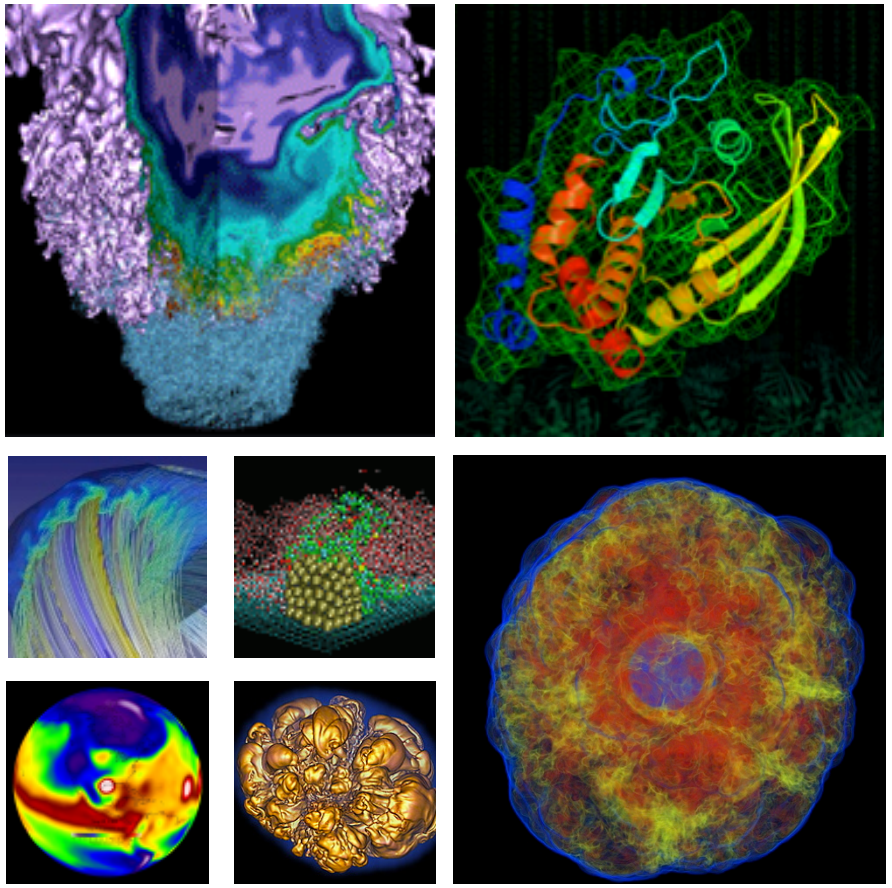
The HPC Data Center of the Future



Jason Hick

Lawrence Berkeley National Laboratory
NERSC Storage Systems Group

HPC Joint Facility User Forum
June 16, 2014

NERSC 40 YEARS at the FOREFRONT 1974-2014

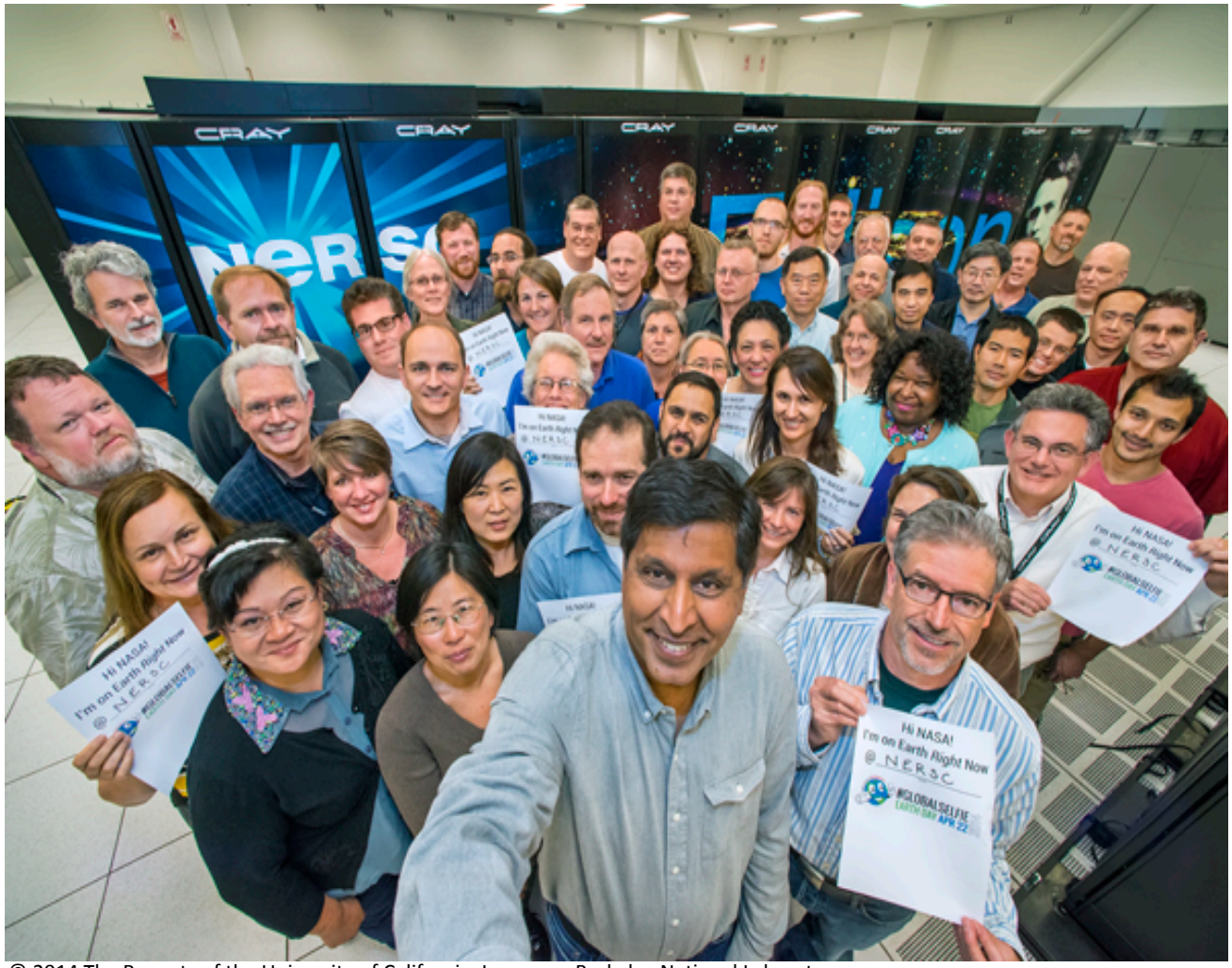
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Agenda

- **NERSC's storage systems & services**
- **Trends of existing storage-class hardware**
  - Flash overtakes disk for $/GB/sec
- **Future storage-class hardware**
  - Memristor, MRAM
- **Storage software advancements**
  - Metadata performance
  - Burst buffer
  - Access to storage systems
- **NERSC in 2020**
- **What this means to the user**

