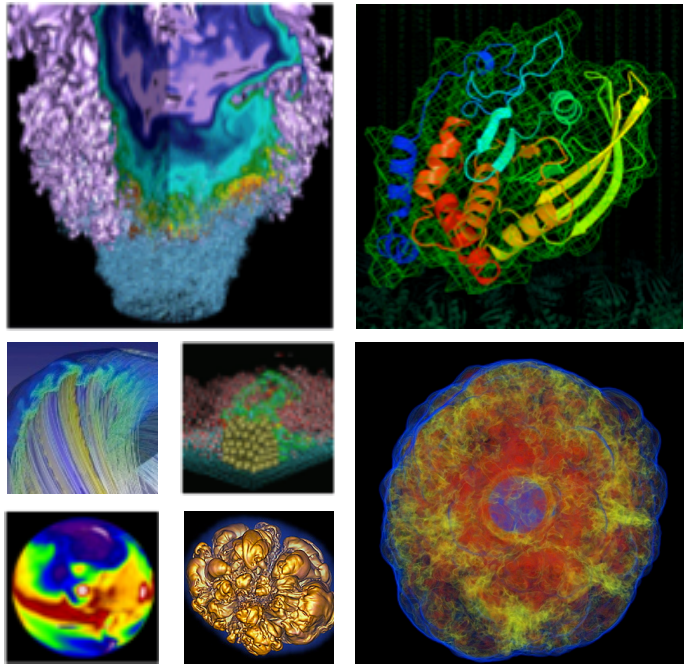
Planning for the Future of Storage for HPC: 2020, 2025, and Beyond

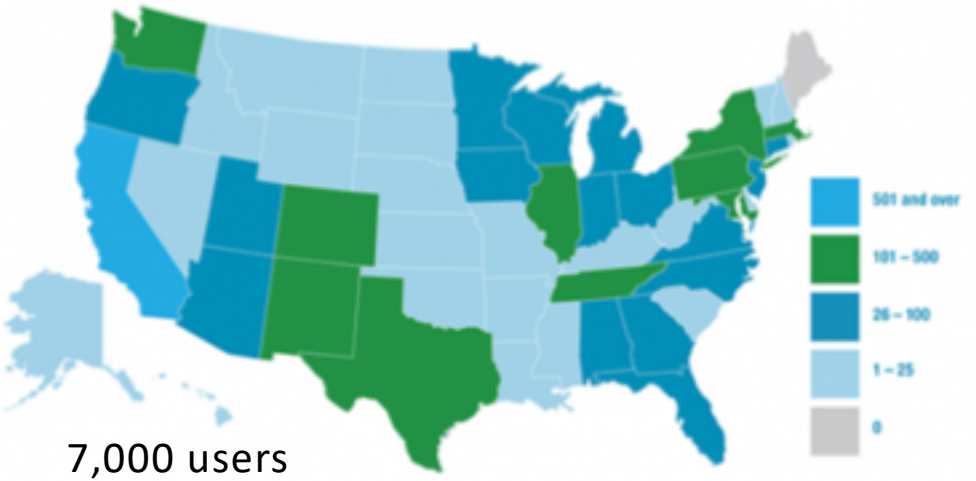


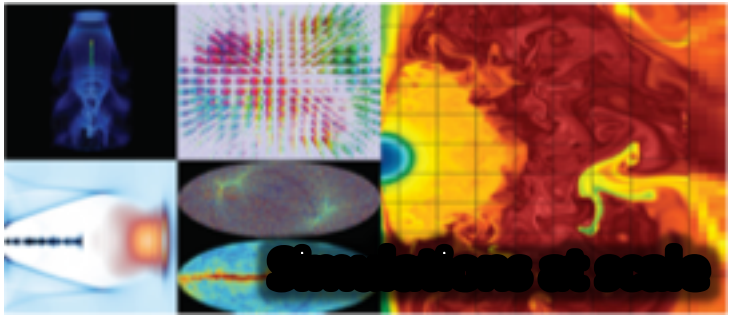
Glenn K. Lockwood, Ph.D.
+ 28 others

June 28, 2018

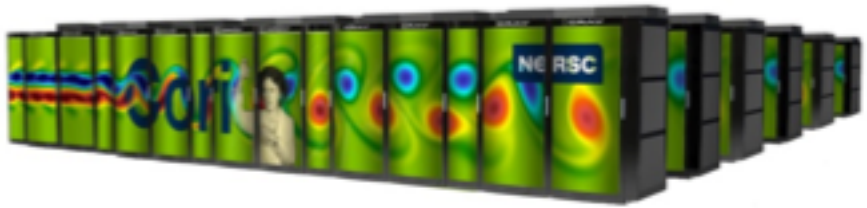
NERSC: the mission HPC facility

for the U.S. Department of Energy Office of Science





7,000 users
800 projects
700 applications
~2000 publications per year



U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Cori – NERSC's Cray XC-30 (2015)

