Storage 2020:
A Vision for the Future of HPC Storage

Glenn K. Lockwood, Damian Hazen, Quincey Koziol, Shane Canon, Katie Antypas, Jan Balewski, Nicholas Balthaser, Wahid Bhimji, James Botts, Jeff Broughton, Tina L. Butler, Gregory F. Butler, Ravi Cheema, Christopher Daley, Tina Declerck, Lisa Gerhardt, Wayne E. Hurlbert, Kristy A. Kallback-Rose, Stephen Leak, Jason Lee, Rei Lee, Jialin Liu, Kirill Lozinskiy, David Paul, Prabhat, Cory Snavely, Jay Srinivasan, Tavia Stone Gibbins, Nicholas J. Wright

National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

Report No. LBNL-2001072

November 2017

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Table of Contents

1. Introduction **6**

2. NERSC Storage Hierarchy **7**
- **2.1. Current Storage Infrastructure at NERSC** **7**
- **2.2. Workflow-based Model for Storage** **8**

3. Requirements **11**
- **3.1. Current I/O Patterns** **11**
- **3.2. NERSC-9 Requirements** **14**
- **3.3. DOE Exascale Requirements Reviews** **15**
- **3.4. Emerging Applications and Use Cases** **16**
- **3.5. Operational Requirements** **17**
  - 3.5.1. Reliability, Durability, Longevity, and Disaster Recovery 17
  - 3.5.2. Space management and curation features 18
  - 3.5.3. Availability 18
- **3.6. Gaps and Challenges** **19**
  - 3.6.1. Tiering 19
  - 3.6.2. Data Movement 19
  - 3.6.3. Data Curation 19
  - 3.6.4. Workload Diversity 20
  - 3.6.5. Storage System Software 20
  - 3.6.6. Hardware Concerns 21
  - 3.6.7. POSIX and Middleware 21

4. Technology Landscape and Trends **21**
- **4.1. Hardware** **21**
  - 4.1.1. Magnetic Disk 22
  - 4.1.2. Solid-State Storage 23
  - 4.1.3. Storage Class Memory and Nonvolatile RAM 24
  - 4.1.4. Magnetic Tape 25
  - 4.1.5. Storage System Design 26
- **4.2. Software** **27**
  - 4.2.1. Non-POSIX Storage System Software 27
  - 4.2.2. Application Interfaces and Middleware 28

5. Next Steps **28**
- **5.1. Vision for the Future** **28**
- **5.2. Strategy** **30**
  - 5.2.1. Near Term (2017 – 2020) 30
  - 5.2.2. Long Term (2020 – 2025) 32
  - 5.2.3. Opportunities to Innovate and Contribute 34

6. Conclusion **36**

Executive Summary

The explosive growth in data over the next five years that will accompany exascale simulations and new experimental detectors will enable new data-driven science across virtually every domain. At the same time, new nonvolatile storage technologies will enter the market in volume and upend long-held principles used to design the storage hierarchy. The disruption that these forces will bring to bear on high-performance computing (HPC) will also create significant opportunities to innovate and accelerate scientific discovery. To ensure that NERSC fully capitalizes on these opportunities, we have developed a comprehensive vision for the future of storage in HPC and identified short- and long-term strategic goals to effectively realize this vision. This report presents the results of this effort and offers a blueprint for designing a storage infrastructure for supporting HPC through 2025 and beyond.

At a high level, a broad survey of scientific workflows and user requirements reviews identified four logical tiers of data storage with different performance, capacity, shareability, and manageability requirements:

- **Temporary storage**, which contains data being actively used by simulation and data analysis applications over the course of hours to days.
- **Campaign storage**, which contains data being actively used by larger workflows and science projects over the course of weeks to months.
- **Community storage**, which contains larger datasets that are shared among different projects within a scientific community over the course of years.
- **Forever storage**, which contains high-value or irreplaceable datasets indefinitely.

These four tiers do not neatly map to the physical storage hierarchy presently deployed at NERSC today, but over the next several years, NERSC will use tactical deployments to closely align storage resources with these requirements. By 2020, our aim is to accommodate Temporary storage data and much of the Campaign storage data onto a single, flash-based storage system that is tightly integrated with the NERSC-9 compute platform that will be deployed that year. Simultaneously, disk-based Community and tape-based Forever tiers will be more closely coupled and provide a single, seamless user interface that will simplify the management of long-lived data for both users and center staff. These tiers will be implemented off-platform to enable them to grow in response to user needs and persist beyond the lifetime of the NERSC-9 compute system.

By 2025, the nonvolatile media underpinning the converged Temporary/Campaign storage tier will expose extreme performance and scalability through a high-performance object interface. Users who want to use a familiar POSIX file system interface to access data on this system will use POSIX middleware that provides compatibility at the cost of performance. Similarly, the off-platform Community/Forever tiers will converge into a single mass storage system by 2025, and data access will occur through industry-standard object storage interfaces that more naturally map to the use patterns of long-lived data. Today's file system interfaces and custom HPSS client software will be alternate access modes, but the underlying storage system will transparently combine the economics of tape and the accessibility of disk into one seamless data repository.