National Energy Research Scientific Computing Center (NERSC)

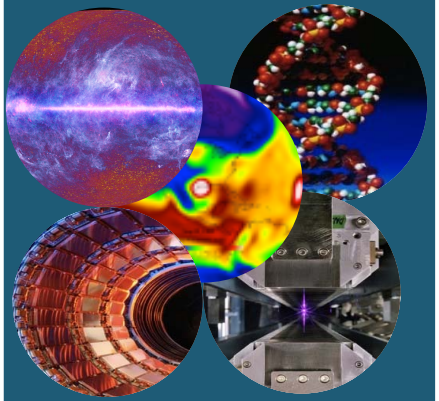


- **Located at Berkeley Lab**
- **User facility supports 6 DOE Offices of Science:**
  - 5000 users, 600 research projects
  - 48 states; 65% from universities
  - Hundreds of users each day
  - ~1500 publications per year
  - With services for consulting, data analysis and more

Types of computing at NERSC

Data Intensive

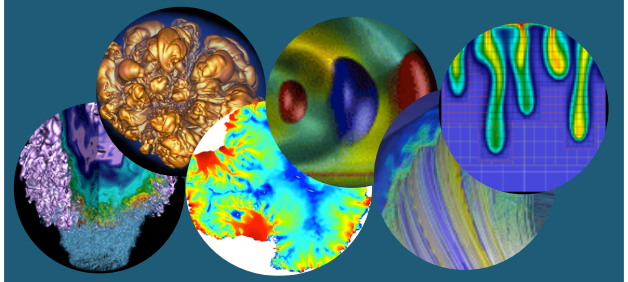
Experiments and Simulations



NERSC ingests, stores and analyzes data from Telescopes, Sequencers, Light sources, Particle Accelerators (LHC), Microscopes, and other scientific instruments

