Planning for the Future of Data, Storage, and I/O at NERSC

Glenn K. Lockwood, Ph.D
Advanced Technologies Group

August 23, 2018
NERSC: the mission HPC facility
for the U.S. Department of Energy Office of Science

7,000 users
800 projects
700 applications
~2,000 publications per year

Simulations at scale

Experimental/observational
data analysis at scale
Cori – NERSC’s Cray XC-30

Compute
• 9,688 Intel KNL nodes
• 2,388 Intel Haswell nodes

Storage
• 30 PB, 700 GB/s scratch
  – Lustre (Cray ClusterStor)
  – 248 OSSes × 41 HDDs × 4 TB
  – 8+2 RAID6 declustered parity
• 1.8 PB, 1.5 TB/s burst buffer
  – Cray DataWarp
  – 288 BBNs × 4 SSDs × 1.6 TB
  – RAID0
NERSC’s storage hierarchy

- **1.8 PB**
- **1.5 TB/s**

- **30 PB**
- **700 GB/s**

- **12 PB**
- **130 GB/s**

- **140 PB**
More data, more problems

- Burst Buffer
- Scratch
- Campaign
- Archive

1.8 PB
1.5 TB/s

30 PB
700 GB/s

12 PB
130 GB/s
More data, more problems

Challenges:
- Inefficient pyramid
- New detectors, experiments
- Exascale == massive data
- HPC tech landscape changing
NERSC Data Strategy:

- Support large-scale data analysis on NERSC-9, NERSC-10
- Start initiatives to begin addressing today’s address pain points

Storage 2020:

- Define storage roadmap for NERSC
- Define architectures for milestones:
  - 2020: NERSC-9 deployment
  - 2025: NERSC-10 deployment
NERSC’s approach to strategic planning

User Requirements

Workload Analysis

Technology Trends

NERSC Strategy
User requirements: Workflows

Survey findings:
- Data re-use is uncommon
- Significant % of working set must be retained forever

Insight:
- Read-caching burst buffers require prefetching
- Need large archive
- Need to efficiently move data from working space to archive

APEX workflows white paper - https://www.nersc.gov/assets/apex-workflows-v2.pdf
Large working sets: “Storage requirements are likely to be large; they are already at the level of 10 PB of disk storage, and they are likely to easily exceed 100 PB by 2025.” (HEP)

High ingest rates: “Next generation detectors will double or quadruple these rates in the near term, and rates of 100 GB/sec will be routine in the next decade.” (BER)
**Workload analysis:**

**Read/write ratios**

**Burst Buffer:** 4:6  
**Scratch:** 7:5  
**Archive:** 4:6  

- Checkpoint/restart is not the whole picture  
- Read performance is very important!
Workload analysis: File interactions

File size distribution on project

Metadata ops issued in a year to scratch

- open, 41.1%
- close, 30.6%
- stat, 25.9%
- unlink, 2.27%
- rmdir, 0.0431%
- link, 0.00861%

U.S. DEPARTMENT OF
Energy
Office of Science
- 12 -
Technology trends: Tape

- Industry is consolidating
- Revenue is shrinking
- Tape advancements are driven by profits, not tech!
  - Re-use innovations in HDD
  - Trail HDD bit density by 10 yr
- Refresh cadence will slow
- $/GB will no longer keep up with data growth

LTO market trends; Fontana & Decad, MSST 2016
Technology trends: Tape

- Industry is consolidating
- Revenue is shrinking
- Tape advances are driven by profits, not tech!
  - Re-tape innovations in HDD
  - Trail HDD bit density by 10x yr
- Refresh cadence will slow
- $/GB will no longer keep up with data growth

NERSC’s archive cannot grow as fast as it has historically!
Technology trends: Magnetic disk

- Bit density increases slowly (10%/yr)
  \[ \frac{MB}{s} \sim \sqrt{\frac{bits}{in^2}} \]
- HDDs for capacity, not performance
Technology trends: Magnetic disk

