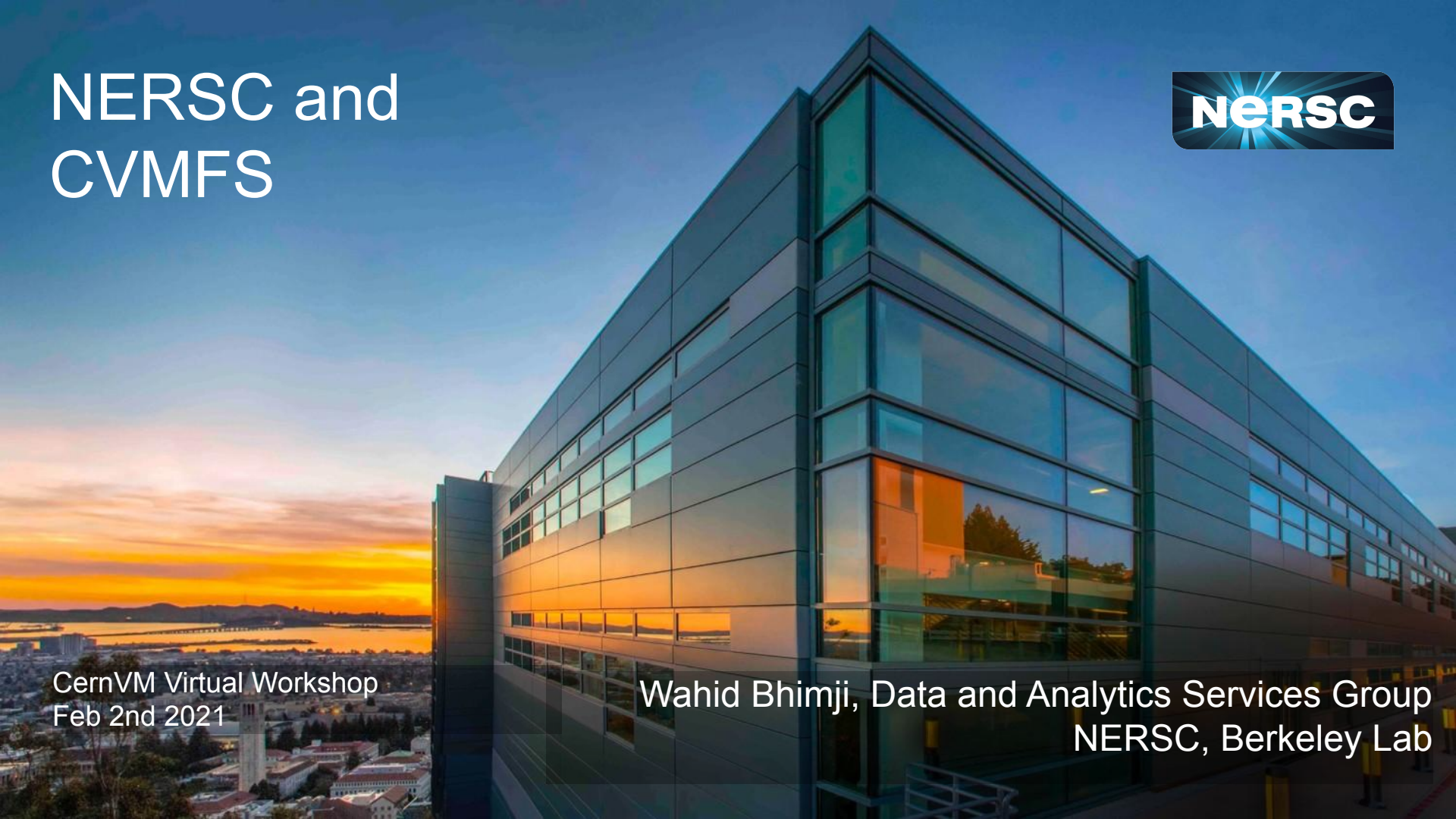
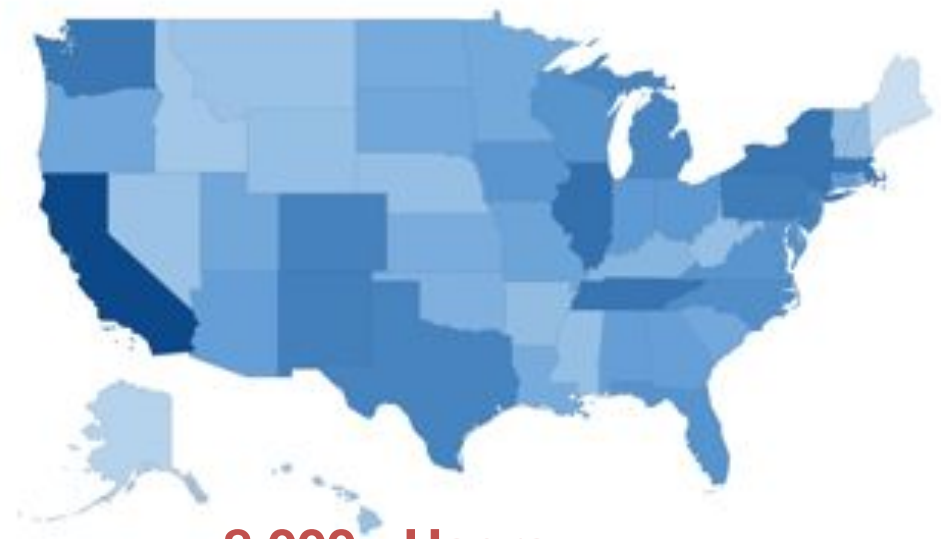
NERSC and CVMFS