Large Scale

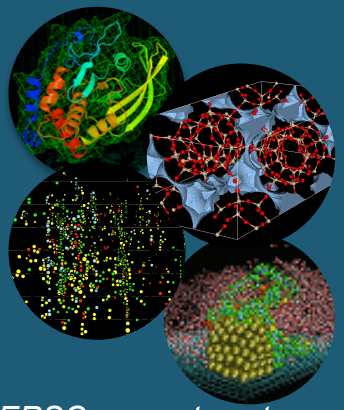
Capability Simulations



Petascale systems run simulations in Physics, Chemistry, Biology, Materials, Environment and Energy at NERSC

High Volume

Job Throughput



NERSC computer, storage and web systems support complex workflows that run thousands of simulations to screen materials, proteins, structures and more; the results are shared with academics and industry through a web interface

NERSC

Petascale Computing, Petabyte Storage, and Expert Scientific Consulting

The compute and storage systems 2014

Hopper: 1.3PF, 212 TB RAM

Cray XE6, 150K Cores

Edison: 2.5PF, 357 TB RAM

Cray XC30, 130K Cores

Production Clusters
Carver, PDSF, JGI,KBASE,HEP
14x QDR

Vis & Analytics, Data Transfer Nodes, Adv. Arch., Science Gateways

2.2 PB Local Scratch 70 GB/s

16 x QDR IB

6.4 PB Local Scratch 140 GB/s

16 x FDR IB

80 GB/s

50 GB/s

5 GB/s

12 GB/s

Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN

/global/ scratch — ***3.6 PB 5 x SFA12KE***

/project — **5 PB DDN9900 & NexSAN**

/home — ***250 TB NetApp 5460***

HPSS — ***65 PB stored, 240 PB capacity, 40 years of community data***

2 x 10 Gb

1 x 100 Gb

Science Data Network

ESnet Energy Sciences Network

U.S. DEPARTMENT OF ENERGY | Office of Science

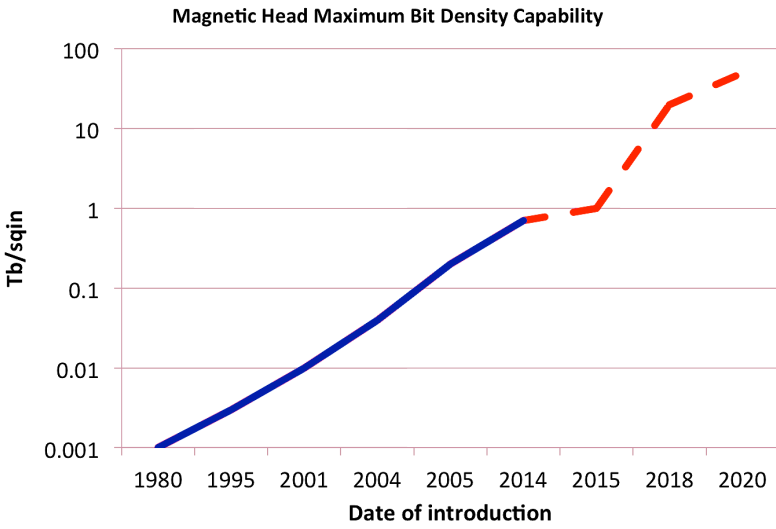
BERKELEY LAB

Storage systems and services

- **Parallel file systems (Lustre and GPFS) are primary storage to supercomputers**
  - Total of over 20 PBs of disk available to users
  - Some multi-PB parallel file systems backed up to HPSS (Parallel Incremental Backup System)
    - Has demonstrated it can identify backup candidates from 500M total files and process over 150TBs of backup data in a single day
- **Archival and backup systems (HPSS) are secondary storage for users**
  - 65 PBs of data stored, growing at >1PB per month
  - 30% of user IOs are read/retrieve requests from archival storage, so a very active archive
  - Focus on reliability of the system for user data by:
    - Deploying solutions to proactively monitor and maintain health of user data, and environmental parameters necessary for tape
    - Actively migrating/moving data within the system
- **Storage services highly utilized**
  - Inter-facility data transfers
  - Science gateways for sharing data with collaborators