The transition from file systems to object stores as exascale becomes widespread in 2025 will require users to change their applications or adopt I/O middleware that abstracts away the interface changes. Ensuring that users, applications, and workflows will be ready for this transition will require immediate investment in testbeds that incorporate both new nonvolatile storage technologies and advanced object storage software systems that effectively use them. These testbeds will also provide a foundation on

which a new class of data management tools can be built to leverage the flexibility of user-defined object-level metadata.

As the DOE Office of Science's mission computing facility, NERSC will follow this roadmap and deploy these new storage technologies to continue delivering storage resources that meet the needs of its broad user community. NERSC's diversity of workflows encompass significant portions of open science workloads as well, and the findings presented in this report are also intended to be a blueprint for how the evolving storage landscape can be best utilized by the greater HPC community. Executing the strategy presented here will ensure that emerging I/O technologies will be both applicable to and effective in enabling scientific discovery through extreme-scale simulation and data analysis in the coming decade.

1. Introduction

The National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory is the mission scientific computing facility for the Office of Science (SC) in the U.S. Department of Energy (DOE). As one of the largest facilities in the world devoted to providing computational resources and expertise for basic scientific research, NERSC is a world leader in accelerating scientific discovery through high performance computing (HPC) and data analysis. Storage systems play a critical role in supporting NERSC's mission by enabling the retention and dissemination of science data used and produced at the center. Over the past 10 years, the total volume of data stored at NERSC has increased from 3.5 PiB to 146 PiB and continues to grow at an annual rate of 30%, driven by a 1000x increase in system performance and 100x increase in system memory. In addition, there has been dramatic growth in experimental and observational data, and experimental facilities such as the Large Synoptic Survey Telescope (LSST)[1] and Linac Coherent Light Source (LCLS)[2] are increasingly turning to NERSC to meet their data analysis and storage requirements.

As these data requirements continue to grow, the technologies underpinning traditional storage in HPC are rapidly transforming. Solid-state drives are now being integrated into HPC systems as a new tier of high-performance storage, shifting the role of magnetic disk media away from performance, and tape revenues are on a slow decline. Economic drivers coming from cloud and hyperscale data center providers are altering the mass storage ecosystem as well, rapidly advancing the state of the art in object-based storage systems over POSIX-based parallel file systems. In addition to these changing tides, non-volatile storage-class memory (SCM) is emerging as an extremely high-performance, low-latency media whose role in the storage hierarchy remains the subject of intense research. The combination of these factors broadens the design space of future storage systems, creating new opportunities for innovation while simultaneously introducing new uncertainties.

To clarify how the evolving storage requirements of the NERSC user community can be best met given the storage technology landscape over the next ten years, we present here a detailed analysis of NERSC users' data requirements and relevant hardware, middleware, and software technologies and trends. From this we propose a reference storage architecture that addresses the increasing data demands from external experimental facilities, data science, and other emerging workloads while continuing to support the needs of traditional HPC users. We enumerate the requirements of longer-termed storage resources that enable publication, collaboration, and curation over multiple years.

We lay out a roadmap for the center to deploy storage resources that best serve NERSC users in 2020 and identify the actions required to realize this strategy. We then describe the evolution of storage systems beyond 2020 and how advances in storage hardware and innovation within DOE and in industry will impact our long-term storage strategy through 2025. With this roadmap and long-term strategy, we identify areas where NERSC is positioned to provide leadership in storage in the coming decade to ensure our users are able to make the most productive use of all relevant storage technologies. Because of the NERSC workload's diversity across scientific domains, this analysis and the

---

[1] Ivezić, Z et al. 2011. Large Synoptic Survey Telescope (LSST) Science Requirements Document. https://docushare.lsst.org/docushare/dsweb/Get/LPM-17. Accessed September 11, 2017.
[2] 2016. LCLS Data Analysis Strategy. https://portal.slac.stanford.edu/sites/lcls_public/Documents/LCLSDataAnalysisStrategy.pdf. Accessed September 11, 2017.

reference storage architecture should be relevant to HPC storage planning outside of NERSC and the DOE.

2. NERSC Storage Hierarchy

NERSC has more than 6,000 active users with more than 700 active projects that span a broad range of science disciplines, such as materials science, astrophysics, bioinformatics, and climate science. The diversity of workflows at NERSC result in a wide range of I/O patterns, data volumes, and retention requirements; for example, a number of projects use data from experimental and observational facilities as part of their workflow and need high-capacity storage at NERSC to ingest observational data that is transferred over the wide-area network. A growing number of projects also combine modeling and simulation with experimental or observational data, which is increasing the complexity of workflows and the demand for storage resources accessible from both extreme-scale compute systems and the wide-area network. To meet these diverse needs, NERSC maintains different tiers of storage, each optimized for a different balance of performance, capacity, and manageability.

2.1. Current Storage Infrastructure at NERSC

As of 2017, the NERSC storage hierarchy consists of a 1.6 PiB flash-based burst buffer, a 27 PiB Lustre scratch file system built using hard disk drives (HDDs), a 10.7 PiB disk-based project file system that provides medium term storage, and a 130 PiB enterprise tape-based archive. These tiers, depicted schematically in Figure 1, vary in capacity, performance, reliability, and data management policies.