CernVM Virtual Workshop
Feb 2nd 2021

Wahid Bhimji, Data and Analytics Services Group
NERSC, Berkeley Lab

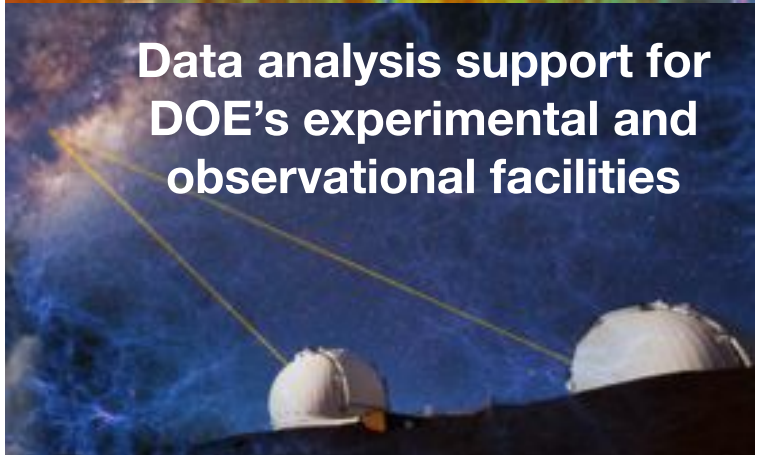
NERSC is the **mission** High Performance Computing facility for the DOE Office for Science