Trends in storage-class hardware

- **Through innovation, the storage industry has delivered significantly more capacity and bandwidth at reduced cost for decades**
- **Magnetic storage is the primary method**
  - Hard disk drives – primary innovation is in head technology
  - Tape – primary innovation is in materials technology (new tape media), leverage head technology from disk
- **Optical storage used occasionally but slow performance limits it to only work well for write-once read many (WORM) or "cold storage" scenarios**
- **Innovation in magnetic storage bit densities is getting very difficult**
  - Superparamagnetic effect (random bit flips at high bit densities) requires high R&D and longer time to market for production products
  - Head design of magnetic storage currently use tunneled magneto-resistive (TMR)
- **Transistor and circuit based storage technologies look promising**
  - But capacity and productization are lagging magnetic storage
- **Today, NAND flash technology has edged magnetic storage for $/GB/s**



Bit-patterned media & Heat-assisted Magnetic Recording (HAMR) are being developed to help break the 1Tb/sqin limit, but are difficult to mass produce

Future storage-class hardware

- **Magnetoresistive random access memory (MRAM)**
  - Micron and Tokyo Electron are leading a 20 company development effort to mass produce by 2018
  - DRAM speed with endurance of Flash
  - Faster computing with reduced energy usage
  - 64Mb sized devices today, working towards DRAM capacities
- **Memristors**
  - Memory resistor that remembers its most recent resistance even with the power turned off
  - Capable of performing logic operations
  - Developing an integrated circuit to replace DRAM and Flash
- **Future systems will have onboard storage-class memory**
- **Need software to evolve to exploit this capability**
  - Features for data locality and data staging