Compute

- **9,688 Intel KNL nodes**
- **2,388 Intel Haswell nodes**

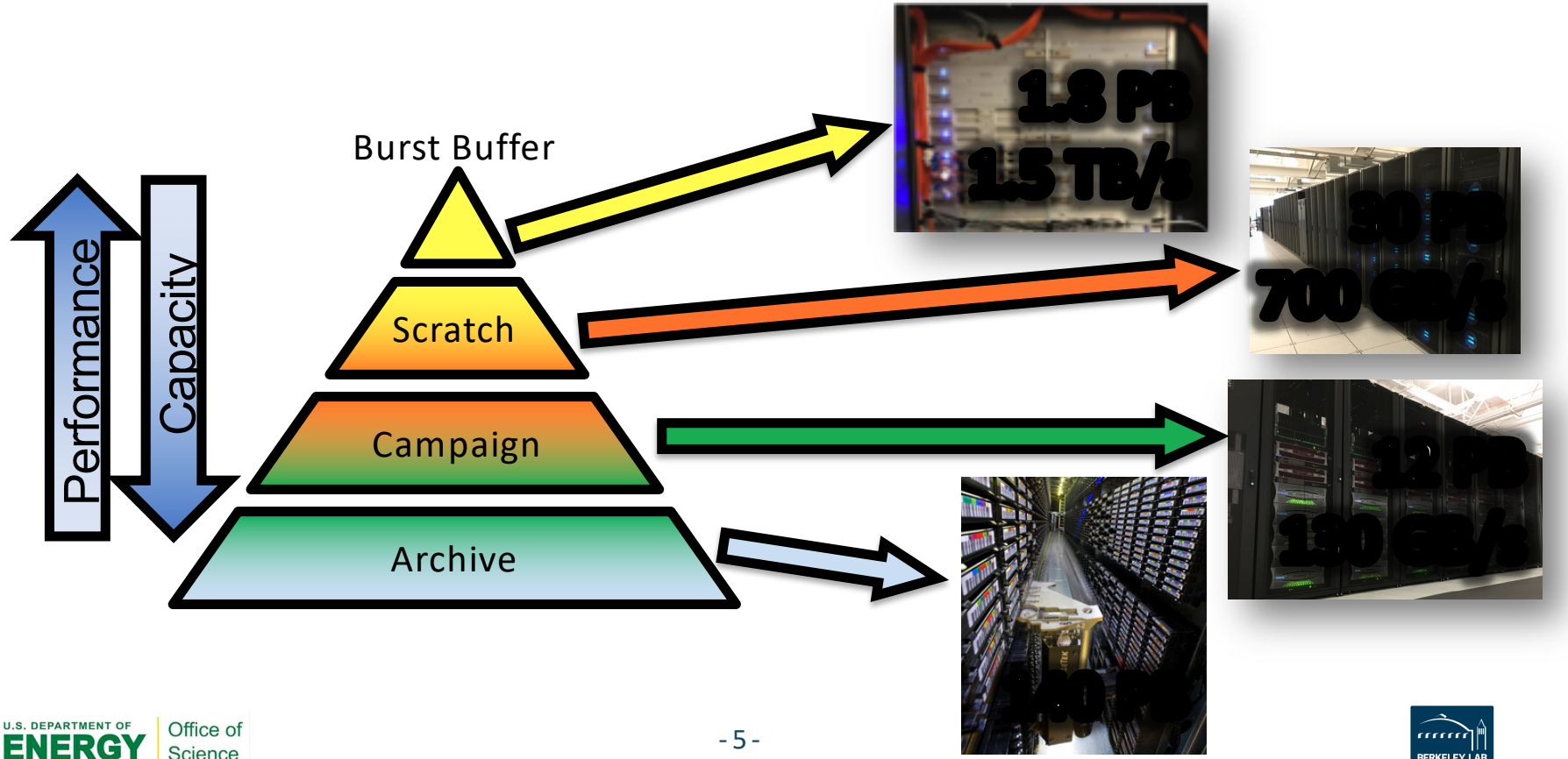
Storage

- **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

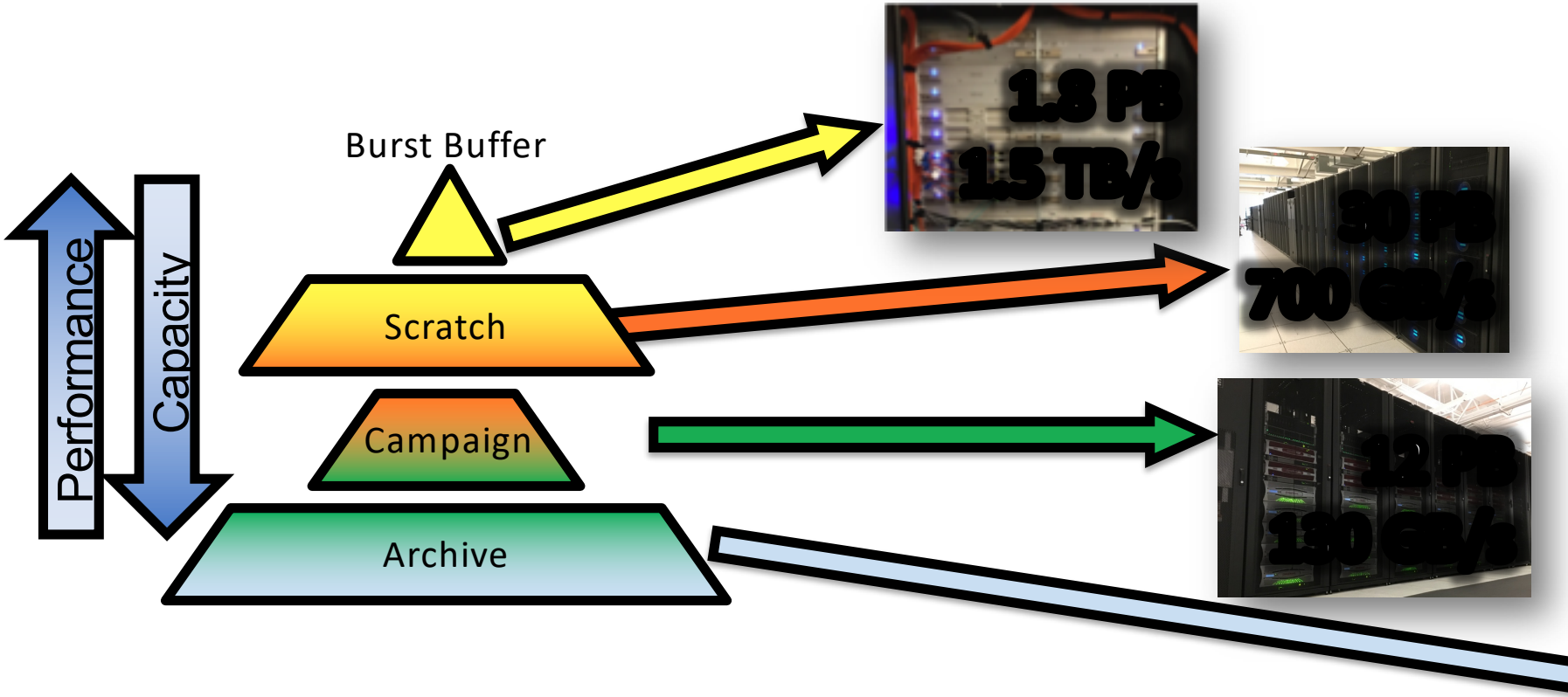
Wang Hall – NERSC's new data center (2015)

- **12.5 MW** in 2015
- **20 MW** planned for 2020
- **1,850 m²** floor space
- **PUE < 1.1**
- **Free cooling – no chillers!**
  - 4x 3.4 MW cooling towers
  - < 25°C water to racks
  - 4x ~0.5 MW air handlers
  - 30% - 70% humidity

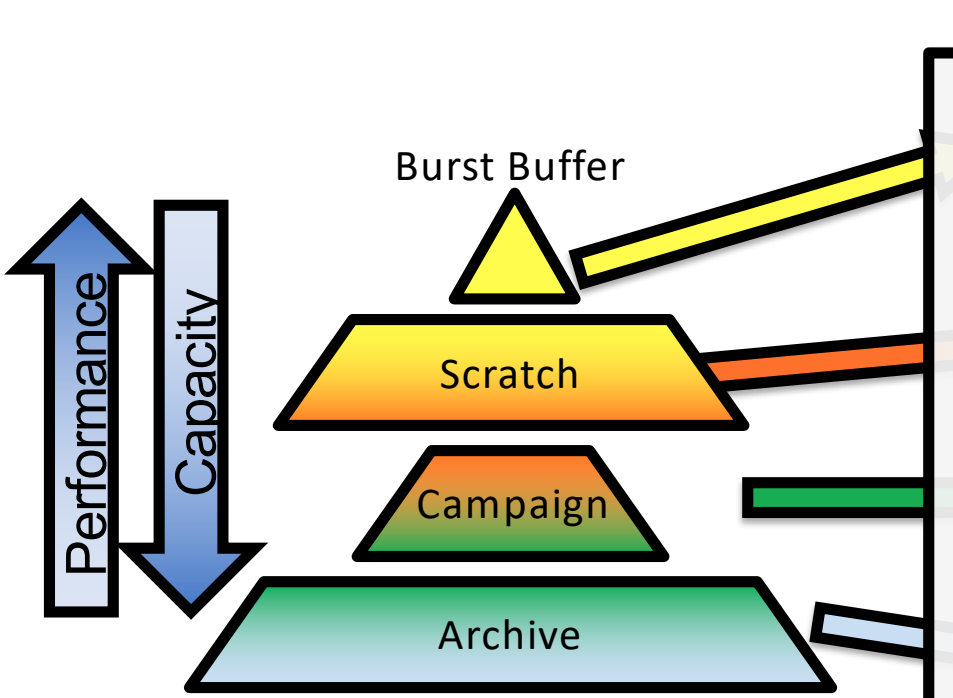
NERSC's storage infrastructure (2015)



More capabilities, more problems



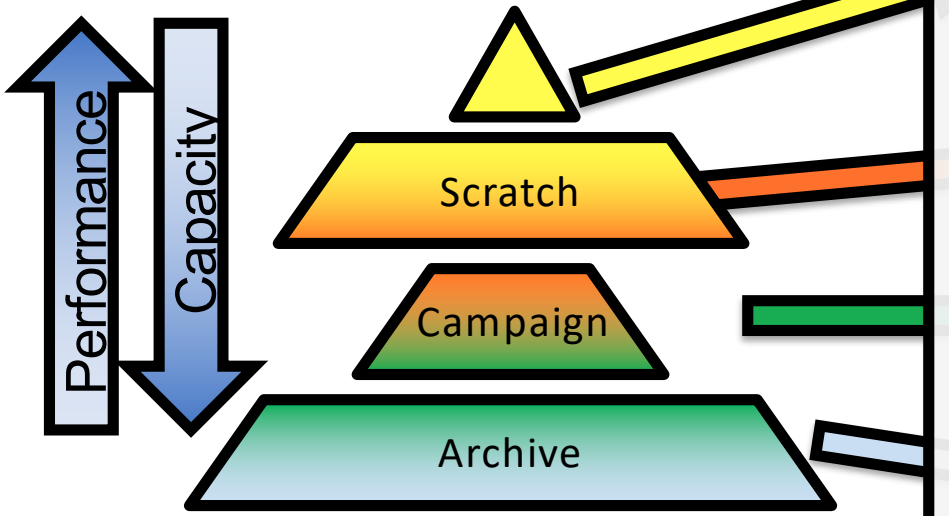
More capabilities, more problems



Tactical challenges

- Partly inverted pyramid
- Wang Hall could not host tape archives
- Tape technology was discontinued by vendor

More capabilities, more problems



Strategic challenges

- Exascale problems are massive data problems
- New instruments have huge data analysis needs
- Storage tiers are hard to use and manage
- Hardware landscape is rapidly changing

Storage 2020: NERSC's strategic plan for storage

Storage 2020 goals:

- **Define roadmap for NERSC storage infrastructure for the next ten years**
- **Define architectures for specific NERSC milestones:**
  - 2020: NERSC-9 deployment
  - 2025: NERSC-10 deployment
- **Define blueprint for HPC storage planning**

NERSC's approach to strategic planning

User Requirements

Workload Analysis

Technology Trends

NERSC Strategy

User requirements: Workflows

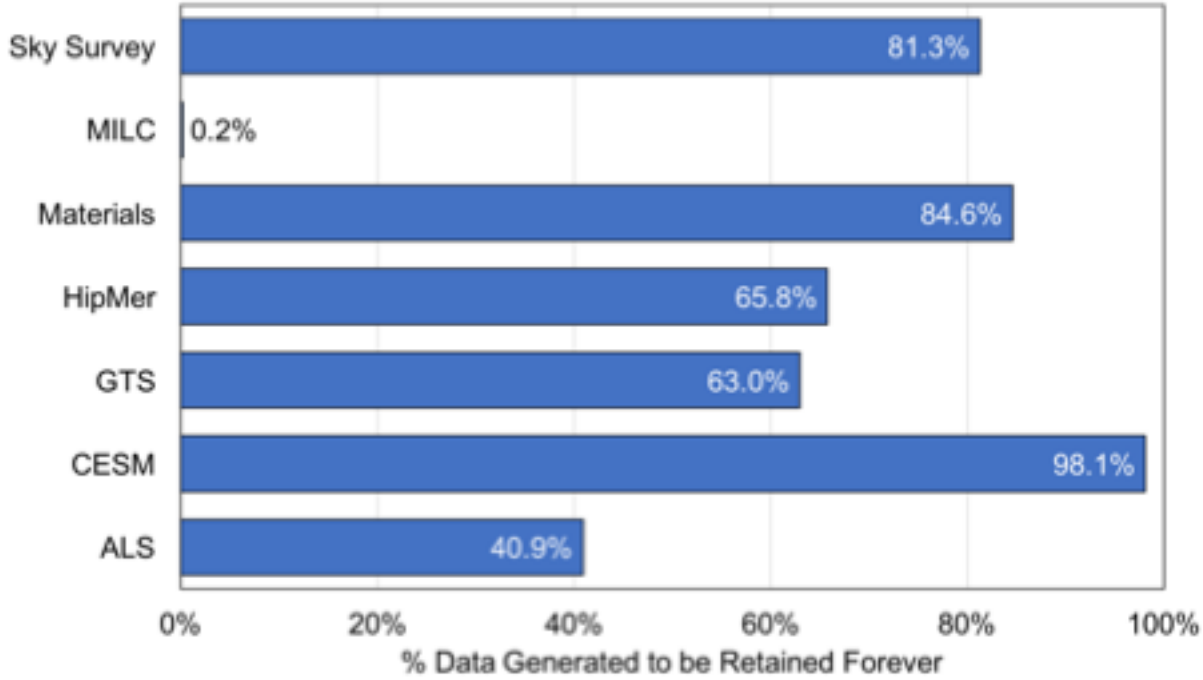
Survey findings:

- 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



User requirements: Exascale

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." (BES)

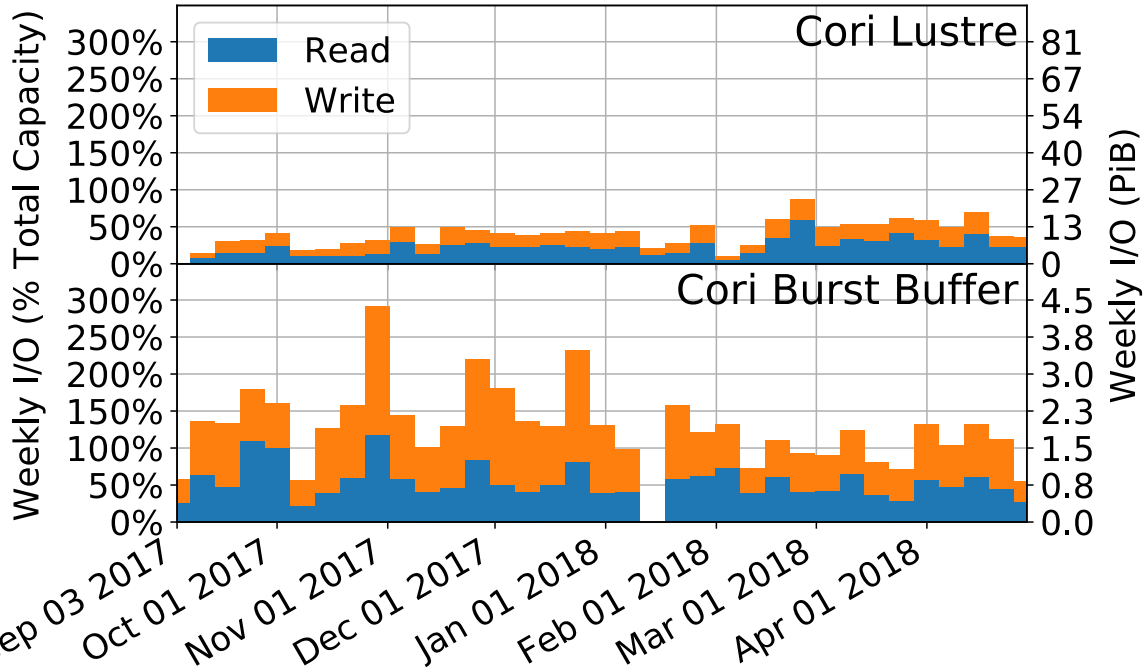
Workload analysis: Read/write ratios

Burst Buffer: 4:6

Scratch: 7:5

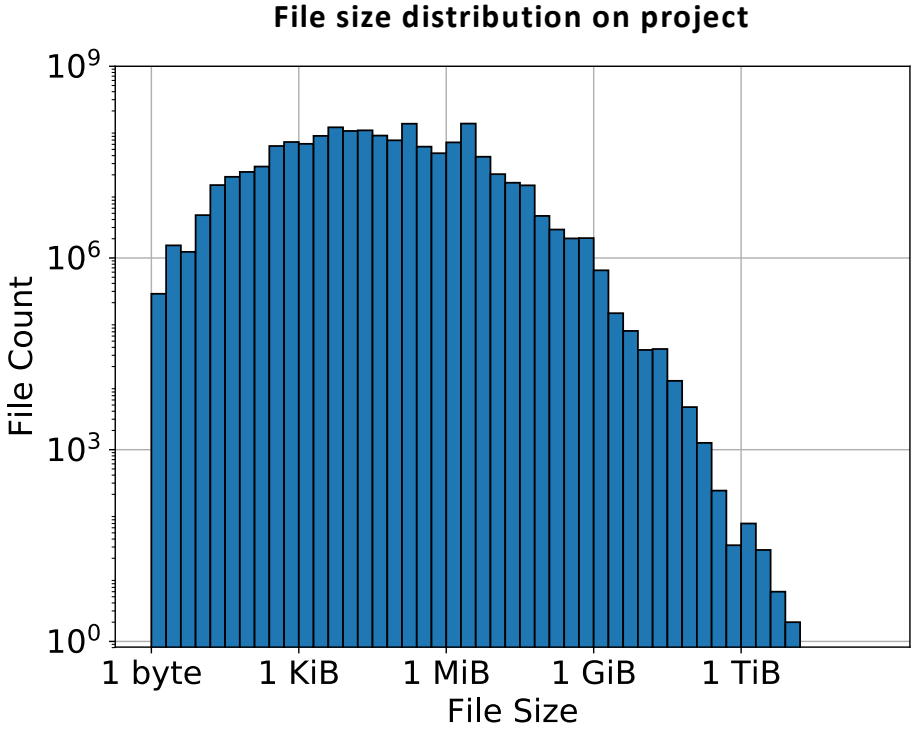
Archive: 4:6

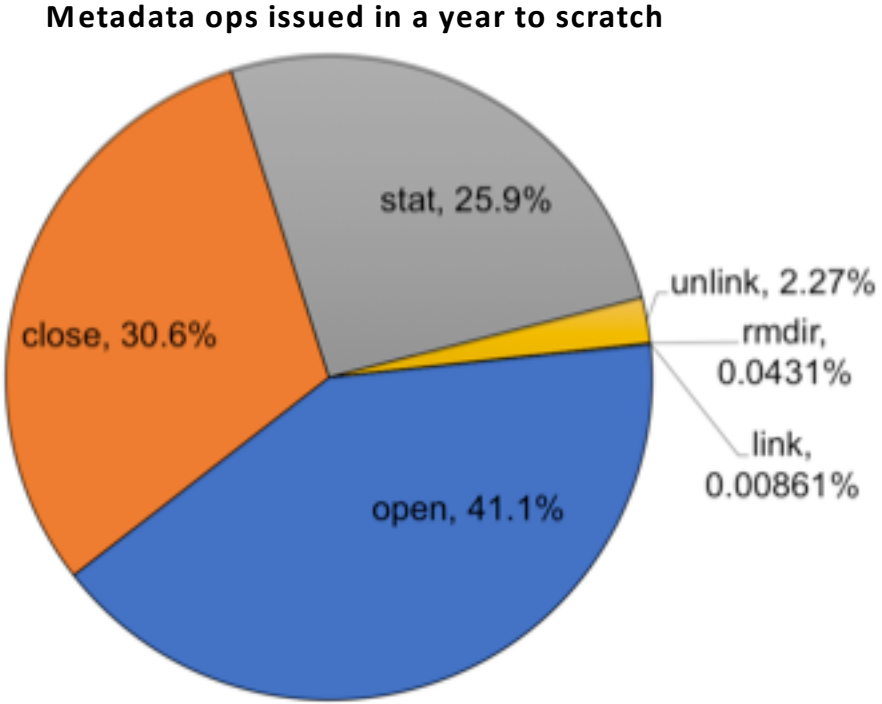
- Checkpoint/restart is not the whole picture
- Read performance is very important!



Lockwood et al., *TOKIO for ClusterStor*. Proceedings of 2018 Cray User Group. Stockholm, SE. May 2018.