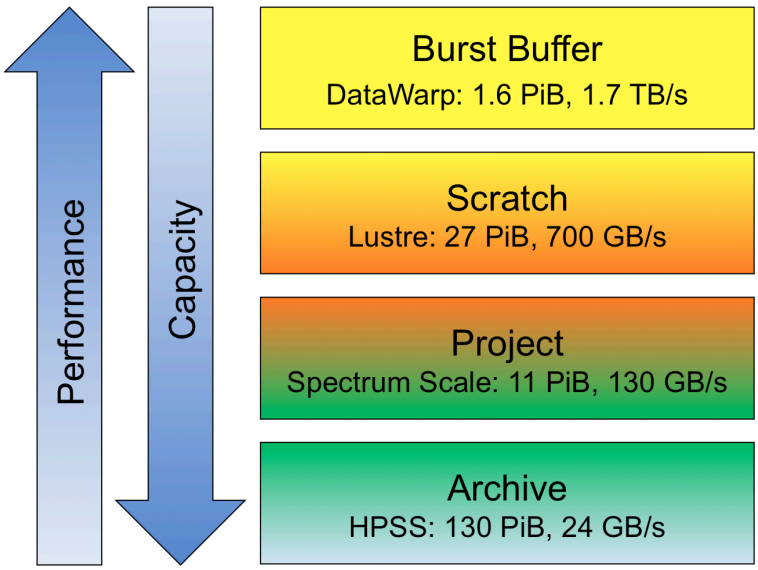

FIGURE 1. STORAGE HIERARCHY AT NERSC IN 2017

The top two tiers (burst buffer and scratch) are optimized for performance and provide sufficient capacity to support typical active workloads in the system. These storage systems are either actively purged or require users to request resources as part of their job. They are advertised as scratch space and managed as more volatile and less robust resources, and users are encouraged to save critical data and results to the other tiers. The disk-based scratch tier is currently implemented using the Lustre parallel file system, and the burst buffer currently uses Cray's DataWarp file system and infrastructure.

The project and archive tiers are optimized for capacity and durability, but still provide sufficient performance to allow users to move data effectively in and out. These tiers are not actively purged but instead managed via quotas. The project tier is disk-based and runs IBM's Spectrum Scale parallel file system (previously known as GPFS), while the archive tier uses a combination of disk and tape that are managed by the HPSS software developed by a collaboration between DOE labs and IBM.

Reliability and manageability are a major concern for the project and archive tiers since they are often the repositories for users' most critical data. Data stored in these systems are critical to support the scientific process itself, since scientific results must be maintained for long periods of time and are often shared through the community via data portals[3] associated with these storage systems. Consequently, the storage software technologies used for these tiers must be highly robust. These tiers must also be able to grow over time to allow for external projects to sponsor additional space to meet mission or science requirements. For example, various experimental projects such as STAR[4] and ALICE,[5] along with experimental facilities such as the ALS[6] and JGI,[7] have augmented NERSC's project file system to store their data. This contrasts sharply with the burst buffer and scratch tiers, which are typically designed specifically to meet the needs of the computational platform with which they are procured.

2.2. Workflow-based Model for Storage

In preparation for NERSC's next major system, to be deployed in 2020, and as part of the Alliance for Application Performance at Extreme Scale (APEX),[8] the NERSC division of Lawrence Berkeley National Laboratory, Los Alamos National Laboratory (LANL), and Sandia National Laboratory (SNL) surveyed their users' scientific workflows to inform the technical requirements for the procurement of the NERSC-9 and Crossroads systems. The results of this analysis, summarized in the APEX Workflows white paper,[9] presents the data movement between different stages of workflows as workflow diagrams to help reason about system architecture; an example of such a diagram is shown in Figure 2. The vertical axis captures the required retention time for the data inputs and outputs and is a major contributor to storage system capacity requirements. The vertical axis also speaks to the performance requirements of each tier, as data that is generated (and deleted) more frequently will require higher performance than those data products that are generated much less frequently.

---

[3] ALS Data and Simulation Portal. https://spot.nersc.gov/. Accessed September 4, 2017.
[4] Adams, J. et al. 2005. Experimental and theoretical challenges in the search for the quark–gluon plasma: The STAR Collaboration's critical assessment of the evidence from RHIC collisions. *Nuclear Physics A*. 757, 1–2 (Aug. 2005), 102–183.
[5] Aamodt, K. et al. 2008. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*. 3, 8 (Aug. 2008), S08002–S08002.
[6] Advanced Light Source. https://als.lbl.gov/. Accessed September 3, 2017.
[7] DOE Joint Genome Institute: A DOE Office of Science User Facility of Lawrence Berkeley National Laboratory. https://jgi.doe.gov/. Accessed September 3, 2017.
[8] Alliance for Application Performance at Extreme Scale. http://www.lanl.gov/projects/apex/. Accessed April 30, 2017.
[9] APEX Workflows. http://www.nersc.gov/assets/apex-workflows-v2.pdf. Accessed April 30, 2017.

