

NERSC Site Update



2025 West Coast IBM Storage Scale
User Group Event

Steve Leak
Storage Infrastructure Specialist
Aug 7, 2025

Large-scale storage for large-scale compute

Perlmutter

Cray EX supercomputer

4864 compute nodes

7168 NVidia A100 GPUs

7936 AMD EPYC 7763 (Milan) CPUs

79PFLOPs (HPL), Top500 #25

(Debut #5 in 2021 with 64.6PFLOPS)



Pscratch - Lustre

HPE ClusterStor E1000

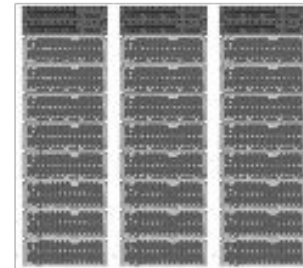
44PB Flash, 370 OSTs

CFS ("Community File System")

IBM ESS5000 SC8 - Spectrum Scale

18 racks, 160PB raw / 147PB usable

15,264 HDDs



HPSS (Tape)

IBM TS4500

4 lbs x 16 frames

~400 PB stored

>650 PB capacity

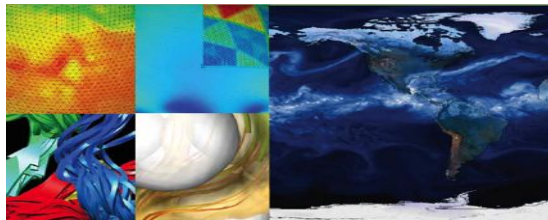
NERSC: Mission HPC for DOE Office of Science Research



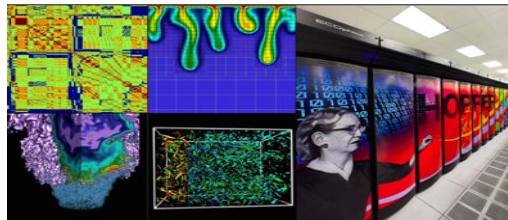
U.S. DEPARTMENT OF
ENERGY

Office of
Science

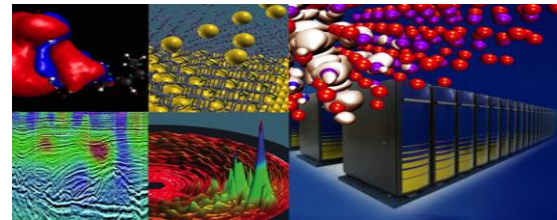
Largest funder of physical science
research in the U.S.



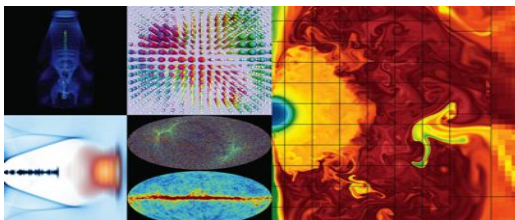
Biological and Environmental Research



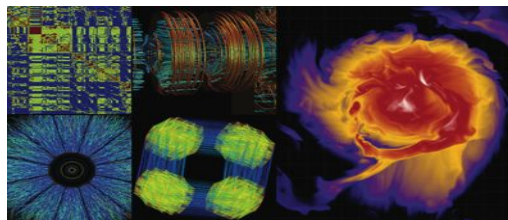
Computing



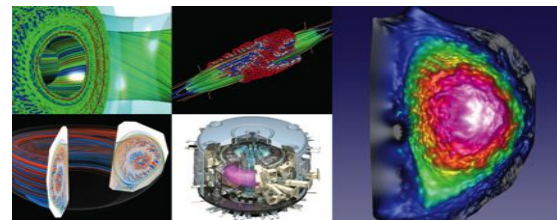
Basic Energy Sciences



High Energy Physics



Nuclear Physics



Fusion Energy, Plasma Physics

NERSC: Mission HPC for DOE Office of Science Research

NERSC

About Us

Mission: science

NERSC's mission is to provide research community with the DOE Office of Science through high performance computing and data analysis.

As the nation's scientific computing facility for the U.S. Department of Energy (DOE) Office of Science (OS), NERSC is the preeminent provider of high-performance computing and data analysis services to the research community.