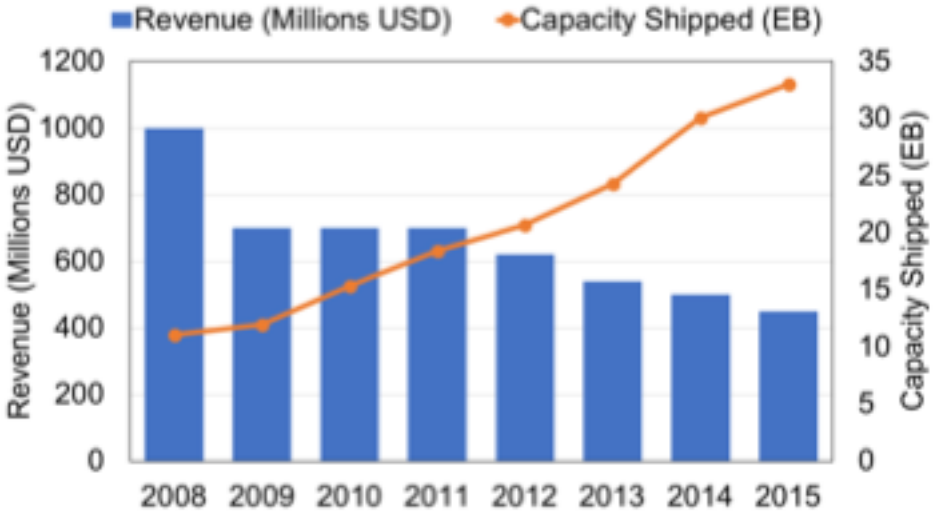
Workload analysis: File interactions





Technology trends: tape

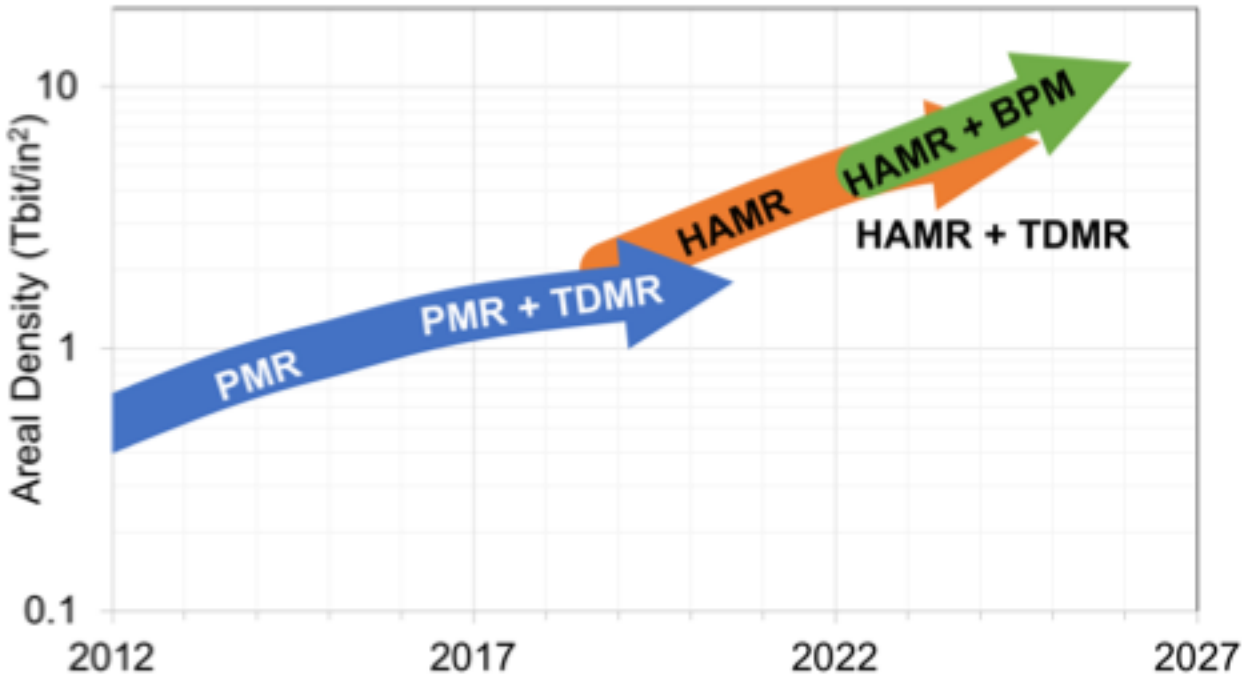
- **Industry is consolidating**
- **Revenue is shrinking**
- **Tape advancements are driven by profits, not tech!**
  - Re-use innovations in HDD
  - Trail HDD bit density by 10 yr
- **Refresh cadence will slow**
- **$/GB will no longer keep up with data growth**



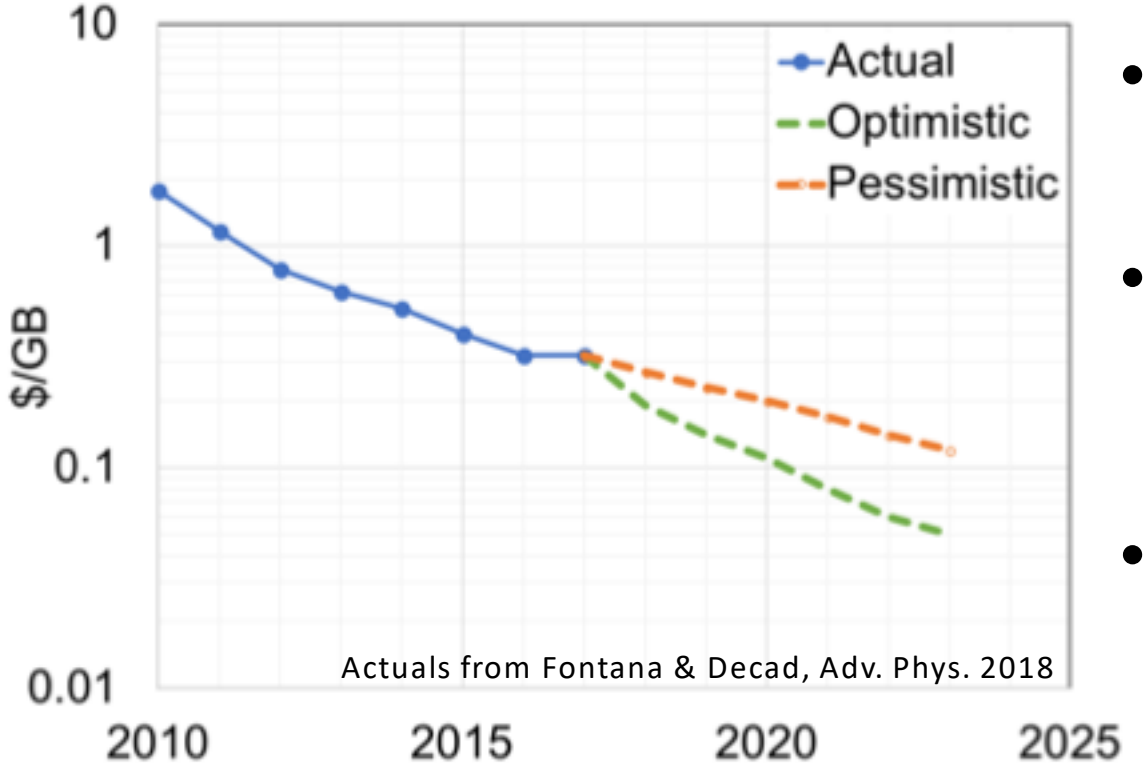
LTO market trends; Fontana & Decad, MSST 2016

Technology trends: magnetic disk

- Bit density increases slowly (10%/yr)
- $\frac{MB}{s} \sim \sqrt{bits/in^2}$
- HDDs for capacity, not performance