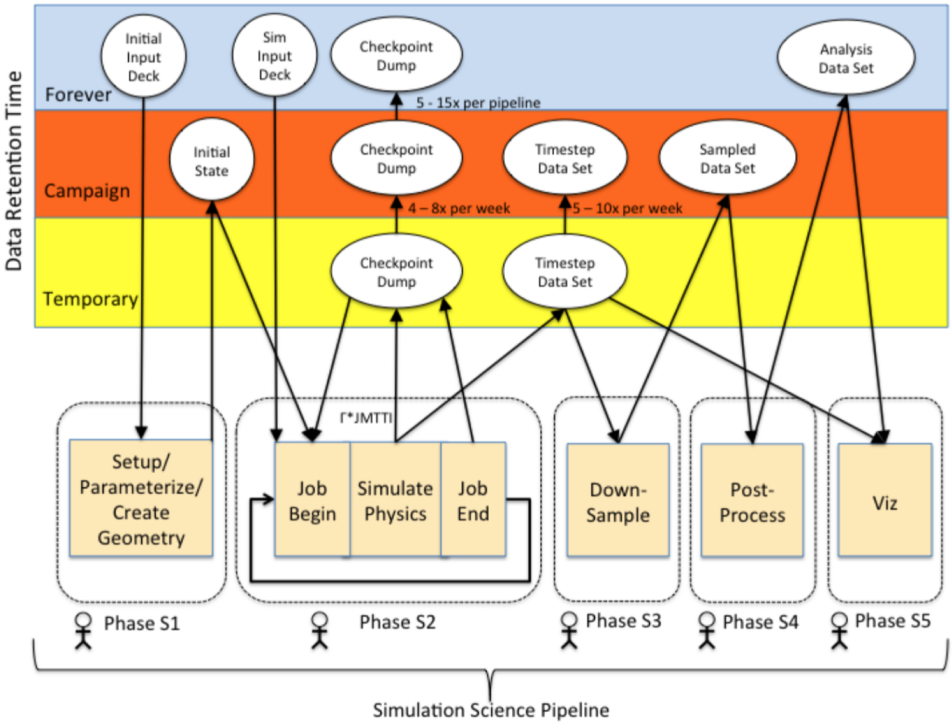

FIGURE 2. DATA MOTION AND RETENTION IN AN ARCHETYPAL SIMULATION SCIENCE PIPELINE. FROM THE APEX WORKFLOWS WHITEPAPER.[10]

Overall, this study found commonality across DOE in compute and storage requirements, and it presented a taxonomy of workflows' storage requirements in the form of three logical storage tiers: Temporary, Campaign, and Forever:

- **Temporary storage**, used for the duration of a single workflow instance, is used to store and deliver working sets, checkpoints, and job outputs. It is the highest performing storage resource, and as such is typically tightly coupled to the compute system.
- **Campaign storage**, used for the duration of a project or allocation, enables collaboration within a group of researchers, provides space for post processing and input sets for subsequent runs, and facilitates data curation for later publication or movement to longer-term storage. It requires greater capacity but less performance than the Temporary storage tier.
- **Forever storage**, used for long-term storage, acts as a repository for high-value data that is irreplaceable or prohibitively expensive to reproduce. It will contain raw datasets, often too large to store in other resources, and may also store golden datasets that are of wider value to scientific communities. Its performance requirements are lower than Campaign storage, but it must be able to reliably hold years or decades worth of data.

In addition to these three tiers formalized in the APEX Workflows document, there are additional design criteria that are critical to NERSC's users: the ability to ingest and store data from remote instruments, the availability of access controls for publishing and sharing, and the ability to efficiently index, search, and describe datasets. Thus, we also identify a fourth, **Community storage,** resource that is optimized

[10] 2016. APEX Workflows Whitepaper. http://www.nersc.gov/assets/apex-workflows-v2.pdf. Accessed April 30, 2017.

to ingest data from experimental and observational facilities, share data with researchers at other centers, and facilitate the curation of data.

Figure 3 summarizes the functionality of these four logical tiers in terms of their balance of capacity and performance and how much optimization is invested in making their contents searchable, shareable, and otherwise easily curated.

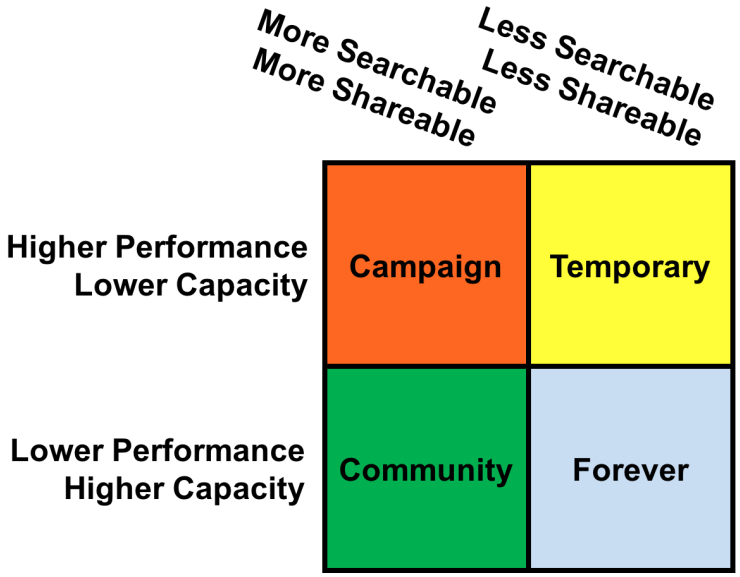

FIGURE 3. FUNCTIONAL VIEW OF STORAGE TIERS

While Figure 3 depicts a functional view of storage, Figure 4 shows how the functional model maps to the NERSC resources shown in Figure 1.