- Advanced Scientific Computing Research
- Advanced Scientific Computing Research
- Advanced Scientific Computing Research
- Advanced Scientific Computing Research
- Advanced Scientific Computing Research
- Advanced Scientific Computing Research
- Advanced Scientific Computing Research
- Advanced Scientific Computing Research

Innovating for the future

NERSC is the National Energy Research Scientific Computing Center (NERSC) established as part of accelerating scientific discovery for the DOE Office of Science (OS). Launched in 1996 as a computing resource for fusion energy research at Lawrence Livermore National Laboratory, NERSC quickly expanded its role to include work from all OS programs.

In 2016, the center moved to Lawrence Berkeley National Laboratory.

Today, NERSC is one of the world's most scientifically productive high performance computing (HPC) centers. Our state-of-the-art produce an average of 100,000 publications annually using NERSC resources. Over the years, NERSC has been privileged to serve some of the world's leading scientific and engineering teams.

NERSC USERS ACROSS US AND WORLD

50 States,
Washington D.C.
& Puerto Rico

53 Countries

~10,000 Annual Users from **~800** Institutions + National Labs

50 years of scientific computing



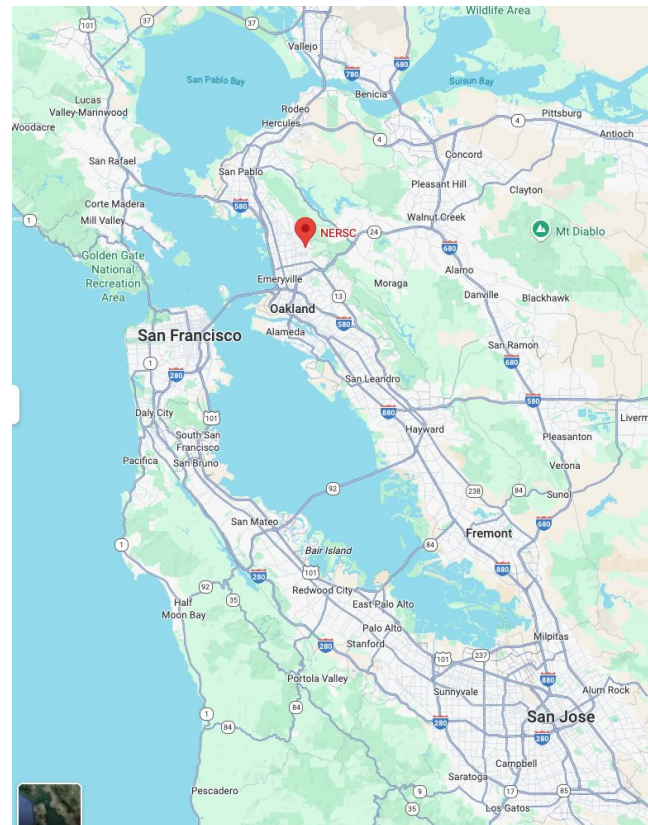
NERSC is one of the world's most scientifically productive high performance computing (HPC) centers. Our state-of-the-art produce an average of 100,000 publications annually using NERSC resources. Over the years, NERSC has been privileged to serve some of the world's leading scientific and engineering teams.