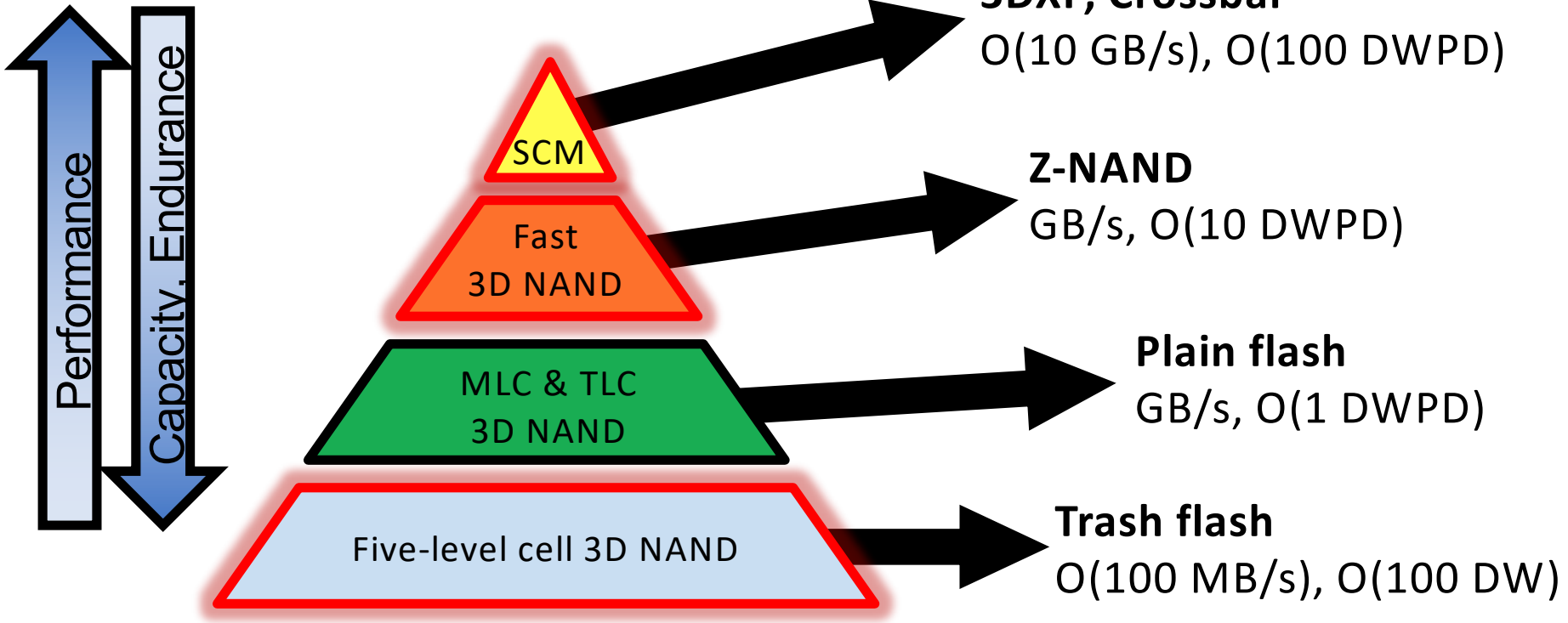


Technology trends: flash



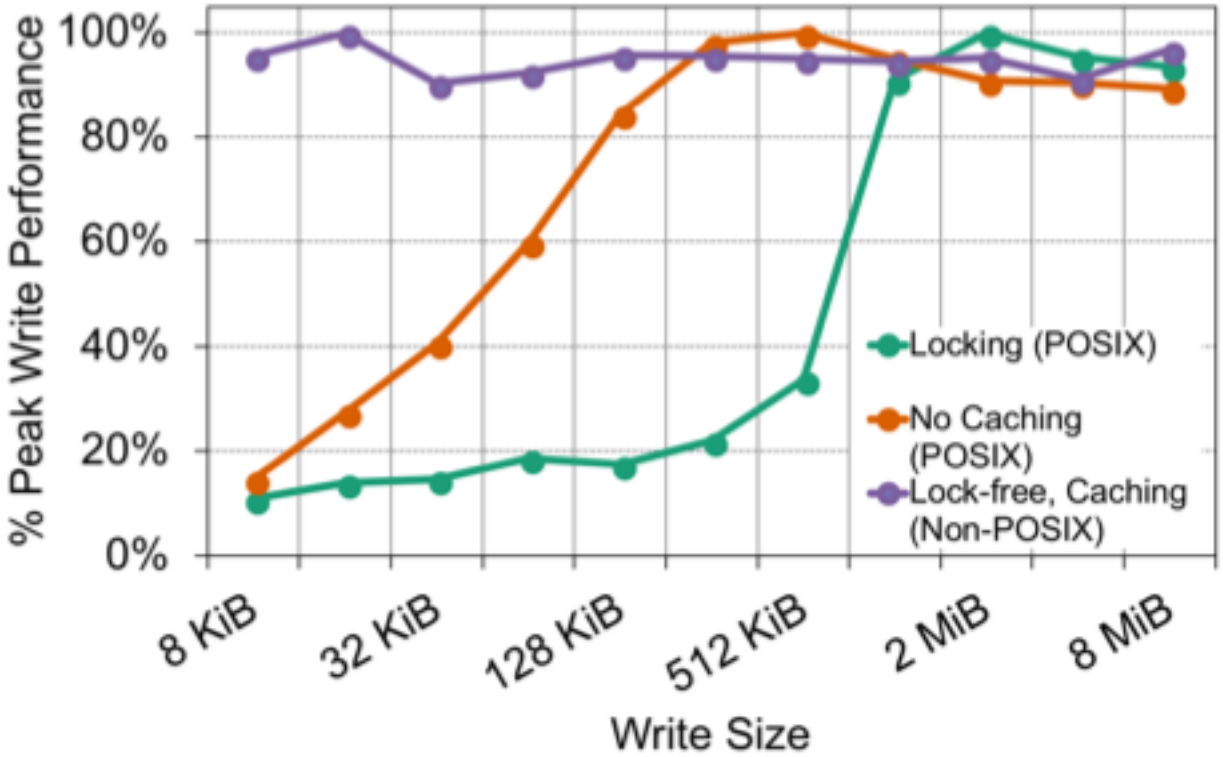
- NAND $/GB dropping fast
- Performance limited by PCIe and power
- O($0.15/GB) by 2020

Technology trends: flash



Technology trends: software

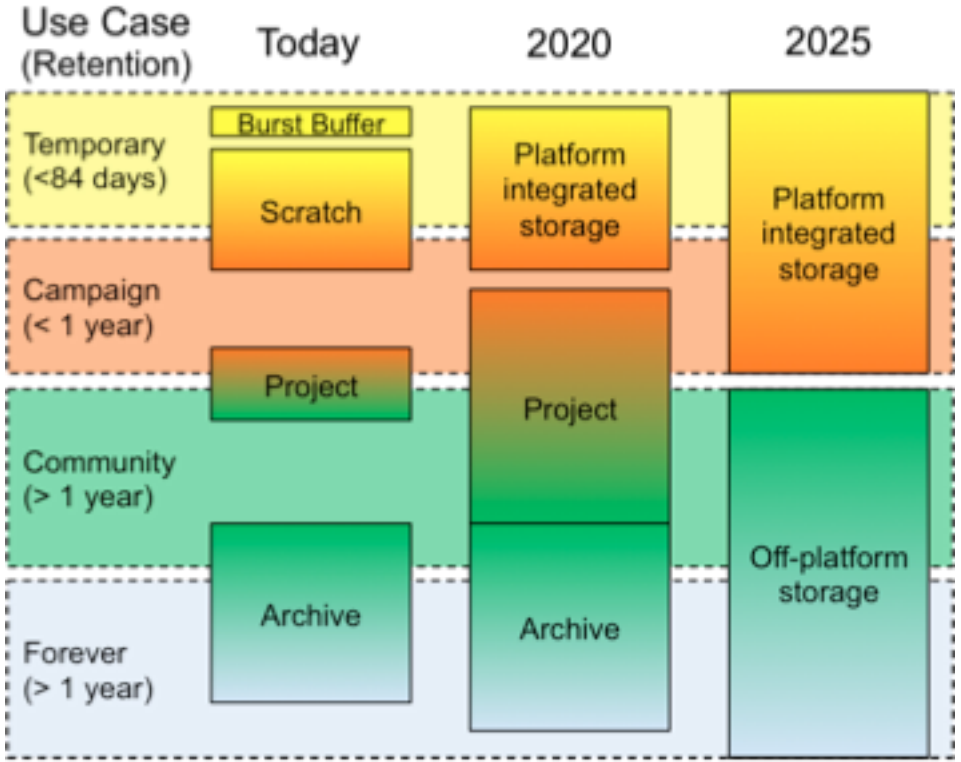
- Workloads demand perf at small I/O sizes
- Don't pay for POSIX consistency
- Object stores are a flexible foundation



Design goals derived from requirements

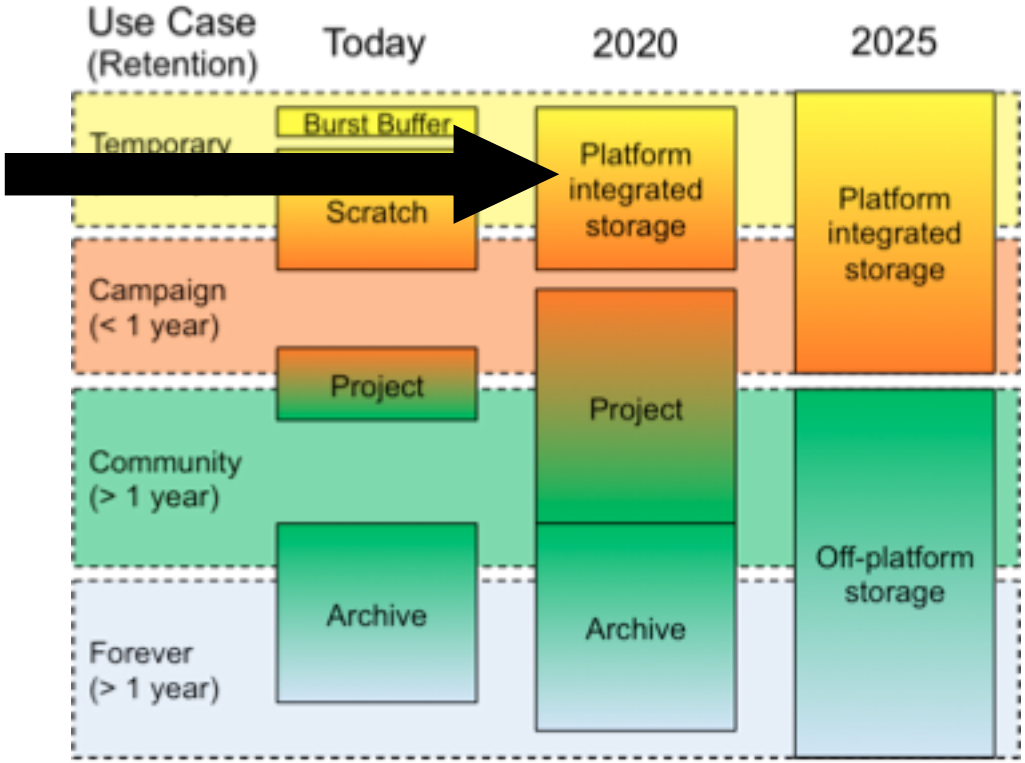
- **Tiering and data movement**
  - Fix the storage pyramid
  - Collapse tiers for simplicity
- **POSIX and file systems**
  - Maintain POSIX API support
  - ...but look beyond POSIX semantics
- **Hardware**
  - Capitalize on falling costs of flash
  - Tape will not be business as usual