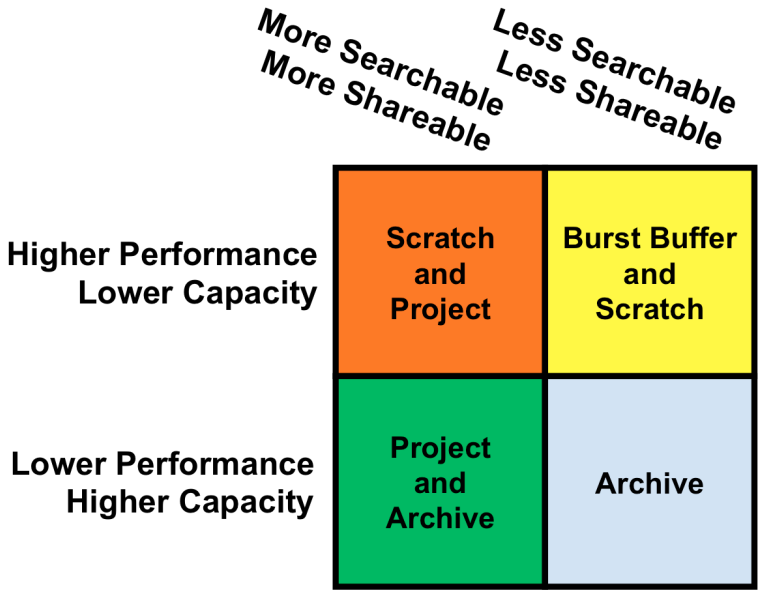

FIGURE 4. MAPPING BETWEEN FUNCTIONAL MODEL AND ACTUAL STORAGE RESOURCES AVAILABLE AT NERSC.

As is clear in this diagram, the storage resources provided by NERSC today do not precisely align with the four logical tiers we have identified. However, with the understanding that four logical tiers need not necessarily map to four physical storage resources, this serves as a sound approach to defining the design optima and goals for future physical storage resources.

3. Requirements

As previously indicated, the NERSC workload is evolving as a result of a variety of scientific and technological changes. To ensure that future compute and storage resources will meet these evolving needs, we draw on a variety of requirements studies that include current workloads, the APEX Workflows white paper,[11] the DOE Exascale Requirements Reviews,[12] and NERSC staff experiences.

3.1. Current I/O Patterns

Examining current user and application I/O behavior targeting scratch file systems (the Temporary storage tier) at NERSC shows that the volume of data read from and written to these scratch file systems are approximately equal, as shown in Figure 5. This is likely due to a balance between checkpoint-heavy workloads (many write-heavy checkpoint operations for each read-heavy restart operation), common experimental and simulation datasets being re-read multiple times, and write-once, read-once intermediate files generated by scientific workflows, as noted in Figure 2.