8,000+ Users

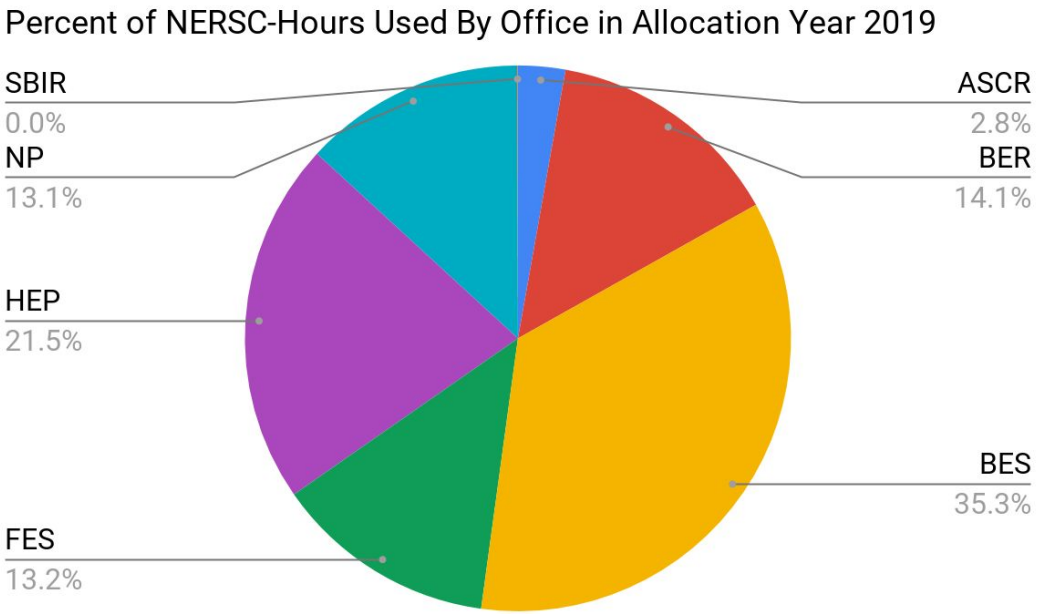
800+ Projects

2000+ NERSC citations per year



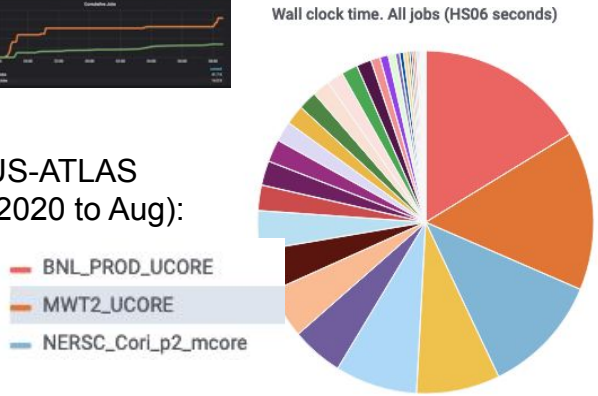
NERSC

NERSC is not HEP/NP dedicated - but can be a valuable HEP resource

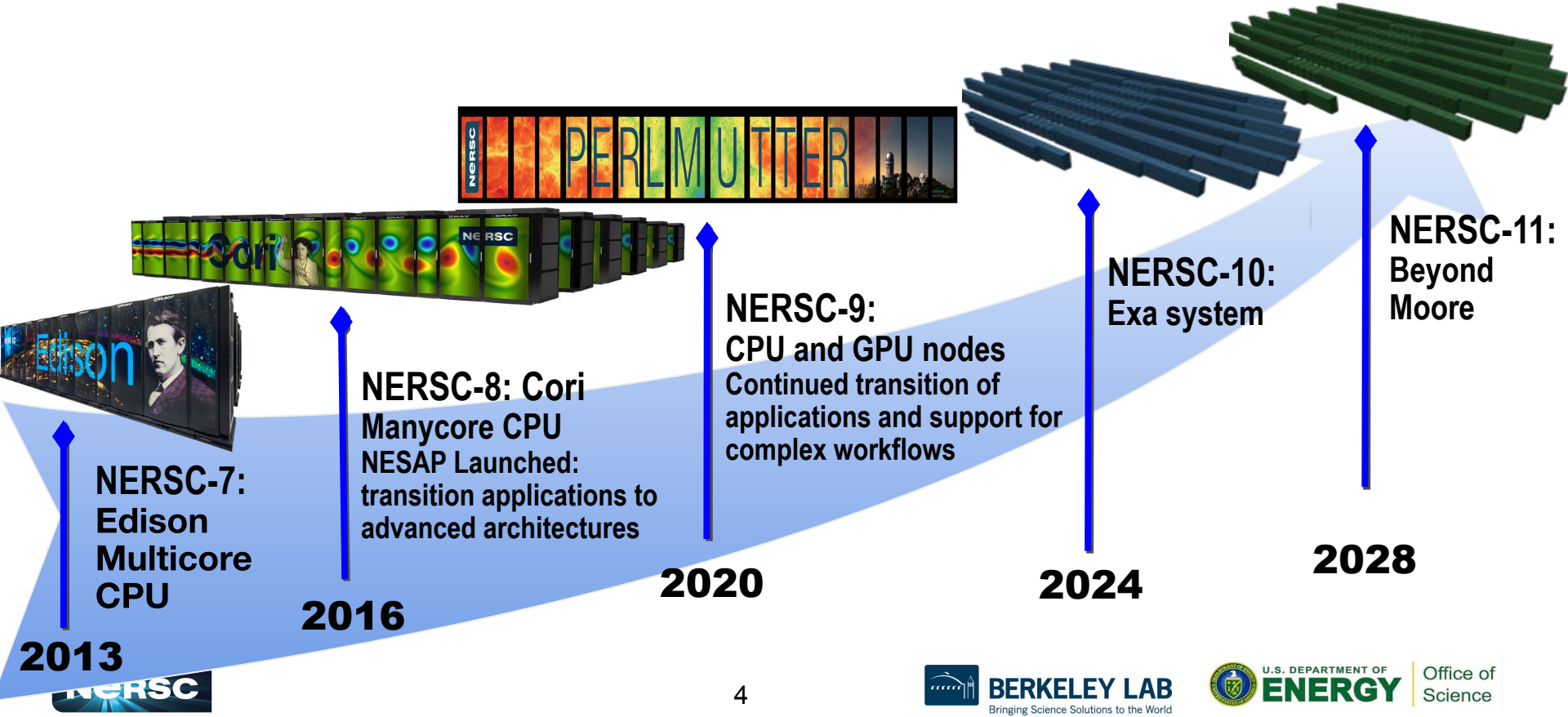


E.g. [NoVA '1m (CPU) cores'](#)