Shyh Wang Hall

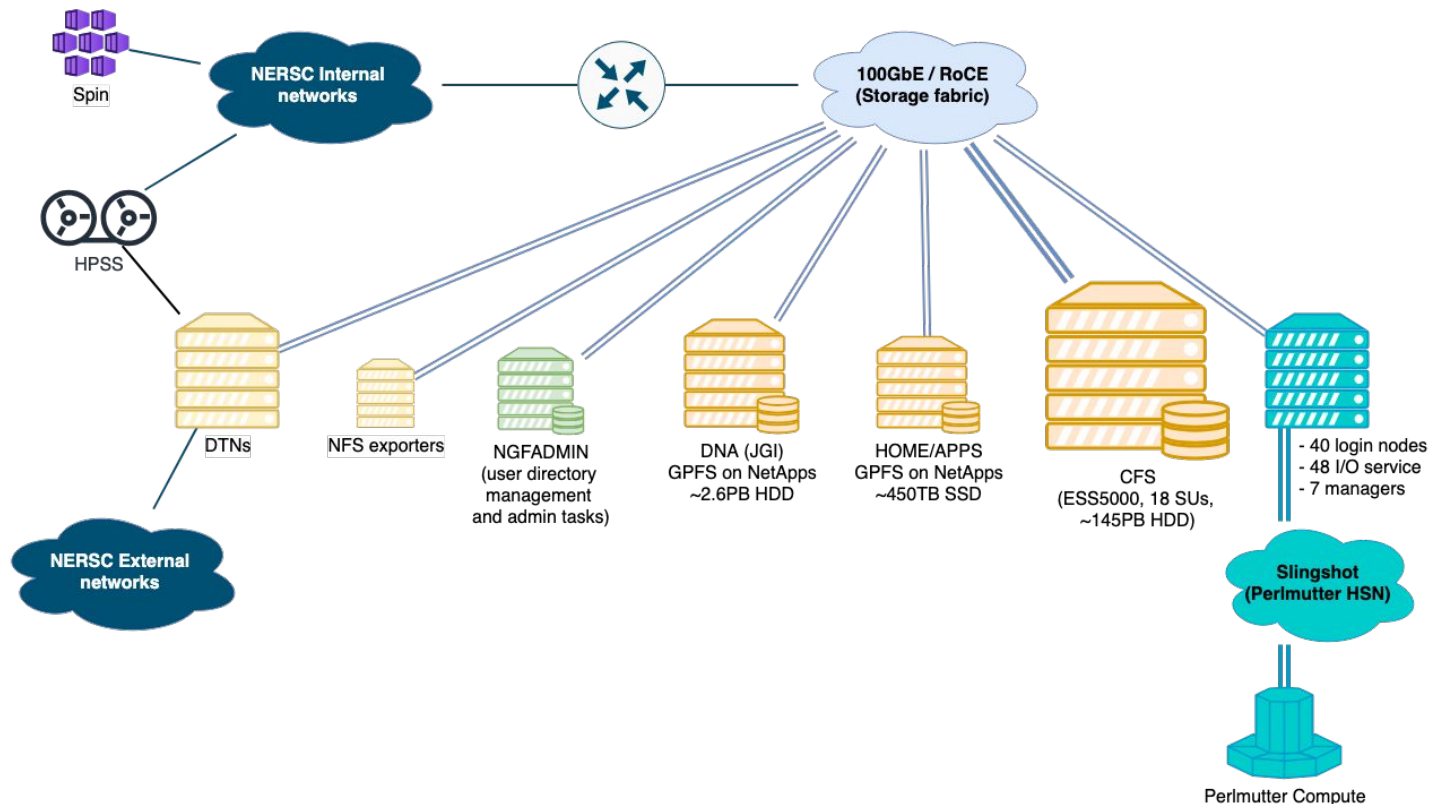
Shyh Wang Hall is the new home for NERSC, located at the Berkeley Lab campus. Shyh Wang Hall is a 12-story building with state-of-the-art safety and energy efficiency features.



www.nersc.gov



Mostly Storage Scale, Multi-cluster, RoCE



Community File System (CFS)

CFS is our largest filesystem: currently 145 PB, 7B inodes

- ESS5000 SC8 (capacity oriented)
- 18 SUs (>15k disks!)
- 8x100GbE RoCE links/SU (traffic mostly outbound)
- Campaign storage: multi-year, fairly-actively used
- ~1000 projects:
 - Many users are in multiple projects
 - Some projects have subprojects, with own quotas etc
 - Imagine 1000 data lakes
- Growth in space and inode usage is ~linear