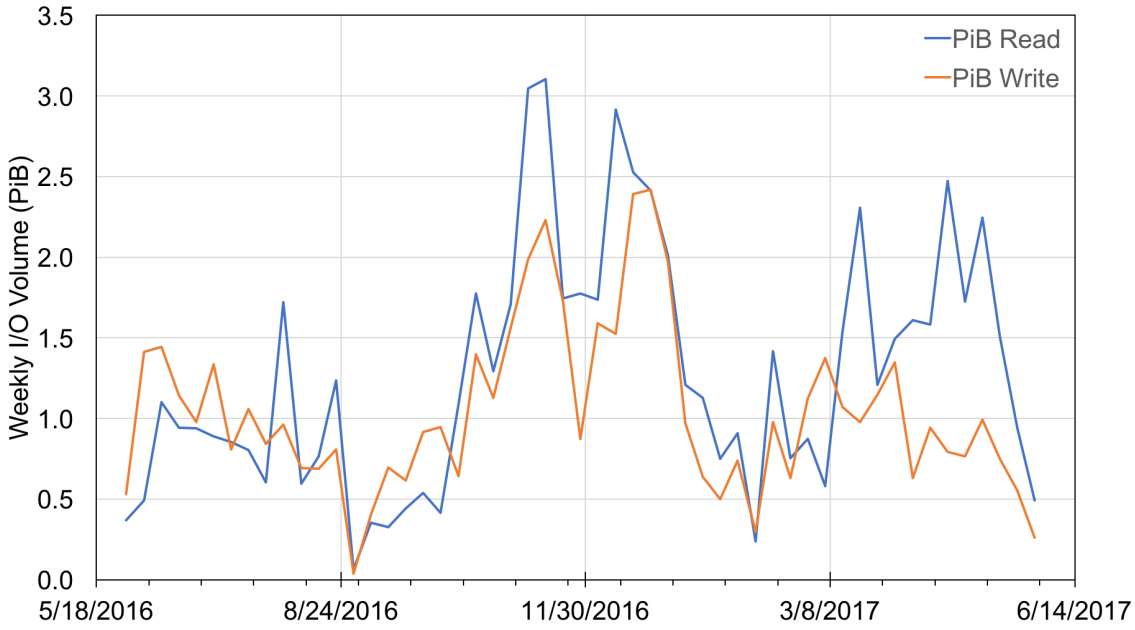

FIGURE 5. WEEKLY I/O READ AND WRITE VOLUMES ON NERSC EDISON'S SCRATCH1 AND SCRATCH2 LUSTRE FILE SYSTEMS. OVERALL ANNUAL AVERAGE READ/WRITE RATIO IS 11/9.

This analysis indicates that Temporary storage needs to provide balanced read and write capabilities and that storage media, APIs, or access semantics that emphasize one over the other would not be

[11] APEX Workflows. http://www.nersc.gov/assets/apex-workflows-v2.pdf. Accessed April 30, 2017.
[12] DOE Exascale Requirements Review. http://www.exascaleage.org/. Accessed August 31, 2017.

suitable for the NERSC workload. In addition, the Temporary and Campaign storage tiers should be strongly coupled to streamline data motion of hot datasets between the working space and a storage resource that facilitates data management over the course of the larger scientific study.

As shown in Figure 6, NERSC applications also use a variety of POSIX metadata calls within Temporary and Campaign storage systems, with the vast majority being opens, closes, and stats. It is therefore essential that the Temporary storage resource's system software implement these calls in a highly scalable fashion; for example, calculating the size of a file that is striped across hundreds of storage servers must be efficient, and allowing users to obtain file handles by which they can access their stored data must incur minimal latency.

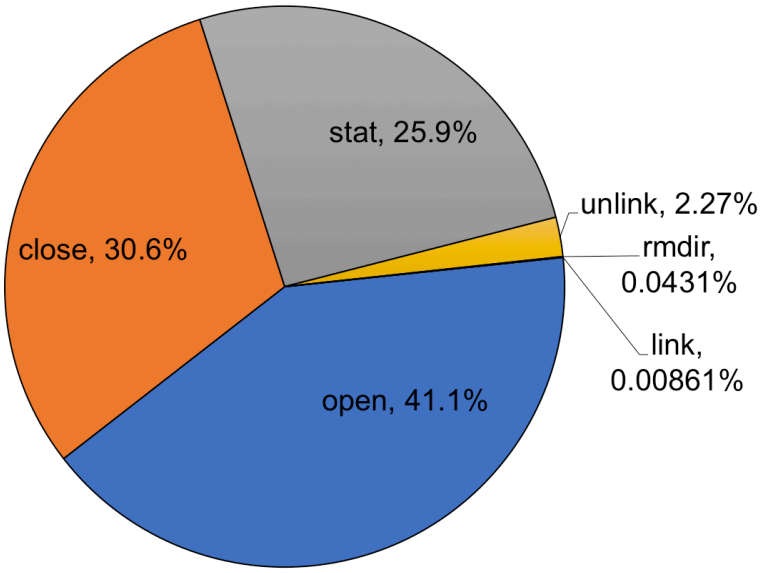

FIGURE 6. DISTRIBUTION OF METADATA OPERATION COUNTS ON NERSC EDISON'S SCRATCH1 AND SCRATCH2 LUSTRE FILE SYSTEMS FROM JUNE 2016 TO JUNE 2017.

Intuitively, accesses to Forever storage should skew toward writes, but this is not pronounced at NERSC; 24% of data written to the archive is recalled at some point. In fact, some archived data shows a high skew toward reads as a result of science communities continually accessing large datasets. The net result is that NERSC's archive read-to-write ratio is remarkably balanced, with reads accounting for 40% of system I/O. Given that NERSC's Forever tier is magnetic tape, and tape reads are more difficult to manage and are slower due to volume mount and seek latencies on linear access media, we conclude that this is a result born out of necessity rather than best use of the system or desires of users. Re-reads would be better served from lower-latency Campaign or Community storage layers if capacity allowed.

With sufficiently sized Community and Campaign storage tiers, Forever storage should be optimized for high performance write capabilities rather than read performance, as read duty is mainly fulfilled by Community and Campaign storage resources. However, as shown in Figure 4 (which depicts the reality at NERSC), this is a system design point rather than a statement of how the current storage systems work. The discrepancy is driven in large part by the fact that tape is still the most cost-effective mass storage medium on a dollars-per-bit basis.