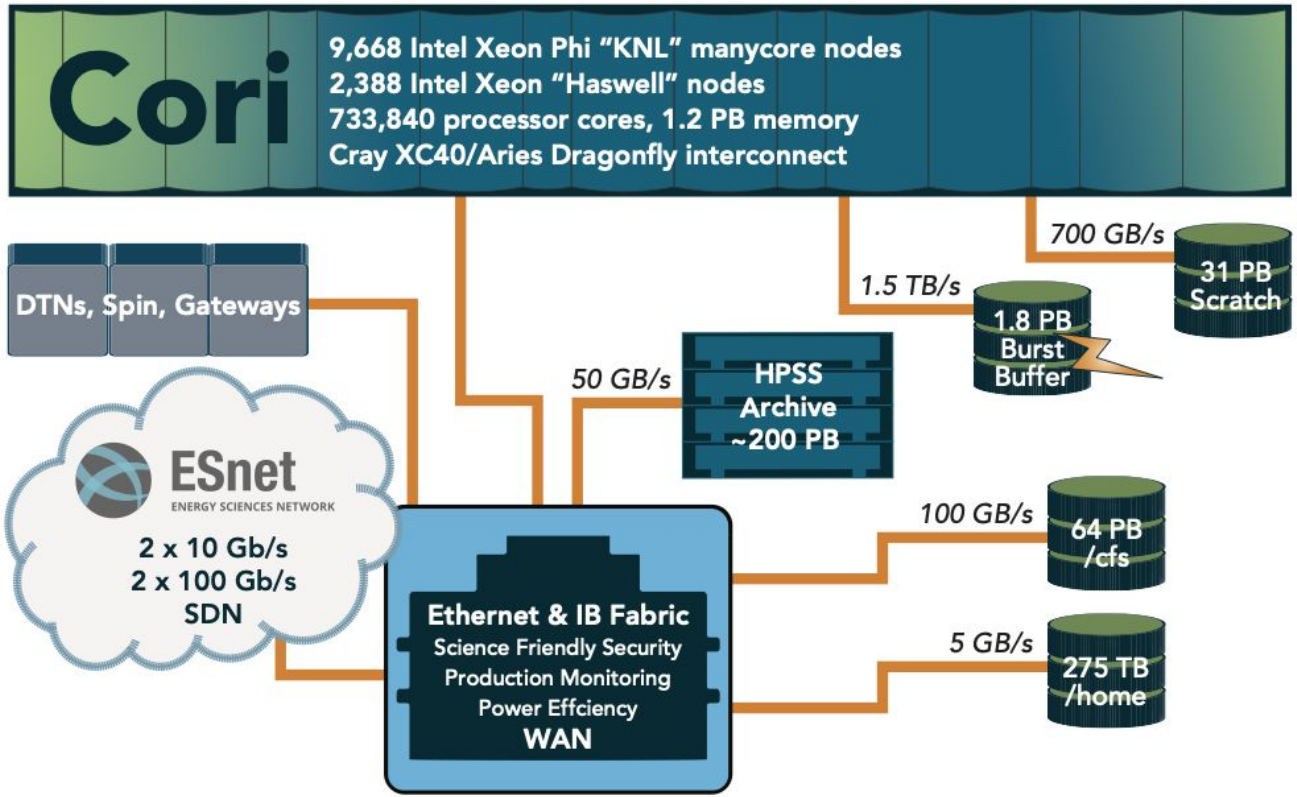
US-ATLAS
(2020 to Aug):



NERSC Systems Roadmap

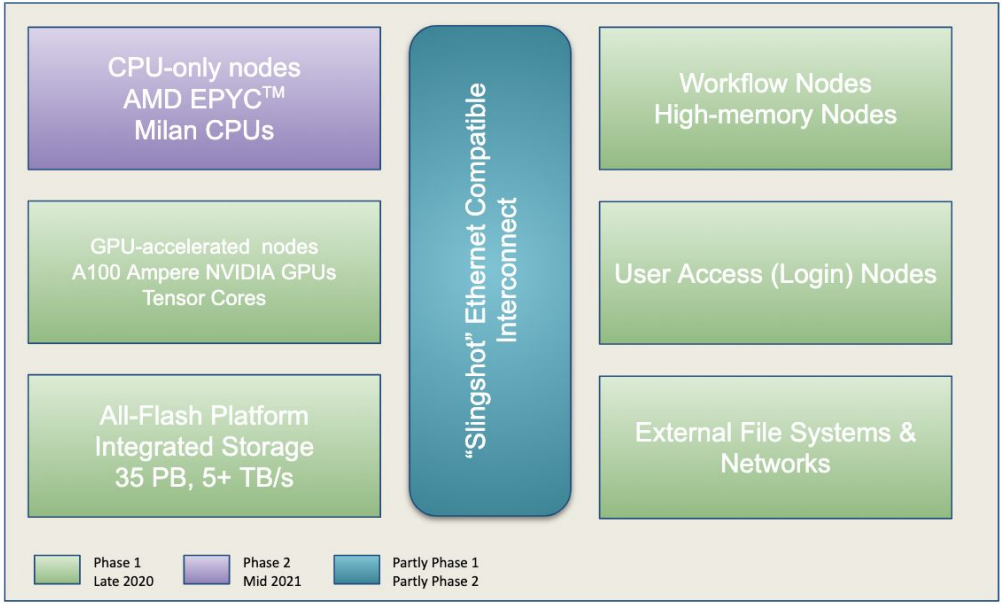


NERSC Centre 2020



Perlmutter: a System Optimized for Science

- AMD/NVIDIA **A100-accelerated** and **CPU-only nodes** meet the needs of large scale simulation and data analysis from experimental facilities
- HPE/Cray "**Slingshot**" - High-performance, scalable, low-latency Ethernet- compatible network
  - seamless connection between inside/outside the machine
- Single-tier **All-Flash Lustre** HPC file system, 6x Cori's bandwidth
- Dedicated login and high memory nodes to support complex workflows