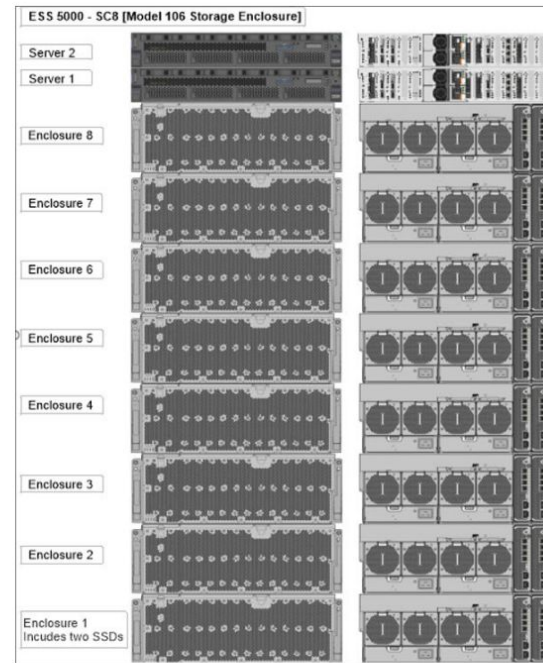


Figure 2-4 IBM ESS 5000 system with Model 106 expansion enclosures

Implementation Guide for IBM Elastic Storage System 5000

Large-scale storage for large-scale compute

Perlmutter

Cray EX supercomputer

4864 compute nodes

7168 NVidia A100 GPUs

7936 AMD EPYC 7763 (Milan) CPUs

79PFLOPs (HPL), Top500 #25

(Debut #5 in 2021 with 64.6PFLOPS)



Pscratch - Lustre

HPE ClusterStor E1000

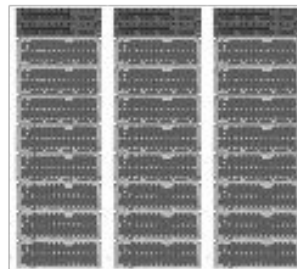
44PB Flash, 370 OSTs

CFS ("Community File System")

IBM ESS5000 SC8 - Spectrum Scale

18 racks, 160PB raw / 147PB usable

15,264 HDDs



HPSS (Tape)

IBM TS4500

4 lbs x 16 frames

~400 PB stored

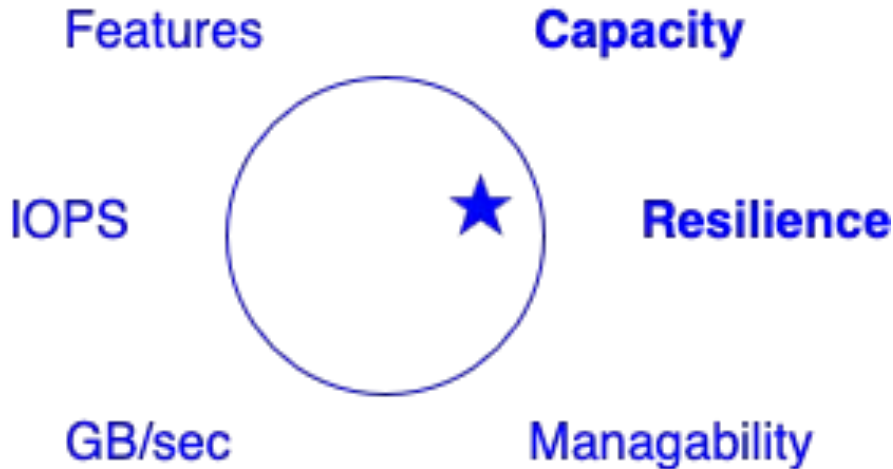
>650 PB capacity

Priorities

CFS: Resilience and Capacity

Platform Storage: Performance and Capacity

Other filesystems: IOPS and Resilience



Design/usage decisions and tradeoffs

- One giant filesystem
 - Easier to manage
 - Parallelism across SUs
 - Maintenance operations take a long time!
 - Limit upgrades to 1-2 per year, favor x.1, x.2 etc (avoid x.0)
 - Filesystem itself can be a SPOF
 - fsck likely requires several days downtime
- Limited feature use
 - No tiering, AFM, CSI, etc
 - Filesets , Multiclusterv , Callhome , Snapshots , ILM (listall) 
 - ESS Web GUI, mmhealth (a bit)

Operations

- Rock-solid resilience
 - High component count => regular component failures (pdisks esp.)
 - Redundancy in design => system availability is not affected (no getting paged at midnight!)
 - Callhome! Much ticket-management labor just goes away
 - Great support and SSRs
- Upgrade process - less rock solid
 - With 18 SUs, takes a long time (can do multiple I/O nodes at once, tradeoff is risk of impact or delay if one goes not-smoothly)
 - Would be great to be able to roll back more easily
 - Moving a Recovery Group to partner node is not-quite-invisible - users do see brief I/O pauses
 - One-button upgrade is interesting. Can a highly-risk-averse customer use it?