The coupling between Forever and Community storage can be looser than Temporary and Campaign, as the data in Forever and Community space is principally static. Community storage should be sized such

that data does not migrate frequently to and from Forever storage. Because of the difficulty interacting with tape, Community storage needs to be large enough that it effectively eliminates repeated re-reads from the Forever tier. For evaluating effective technologies, POSIX I/O operations are much simpler in the Community and Forever space, mainly composed of put / get / stat on whole files, with other operations to create and maintain directory hierarchies and very little else. Such a write-once, read-many (WORM) workload is an area where inexpensive capacity storage systems without full POSIX compliance could be deployed; for example, the object storage systems used extensively in the cloud and hyperscale markets are specifically optimized for WORM I/O.

As science teams move from using small numbers of applications during their research to more complex interactions between many applications, scientific workflows are expected to become the dominant mode of operation at NERSC. The compute concurrency of these workflows is diverse and can be extremely low for image or other instrument-analysis workflows. These data-oriented workflows are anticipated to grow more in throughput rather than problem size by 2020, and because many constituent applications do not strong scale well, the increased concurrency of NERSC's future systems will be utilized by bundling multiple workflow pipelines into a single job.[13] Unlike the scaling behavior of traditional simulation science applications, this will demand scalable metadata performance from the storage system as each node processes larger numbers of files concurrently.[14]

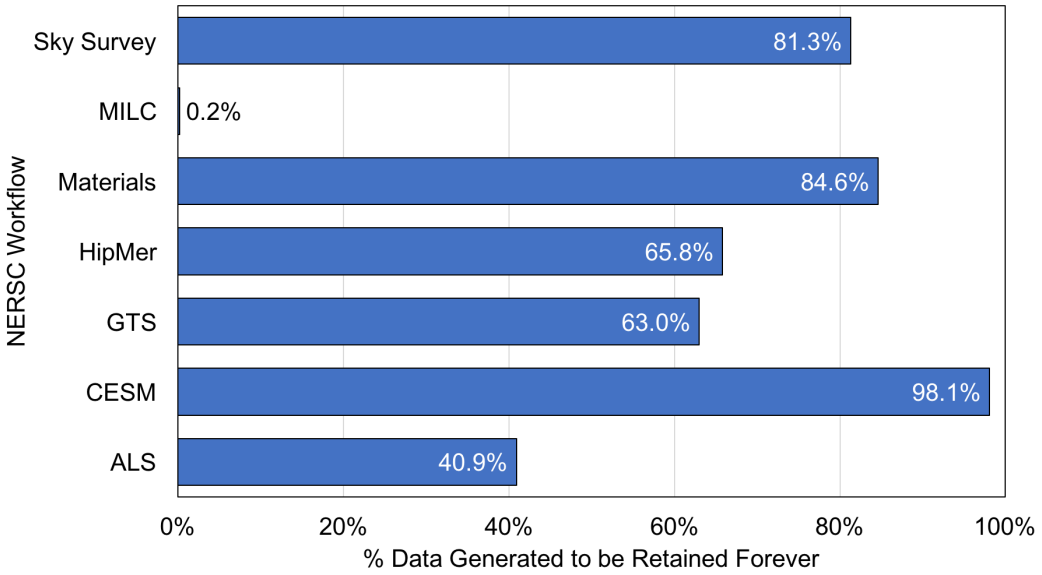

FIGURE 7. PERCENTAGE OF DATA GENERATED BY NERSC WORKFLOWS THAT WILL BE RETAINED IN FOREVER STORAGE

A key finding of the APEX Workflows study was that NERSC users want to save a significant fraction of the data files used and produced by their workflows for a long time, perhaps indefinitely. Figure 7 shows the percentage of I/O generated by the surveyed NERSC workflows that is saved forever. Even if users

[13] Daley, C.S. et al. 2015. Analyses of Scientific Workflows for Effective Use of Future Architectures. *Proceedings of the 6th International Workshop on Big Data Analytics: Challenges, and Opportunities (BDAC-15)* (Austin, TX, 2015).
[14] Daley, C.S. et al. 2016. Performance Characterization of Scientific Workflows for the Optimal Use of Burst Buffers. *Proceedings of the Workshop, Workflows in Support of Large-Scale Science (WORKS 2016)* (Salt Lake City, 2016), 69–73.

are able to make use of in-situ or in-transit analytics to reduce data movement during workflow execution, a large fraction of the generated data is irreducible and must be retained long-term.

Thus, in-flight analytics are not magic bullets that can be relied upon to stem the increasing volumes of data being generated by scientific workflows, and we are rapidly approaching the need for O(exabyte) of capacity storage unless NERSC users re-architect their workflows to save less data. Extrapolating the historic 45% annual growth rate of NERSC's current archive system alone predicts 1 exabyte of user data by 2022. Given the aforementioned observation that NERSC users are currently using the archive as both Community and Forever storage, effectively balancing the capacity of Community storage relative to Forever storage indicates the need for hundreds of petabytes of capacity in the Community storage tier by 2023.

The findings presented above indicate two corollaries:

- Campaign storage is, in a sense, "cold" Temporary storage, and Community storage is "hot" Forever storage.
- The data stored in Temporary/Campaign storage serves the goals of individual research projects and their users, while data in Community/Forever storage may be of interest to broader scientific communities and many research projects.