NERSC

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science

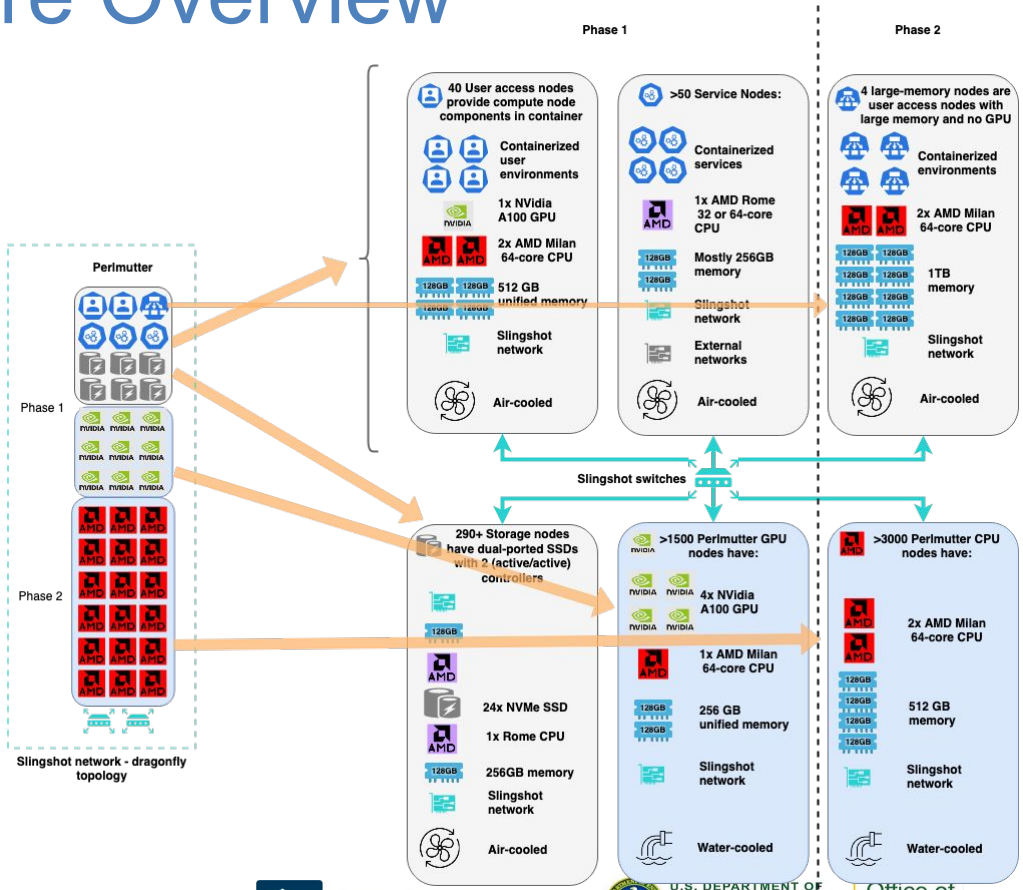
Perlmutter Architecture Overview

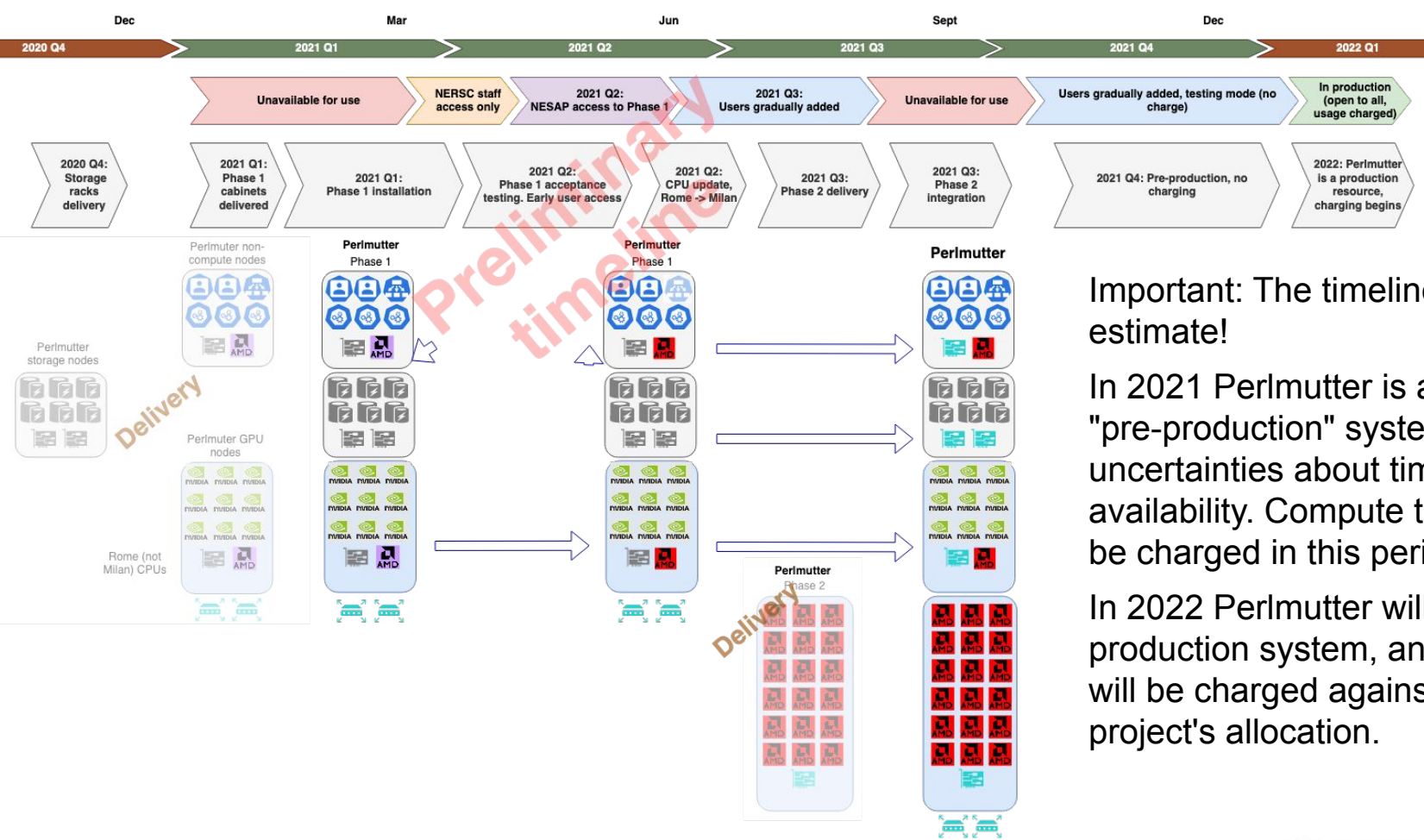
Phase 1:

- Non-compute nodes (First 20 login nodes, service nodes)
- Storage (35 PB Lustre, all flash)
- GPU compute: 1500+ nodes with 4x NVidia A100 GPUs

Phase 2:

- 4 Large memory nodes
- 20 more login nodes
- CPU compute: 3000+ nodes with 2x AMD 64-core CPU





Important: The timeline is an estimate!

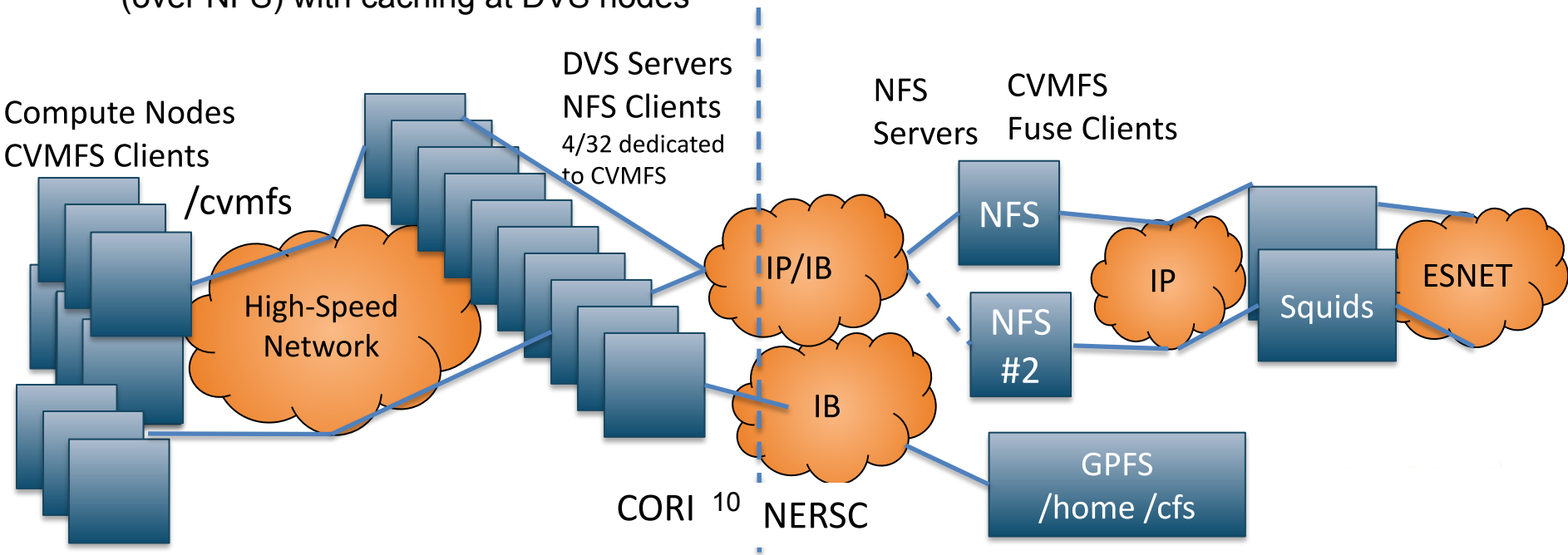
In 2021 Perlmutter is a "pre-production" system: uncertainties about timing and availability. Compute time will not be charged in this period.

In 2022 Perlmutter will become a production system, and usage will be charged against the project's allocation.

CVMFS @ NERSC

CVMFS on Cori..