Operations: Big-impact events

fsck

- With 100+PB and 6B+ inodes, mmfsck take a long time
- Online mmfsck has limited applicability, offline needed
- Plan for 2-3 passes, a few days
- --estimate-only option is valuable in preparation
- Networking, and node specs for the fs manager, matter
 - watch for mmfsd abort at completion, which node is fs manager now?
- Great support from IBM
- Happy outcome: no data loss

Coming up next

SSS6000 C9 (now)

- Expand/update CFS
- 5 SUs: 90PB raw / 67PB usable
- Spinning disk, x86 I/O and EMS nodes

NERSC Doudna (2026)

- Dell, NVIDIA compute
- IBM, VAST Data storage





NVIDIA A100 + AMD EPYC

79 PFLOPs

Lustre platform storage

NERSC Doudna (2026)



NVIDIA Vera-Rubin (Arm)

Target >10x Perlmutter PFLOPs

2 Storage systems:

- Platform Storage System (PSS) - IBM, Storage Scale
- Quality-of-Service Storage System (QSS) - VAST Data)

Target: up to 5x faster than
Perlmutter storage

Doudna Storage

- “Scientific workloads are evolving into complex workflows”
- “The *Doudna* supercomputer will provide more than ten times the performance of *Perlmutter*, NERSC’s current flagship supercomputer. *Doudna* will be built using cutting-edge technology such as Dell’s most advanced ORv3 direct liquid-cooled server technology and the NVIDIA Vera-Rubin CPU-GPU platform.”
(<https://www.nersc.gov/news-and-events/news/doe-announces-new-supercomputer-powered-by-dell-and-nvidia-to-speed-scientific-discovery>)
- “With technology provided by VAST Data and IBM, *Doudna* will feature storage performance up to five times faster than NERSC’s current system and offer performance guarantees for time-sensitive science”
(<https://www.nersc.gov/news-and-events/news/doudna-storage-solutions>)
- “The PSS will be provided by IBM and function as an all-flash high-performance parallel scratch filesystem based on IBM’s software-defined file and object storage platform, IBM Storage Scale.”
- “[The QSS] will allow NERSC to provide deadline-dependent workloads with performance guarantees through either persistent or schedulable performance”

Doudna Storage

Changing needs => Adjustment of priorities

All-flash Storage Scale for PSS (closer in priorities to Perlmutter's Lustre than to CFS)

VAST for schedulable performance

Q: How will PSS and QSS interact with CFS? Some options to explore:

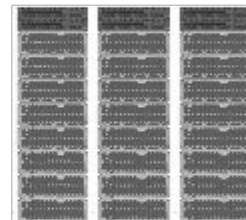
- No interaction, standalone systems
- AFM to move data between PSS and CFS
- PSS as a hot tier storage pool within CFS
- Other?

Large-scale storage for large-scale compute

Pscratch - Lustre
HPE ClusterStor E1000
44PB Flash, 370 OSTs



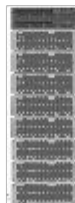
CFS ("Community File System")
IBM ESS5000 SC8 - Spectrum Scale
23 racks, 250PB raw / 214PB usable



Perlmutter
Cray EX supercomputer
4864 compute nodes
7168 NVidia A100 GPUs
7936 AMD EPYC 7763 (Milan) CPUs
79PFLOPs (HPL), Top500 #25
(Debut #5 in 2021 with 64.6PFLOPS)



Doudna
Dell supercomputer
Vera-Rubin compute
IBM (PSS) and
VAST (QSS) storage



HPSS (Tape)
IBM TS4500
4 libs x 16 frames
~400 PB stored
>650 PB capacity

Q&A?



2025 West Coast IBM Storage Scale
User Group Event

Steve Leak
Storage Infrastructure Specialist
Aug 7, 2025