NERSC roadmap – 2015, 2020, and 2025



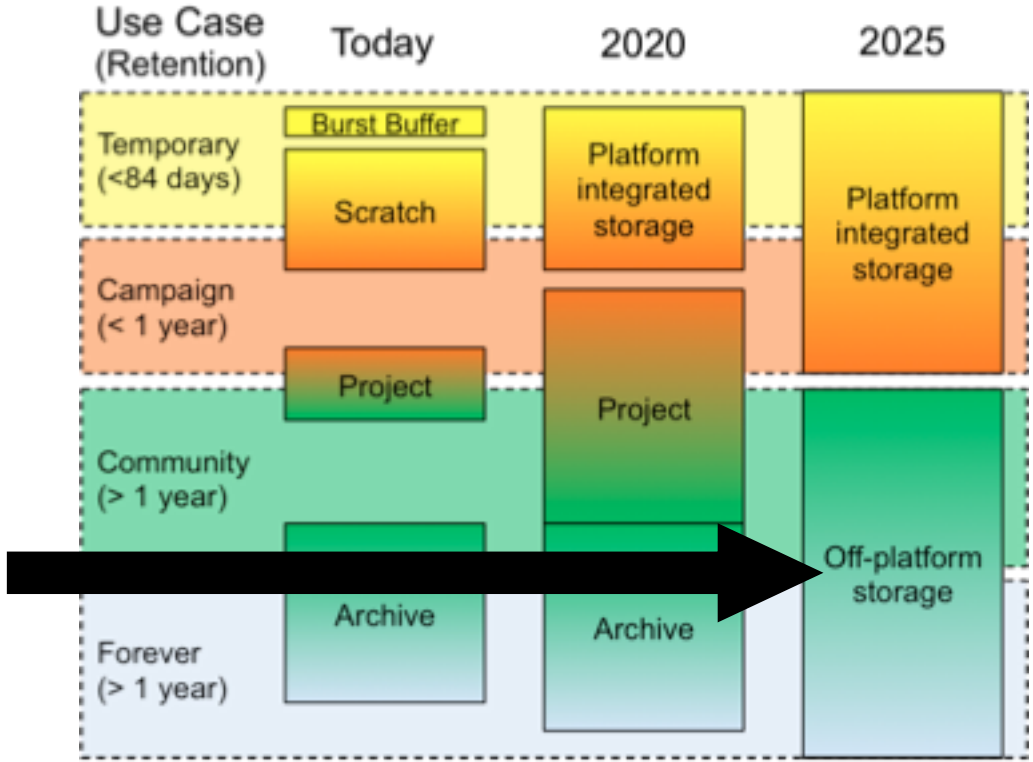
NERSC roadmap – Tiering and data movement

- **Target 2020**
  - Collapse burst buffer and scratch into a single logical tier



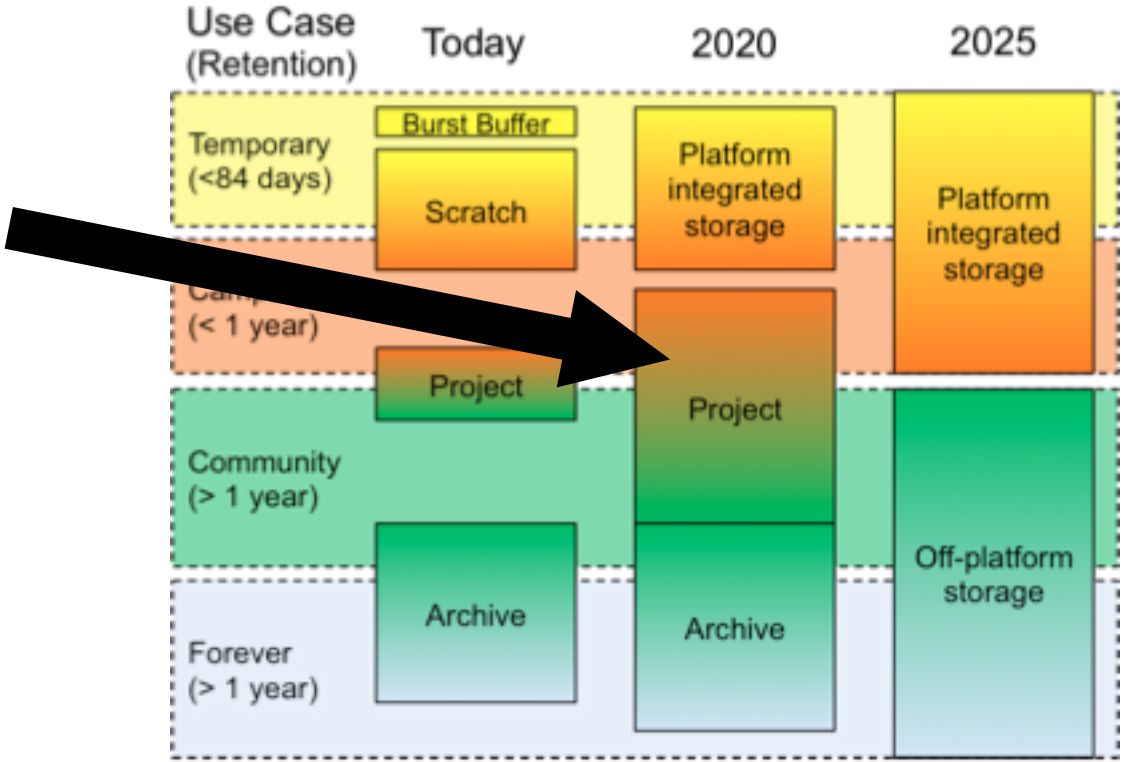
NERSC roadmap – Tiering and data movement

- **Target 2020**
  - Collapse burst buffer and scratch into a single logical tier
- **Target 2025**
  - Collapse longer-term disk and tape into a single logical tier



