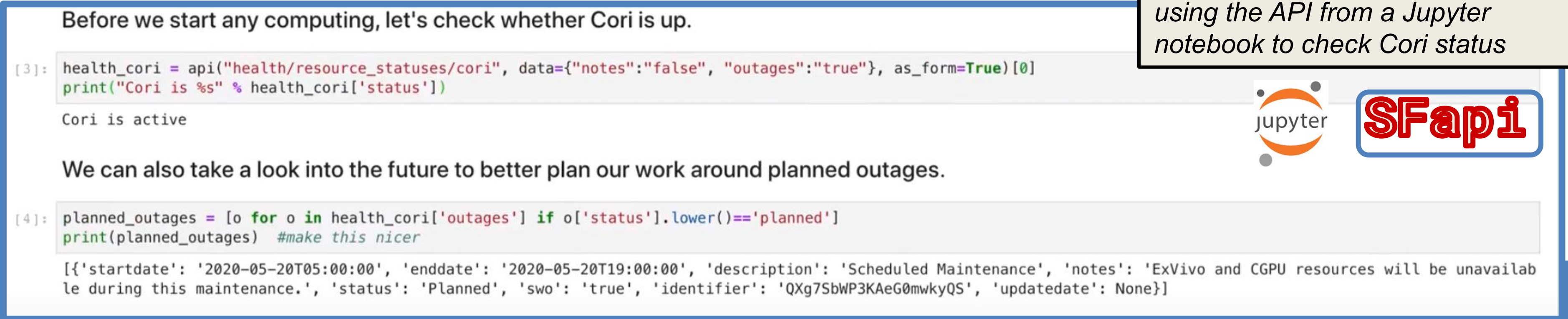
POST `/callbacks/callbacks/` This api requires authentication

https://api.nersc.gov/

The Superfacility API: sustainable, scalable automation

- Less user/staff DIY: simpler, standardized tooling (Python, etc)
  - Stable refactor target for established projects
  - Easier on-ramp for new projects
- Fit (not fight) standard software design patterns
  - Shared libraries and API calls
  - Authentication and security models built on OAuth2 Standard and JSON Web Tokens (JWTs)

using the API from a Jupyter notebook to check Cori status



Before we start any computing, let's check whether Cori is up.

```
[3]: health_cori = api("health/resource_statuses/cori", data={"notes":"false", "outages":"true"}, as_form=True)[0]
     print("Cori is %s" % health_cori['status'])
```

```
Cori is active
```

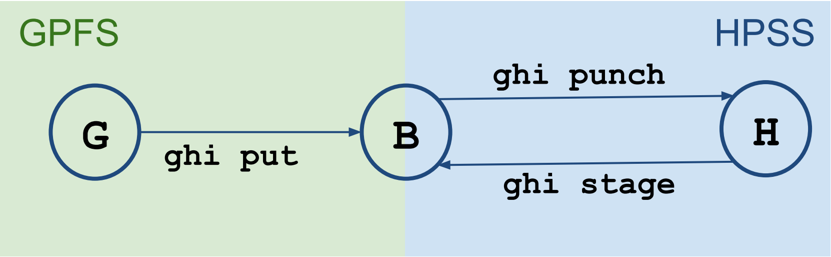
We can also take a look into the future to better plan our work around planned outages.

```
[4]: planned_outages = [o for o in health_cori['outages'] if o['status'].lower()=='planned']
     print(planned_outages)  #make this nicer
```

```
[{'startdate': '2020-05-20T05:00:00', 'enddate': '2020-05-20T19:00:00', 'description': 'Scheduled Maintenance', 'notes': 'ExVivo and CGPU resources will be unavailab
le during this maintenance.', 'status': 'Planned', 'swo': 'true', 'identifier': 'QXg7SbWP3KAeG0mwkyQS', 'updatedate': None}]
```

GPFS-HPSS Interface

exemplary use cases



- **Archiving Complex Directory Structures**
  - Experiments [at the ALS at LBNL] often have complex sets of [microscopy image] data, i.e. large volumes of image data in tens of MBs along with a few kB-sized text file.
  - The files are organized in a complex directory structure that must be maintained.
  - Use a single command (`ghi put`) to archive the entire directory into HPSS.
- **Large volumes of Infrequently Accessed Data**
  - 100TBs of data only accessed a few times a year for reanalysis (DESI)
  - Use `ghi put` to put the data into HPSS
  - In between analyses, frees up disk space by using `ghi punch` to move most data off of GPFS but still leaves a browsable directory structure behind
  - Selectively retrieves some (or all) files with `ghi stage`