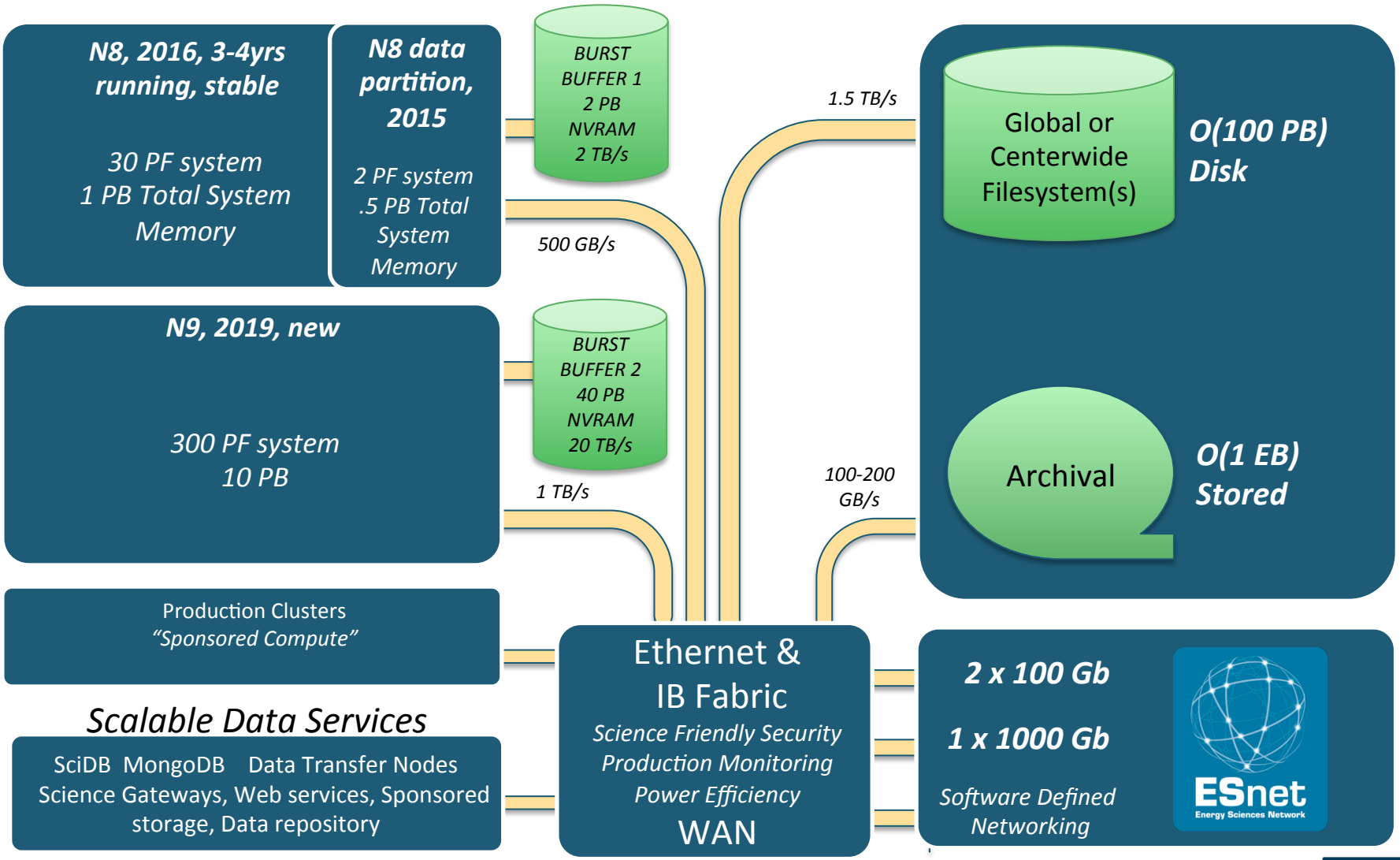
Storage software advancements

- **Improved metadata performance ~2016**
  - Lustre DNE
  - HPSS 7.5
- **FastForward I/O software**
  - Advancements to HDF5, PLFS, and Lustre to provide object-based storage system
- **Burst buffer ~2016-2018**
  - See Trinity/N8 Use Case Scenarios 4/5/2013, but here are some…
  - Checkpoint/restart (pre-stage and drain)
  - Shared read cache (N-to-1)
  - Temporary job data (e.g. swap)
  - Data analysis and visualization (e.g. in-situ analysis)
- **Improving user access to their data**
  - Multisite access through Globus
  - HPSS has file system-like access modifications coming (Fuse)
  - GPFS asynchronous file migration (AFM)
  - Could use LTFS as data format for exchanging data between facilities for massive datasets
- **Storage quality of service (QoS)**
  - Storage systems cross-mounted on key systems (or at multiple sites with unattended data set transfer capability)
    - Measuring current performance relative to workload
    - Ability to guarantee bandwidth
    - Site specific authentication and authorization issues solved

What does your burst buffer do?

- **Burst buffers should improve compute node utilization when they spend significant time in I/O wait**
- **Burst buffers (software + SSDs) provide**
  - Bandwidth with less hardware than disk
  - Energy efficiency over disk
  - Better performance for varied workload or imperfect I/O
- **Enables several options for use**
  - Use scheduler to request job data to be staged into BB read cache
  - Use scheduler to request job data be migrated from BB to file system upon conclusion of the job
  - Use HDF5/FF BB API to perform asynchronous I/O unconstrained by POSIX rules

NERSC 2020


N8, 2016, 3-4yrs running, stable
30 PF system
1 PB Total System Memory
N8 data partition, 2015
2 PF system
.5 PB Total System Memory
BURST BUFFER 1
2 PB
NVRAM
2 TB/s
500 GB/s
1.5 TB/s
Global or Centerwide Filesystem(s)
O(100 PB) Disk
N9, 2019, new
300 PF system
10 PB
BURST BUFFER 2
40 PB
NVRAM
20 TB/s
1 TB/s
100-200 GB/s
Archival
O(1 EB) Stored
Production Clusters
"Sponsored Compute"
Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN
2 x 100 Gb
1 x 1000 Gb
Software Defined Networking
ESnet
Energy Sciences Network
Scalable Data Services
SciDB MongoDB Data Transfer Nodes
Science Gateways, Web services, Sponsored storage, Data repository

What all this means to the storage user

- **Compute systems of 2016+ will have burst buffers, applications will need to consider various optional use cases**
  - Your current I/O workload should perform better, especially "really poor" I/O
  - Do I want node-local or shared BB?
  - How much do I need to request?
  - Do I want to use a custom API for improved I/O control (e.g. relaxing POSIX constraints)?
- **Expect metadata improvements (Lustre, HPSS)**
- **Storage systems will improve their accessibility (GPFS, HPSS) to your data**
  - Easier to understand locality of data and move data between different storage resources
  - Data movement between sites should continue improving

Thank you.