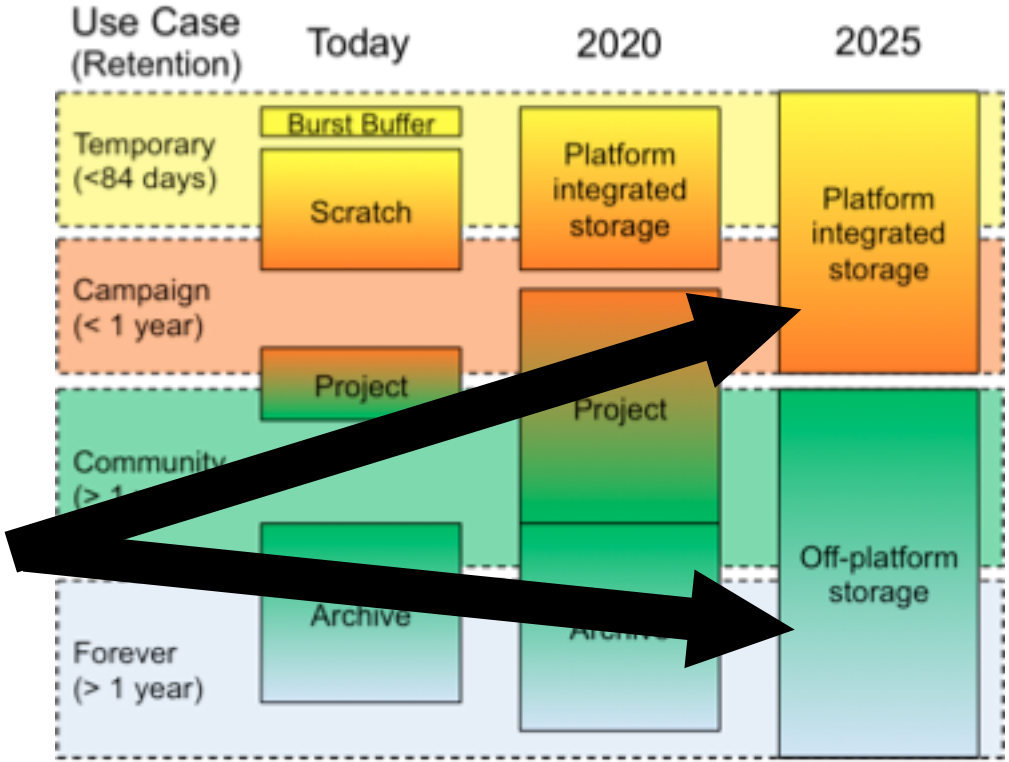
NERSC roadmap – POSIX and file systems

- **Target 2020**
  - Provide object interfaces to file-based storage systems



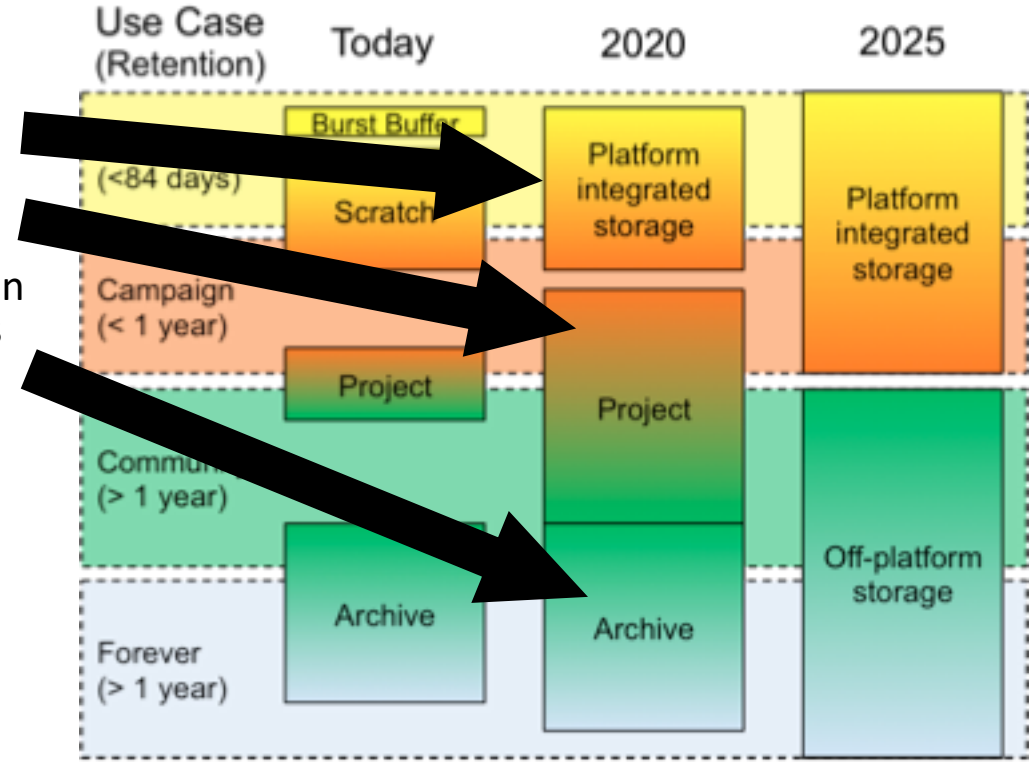
NERSC roadmap – POSIX and file systems

- **Target 2020**
  - Provide object interfaces to file-based storage systems
- **Target 2025**
  - Provide file interfaces into object-based storage systems



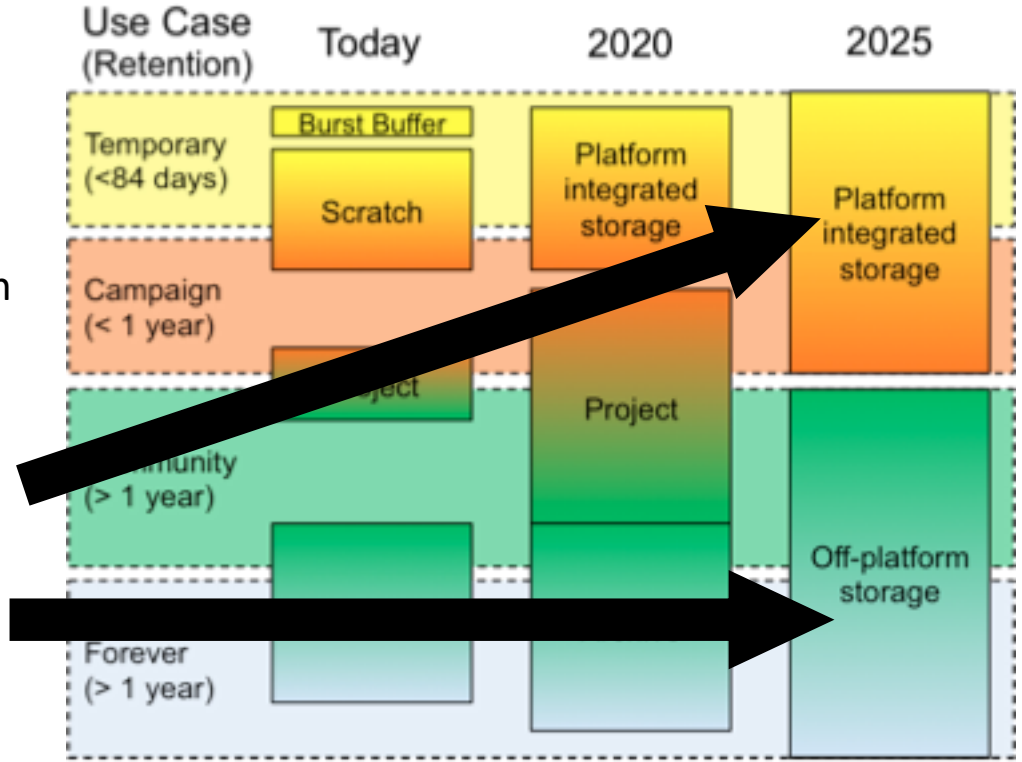
NERSC roadmap – Hardware

- **Target 2020**
  - Collapse burst buffer and scratch into all-flash scratch
  - Invest in large disk tier for capacity
  - Tape isn't dead yet, so invest in it to minimize long-term costs



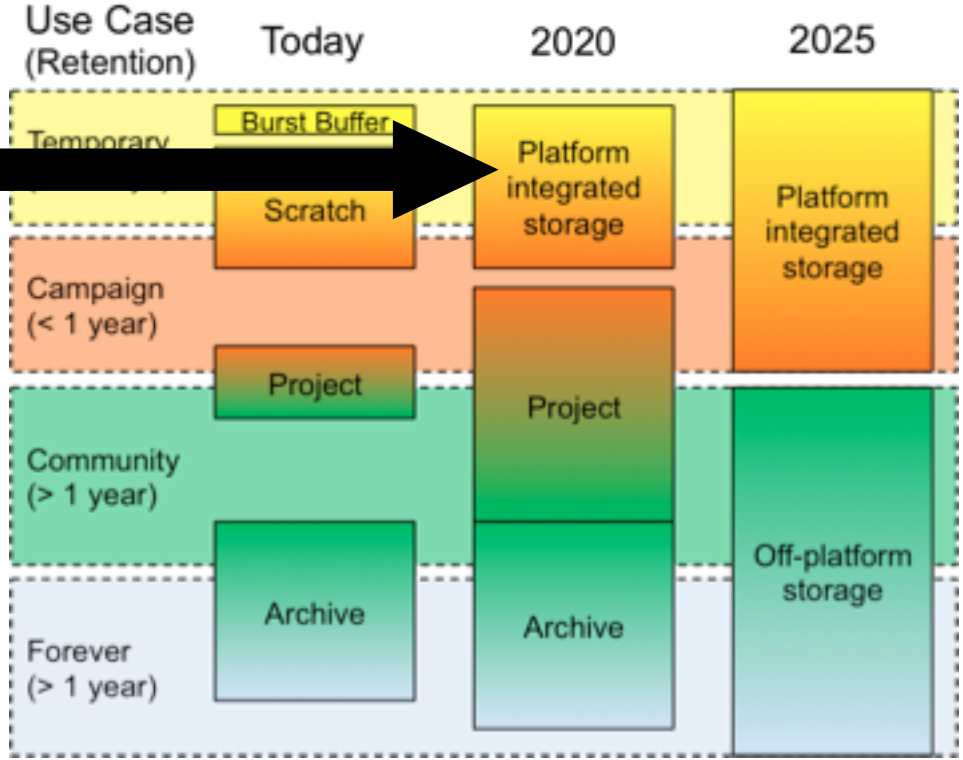
NERSC roadmap – Hardware

- **Target 2020**
  - Collapse burst buffer and scratch into all-flash scratch
  - Invest in large disk tier for capacity
  - Tape isn't dead yet, so invest in it to minimize long-term 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 – 2020, 2025, and beyond

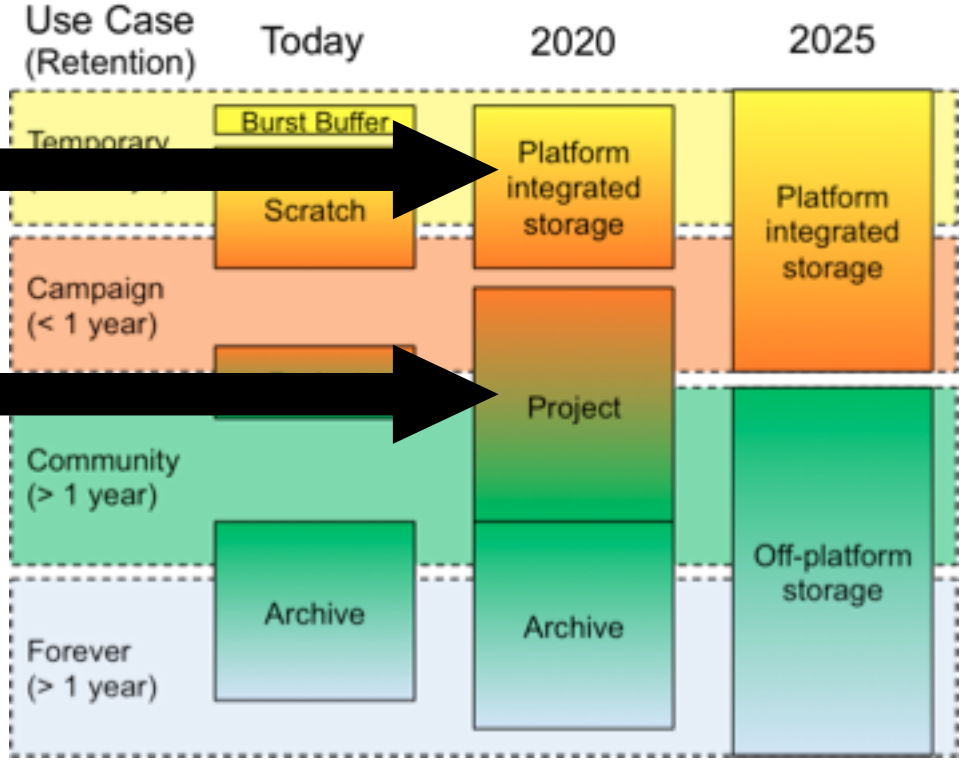
All-flash parallel file system feasible for NERSC-9