- Restrictions with compute OS (FUSE etc.) made cvmfs at NERSC historically challenging
- Could [install into containers] – used in production/automated by ATLAS/CMS
- But large images; non-instant updates; adding other releases/repos not easy etc.
- So we use Cray DVS (IO forwarder for non-lustre filesystems) to provide up-to-date CVMFS (over NFS) with caching at DVS nodes



CVMFS on Cori experiences

```
wbhimji@cori12:~> ls /cvmfs
alice-ocdb.cern.ch              grid.cern.ch
alice.cern.ch                   icecube.opensciencegrid.org
ams.cern.ch                     larsoft.opensciencegrid.org
atlas-condb.cern.ch             lz.opensciencegrid.org
atlas-nightlies.cern.ch         nova-development.opensciencegrid.org
atlas.cern.ch                   nova.opensciencegrid.org
cms.cern.ch                     oasis.opensciencegrid.org
config-osg.opensciencegrid.org  sft.cern.ch
cvmfs-config.cern.ch            spt.opensciencegrid.org
dune.opensciencegrid.org        star.sdcc.bnl.gov
fermilab.opensciencegrid.org    sw.lsst.eu
gm2.opensciencegrid.org         uboone.opensciencegrid.org
wbhimji@cori12:~>
```

24 repositories - multiple diverse projects

Holds up fine at scale

But *many* issues:

- E.g. returning the wrong file due to inode clash - occurred 3 separate times for 3 different reasons. (The last one was because ld.so uses its own caching of inodes to not load the same library (not the VFS layer))

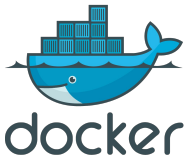
CVMFS will be supported on Perlmutter:

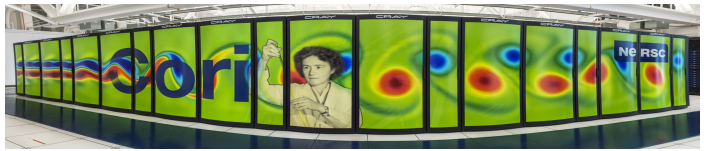
- We are evaluating ‘normal’ FUSE-based solutions
- Don’t have local disk so exploring either tiered-cache or loopback fs (similar to shifter per-node-cache but needs to persist across jobs)

Other communities software approaches

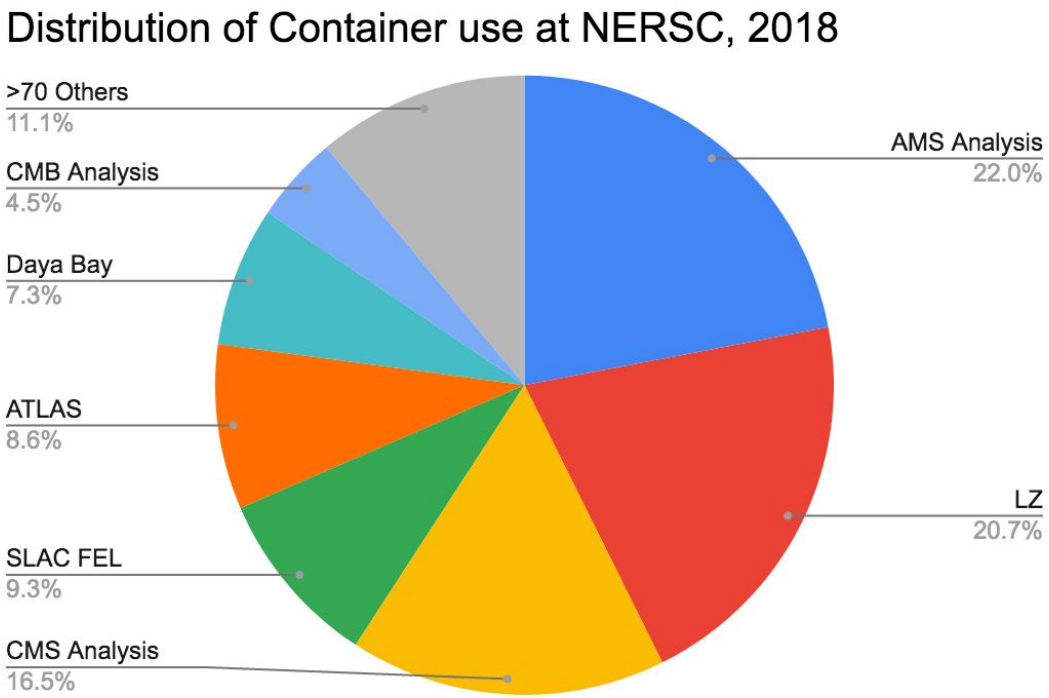
Containers at NERSC currently via Shifter

- **NERSC development effort, in collaboration with Cray, to support Docker Application images**
- **“Docker-like” functionality on Cray and HPC Linux clusters. Enables users to run custom environments on HPC systems.**
- ***Addresses security issues in a robust way***
- ***Efficient job-start & Native application performance***





HPC Container Usage at NERSC



* Only includes jobs that specify image at submission

- Many individual users as well as large collaborations with complex software stacks and dependencies
- Container use has increased dramatically:
  1%* in 2014
  6-8%* in 2018-2020
- 7000+ Unique Image tags
- 900+ Unique Users
- Used at all scales and multiple science domains (HEP, BER, AI/DL)