- Bit density increases slowly (10%/yr)

\[
\frac{MB}{s} \sim \sqrt{\text{bits}}
\]

- HDDs for capacity, not performance

NERSC will not rely on HDDs for performance tiers!
Technology trends: Flash

- NAND $/GB dropping fast
- O($0.15/GB) by 2020
- Performance limited by PCIe and power
- $/GB varies with optimization point

Actuals from Fontana & Decad, Adv. Phys. 2018
Technology trends: Flash

- NAND $/GB dropping fast by 2020
- Performance limited by PCIe and power
- $/GB varies with optimization point

Expect easier performance, more data movement between tiers
Technology trends: Exascale computing

- **Exascale**: power-efficient cores *everywhere* for parallel throughput
- **File-based (POSIX) I/O**: fast cores for serial latency

Exascale will struggle to deliver high-performance, POSIX-compliant file I/O!

**POSIX I/O**

**Exascale @ < 40 MW**

**3.2.2 CORAL System Peak (TR-1)**

The CORAL baseline system performance will be at least 1,300 petaFLOPS (1300x10^{15} double-precision floating point operations per second).

**3.2.5 Maximum Power Consumption (TR-1)**

The maximum power consumed by the 2021 or 2022 CORAL system and its peripheral systems, including the proposed storage system, will not exceed 40MW, with power consumption between 20MW to 30MW preferred.
NERSC roadmap: Design goals

- **Target 2020**
  - Collapse burst buffer and scratch into all-flash scratch
  - Invest in large disk tier for capacity
  - Long-term investment in tape to minimize overall costs

- **Target 2025**
  - Use single namespace to manage tiers of SCM and flash for scratch
  - Use single namespace to manage tiers of disk and tape for long-term repository
NERSC roadmap: Implementation

All-flash parallel file system on NERSC-9

> 150 PB disk-based file system

> 350 PB HPSS archive w/ IBM TS4500 +Integrated Cooling

Performance object store w/ SCM+NAND on NERSC-10

Archival object store w/ HDD+tape (GHI+HPSS? Versity? Others?)
NERSC-9: A 2020, pre-exascale machine

- 3-4x capability of Cori
- Optimized for both
  - large-scale simulation
  - large-scale experimental data analysis
- Onramp to Exascale:
  - heterogeneity
  - specialization

CPUs
Broad HPC workload

Accelerators
Image analysis, Machine learning, Simulation

All-Flash Storage
High bandwidth, High(er) IOPS, Better metadata

Flexible Interconnect

Remote data can stream directly into system
Can integrate FPGAs and other accelerators
Two classes of object storage:

- **Hot archive:**
  - a. Driven by cloud
  - b. Most mature
  - c. Low barrier to entry

Object stores trade convenience for scalability

Performance, \((\text{familiarity})^{-1}\) vs. GB/\$, Durability
Two classes of object storage for science:

- **Hot archive:**
  - a. Driven by cloud
  - b. Most mature
  - c. Low barrier to entry

- **Performance:**
  - a. Driven by Exascale
  - b. Delivers performance of SCM
  - c. High barrier (usability mismatch)

Object stores trade convenience for scalability.

Performance, \((\text{familiarity})^{-1}\)

GB/$, Durability
NERSC’s object store transition plan

2020: new object APIs atop familiar file-based storage
  ○ Spectrum Scale Object Store
  ○ HPSS on Swift

2025: replace file store with object store
  ○ Both object and POSIX APIs still work!
  ○ Avoid forklift of all data
  ○ POSIX becomes middleware
Further reading:

• Storage 2020 report: https://escholarship.org/uc/item/744479dp

• Bhimji et al., “Enabling production HEP workflows on Supercomputers at NERSC” https://indico.cern.ch/event/587955/contributions/2937411/

• Stay tuned for more information on NERSC-9 around SC’18!