NERSC roadmap – 2020, 2025, and beyond

All-flash parallel file system feasible for NERSC-9

> 100 PB disk-based file system



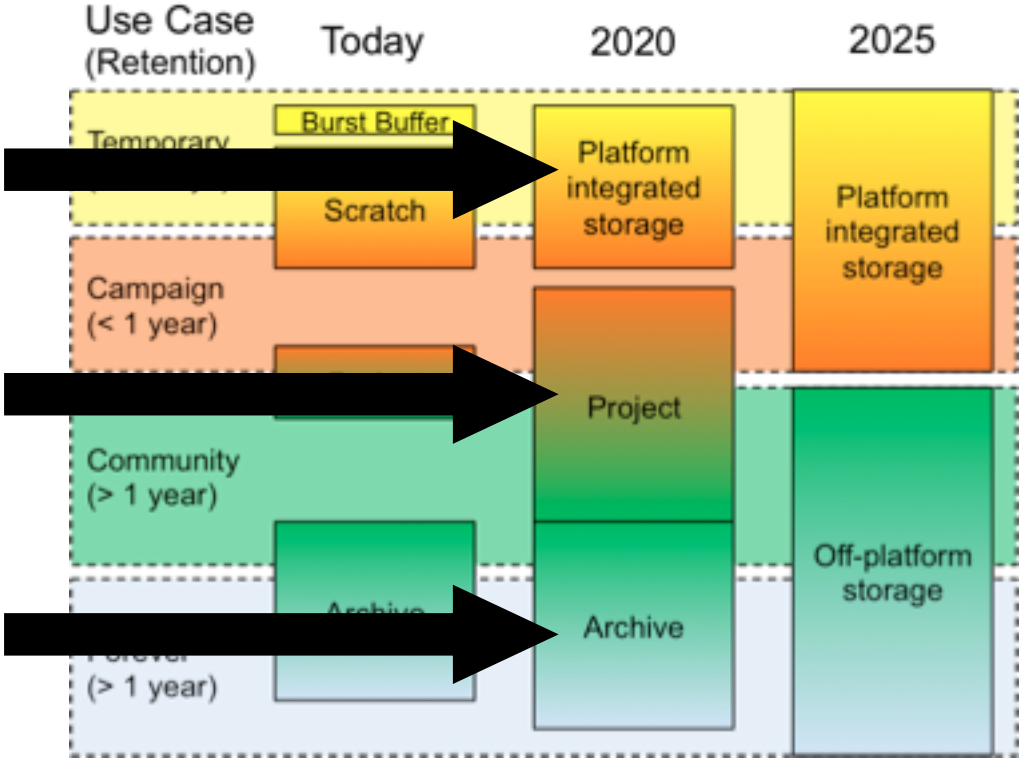
NERSC roadmap – 2020, 2025, and beyond

All-flash parallel file system feasible for NERSC-9

> 100 PB disk-based file system

> 350 PB HPSS archive w/ IBM TS4500

+Integrated Cooling



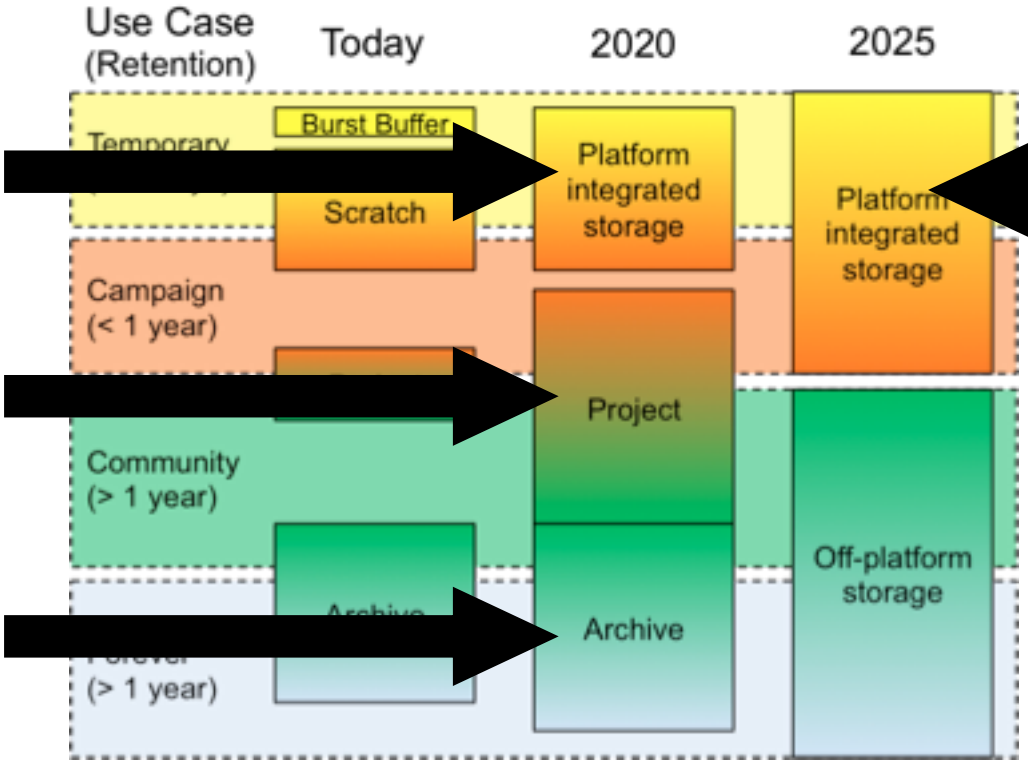
NERSC roadmap – 2020, 2025, and beyond

All-flash parallel file system feasible for NERSC-9

> 100 PB disk-based file system

> 350 PB HPSS archive w/ IBM TS4500

+Integrated Cooling



Performance object store w/ SCM+NAND on NERSC-10 (DAOS? Mero? Others?)

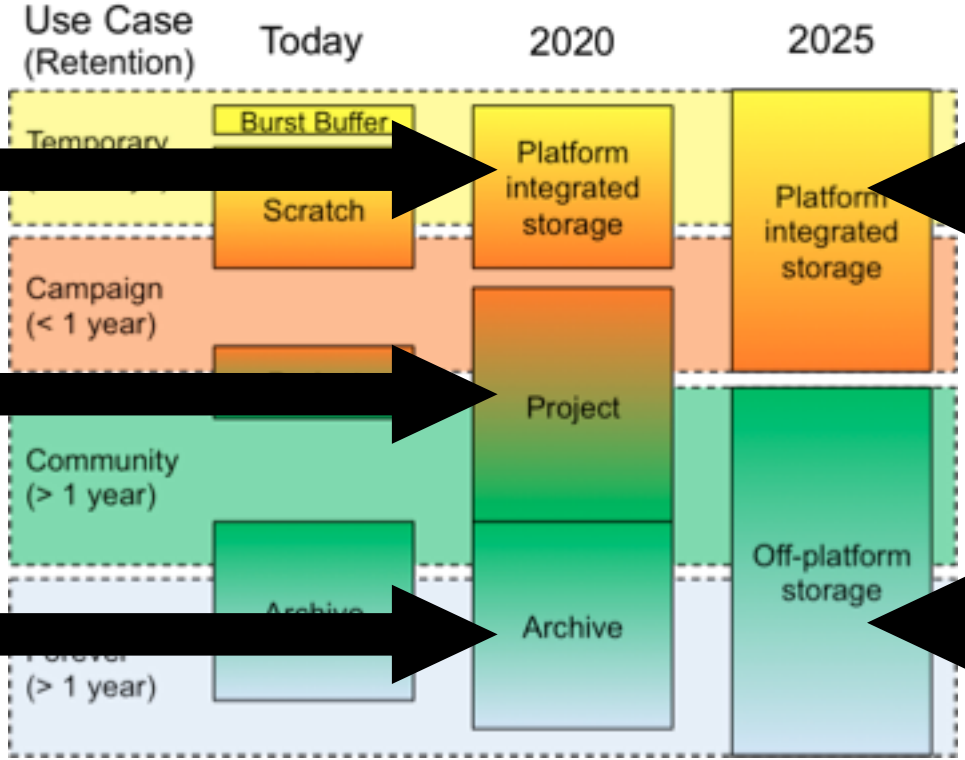
NERSC roadmap – 2020, 2025, and beyond

All-flash parallel file system feasible for NERSC-9

> 100 PB disk-based file system

> 350 PB HPSS archive w/ IBM TS4500

+Integrated Cooling



Performance object store w/ SCM+NAND on NERSC-10 (DAOS? Mero? Others?)

Archival object store w/ HDD+tape (GHI+HPSS? Versity? Others?)

Unresolved issues requiring investment

- **Automatic data movement**
  - Need better metadata capabilities to inform policy
  - Need better software integration between tiers
- **Moving beyond POSIX**
  - Only implement the parts of POSIX people use
  - Limit unnecessary performance penalties (locking, etc)
- **Everything must be tunable to workload of the center**

Following the Storage 2020 blueprint

- **36-page Storage 2020 report online: https://escholarship.org/uc/item/744479dp**
- **Be inclusive and get all staff involved!**
- **Keep up with industry & research**
  - HPC-IODC
  - Massive Storage Systems and Technology conference
  - Flash Memory Summit
  - Large Tape User Group
  - PDSW-DISCS

Thank you!