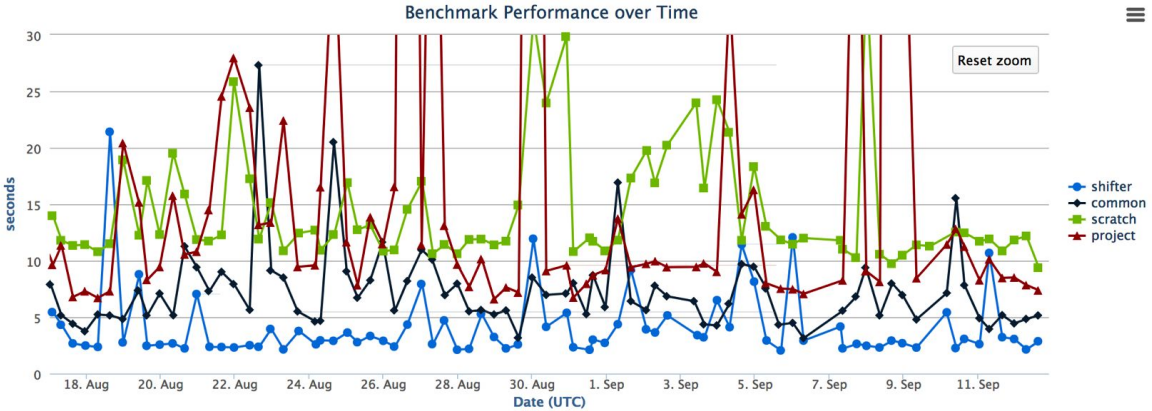
Containers and Perlmutter

- Working to ensure Shifter supported on Perlmutter for continuity
- Exploring several container engine options for the future (e.g. podman) and standard optimized images (e.g. ngc)
- Cray Shasta System, will incorporate containers and Kubernetes into the system management plane
- Also exploring potential for k8s for compute job workflows (in addition to slurm) - kubeflow etc. - interested in use-cases

kubernetes

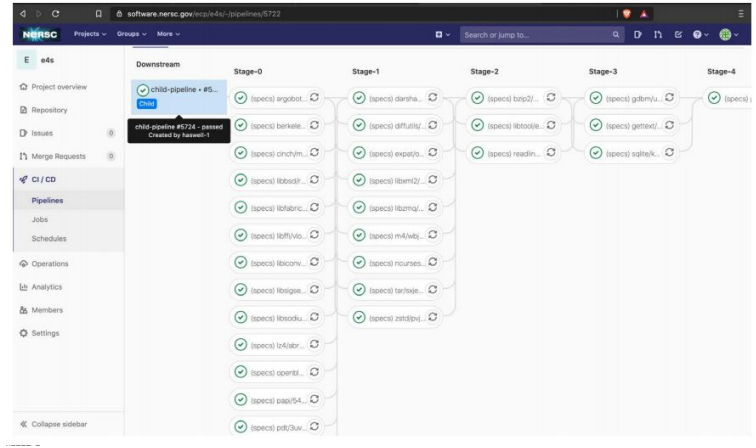
“Global Common”: Software Filesystem

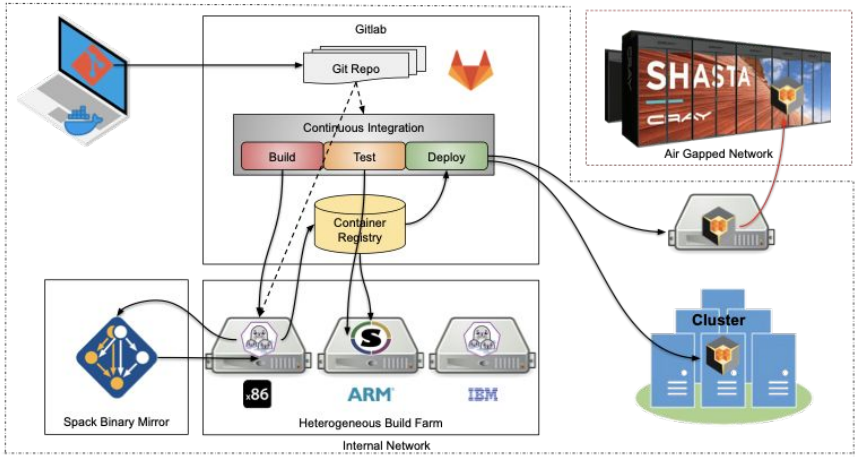
- Group writable directories on GPFS, but with a smaller quota than community (cfs), /global/common/software/<projectname>
  - Write from login node; read-only on compute node
- Smaller block size for faster compiles than project
- Faster library loads than other shared filesystem (but slower than shifter)



Other activity

- NERSC itself builds software for users (loaded via “modules”)
  - Moving to standardizing installation practices
- Exascale Computing project has software technology component(s) https://e4s.io/
  - Spack packages





SUPERCONTAINERS

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

Conclusions

- NERSC is an HPC center with a very varied user base
- HEP/NP have quite unique requirements/approaches

- We support cvmfs: used by several projects/experiments
- Integrating cvmfs at NERSC has had some challenges
- New system *Perlmutter* offers significant resources
  - Shasta software brings opportunities (and challenges)
  - We will make cvmfs available and hope to be closer to ‘normal’
  - Containers on HPC are here to stay with increasing capabilities