These suggest a broad dichotomy between Temporary/Campaign storage and Community/Forever storage in both their data retention times and the breadth of users they serve. It follows that Temporary/Campaign storage is best implemented close to specific compute systems to emphasize high-performance analysis and access by a small cohort of users. Conversely, Community/Forever storage is best maintained closer to the wide-area network and more centrally within a facility to emphasize sharing and broad access by larger user communities.

From these user requirements, several key design criteria become apparent. The Temporary and Campaign tiers should be closely coupled and provide balanced read/write performance and scalable metadata to support the NERSC workload. The Community and Forever tiers do not need such tight coupling, but they should be sized such that most read activity targets data that is stored in the Community tier rather than Forever storage. This would allow Community storage to make use of technologies optimal for WORM workloads, leaving Forever storage for highly valuable but cold data.

3.2. NERSC-9 Requirements

In 2020, NERSC plans to deploy its NERSC-9 system, which is targeted to increase the processing capability of the center by 4-5x over the NERSC-8 system, Cori. With the potential for dramatic data growth as emerging areas in data sciences matures, this increase in computing capability is expected to be accompanied by at least a proportional increase in the rate and volume of data generation within NERSC. The NERSC-9 system will include platform storage that is explicitly designed to:

"[retain] all application input, output, and working data for 12 weeks (84 days), estimated at a minimum of 36% of baseline system memory [3 PiB] per day."[15]

[15] APEX 2020 Technical Requirements Document for Crossroads and NERSC-9 Systems. http://www.lanl.gov/projects/apex/request-for-proposal.php. Accessed April 30, 2017.

as well as deliver sufficient performance to absorb checkpointing. The technical requirements for the NERSC-9 system were specified such that platform-integrated storage will fulfill the role of Temporary storage and a portion of Campaign.

While the utility and capability of this platform storage will be well-defined in the 2020 timeframe, it is designed to retain data for only 84 days. Therefore, users and projects that wish to retain data long-term must store it on alternate, longer-term storage resources that fall outside of the NERSC-9 procurement. However, in the NERSC-9 storage technical requirements, vendors were given the flexibility to respond with innovative solutions surrounding features that are more relevant to longer-term data management, including background data integrity verification, detailed monitoring of storage performance and utilization, fast metadata traversal, and connectivity to external file systems and other data sources. As a result, the NERSC-9 procurement could be used as a vehicle for procuring and satisfying the requirements of Campaign, Community, and Forever storage tier as well.

3.3. DOE Exascale Requirements Reviews

The DOE Advanced Scientific Computing Research (ASCR) program has conducted a number of requirements gathering efforts with other DOE SC programs to ensure that the exascale systems to be fielded in 2021-2023 are aligned with the mission needs of each DOE SC program office. These efforts build on a long history of engagement with the scientific community that help drive future system requirements and architectures, going back to the NERSC's Green Book[16] review in 2002 and extending to the recent DOE Exascale Requirements Reviews.[17] The output of these efforts directs the planning and acquisition strategies for NERSC, the Leadership Computing Facilities and ESnet.

These comprehensive reports span a broad range of areas, including computational requirements, software and middleware needs, networking, data management, and data analysis. Some of the common data and storage requirements that emerged from those efforts that are relevant to NERSC's storage strategy are as follows:

1. Many of the program offices anticipate exabyte-scale storage needs in the coming decade, with many projects generating and processing hundreds of terabytes of data today and projecting 10-50x growth during that decade. Multiple projects are predicting 100 petabyte or greater datasets in the 2025 timeframe. These use cases underline the need for cost-effective, capacity-optimized Community and Forever storage.
2. There is an increasing need to integrate observational and simulation data in workflows that require data to be co-located for effective analysis. This is, in part, a direct result of typical observational and simulation results now surpassing the analysis capabilities of computing systems at users' home institutions. This will drive the need to improve data movement tools, increase storage capacity, and provide high-bandwidth, wide-area networking connectivity. This speaks to the need for effective integration between all storage tiers to minimize the complexity of data movement during workflows.
3. Data management needs to extend beyond NERSC to the wide-area network, as other compute and experimental facilities integrate more closely with NERSC. External connectivity requirements are also being driven by a growing demand to share common, curated datasets with the wider community, driving the need for a robust Community storage resource.

---

[16] Greenbook – Needs and Directions in High-Performance Computing for the Office of Science. https://www.nersc.gov/assets/For-Users/DOEGreenbook.pdf. Accessed April 27, 2017.
[17] DOE Exascale Requirements Review. http://www.exascaleage.org/. Accessed August 31, 2017.

4. Users have a strong need for integrated data tracking and provenance within the storage system. This includes expanded capabilities around metadata storage, searching and querying, and event triggering. These are features that are principal to Campaign and Community storage tiers.
5. There is a transition from individual large-scale simulations toward ensembles, uncertainty quantification, and more complex workflows that must connect and integrate simulation and analysis. This shift towards ensemble workflows will require that Campaign storage simplify data management across large projects and the other tiers.
6. The dramatic growth in data storage demands is accompanied by a desire to apply new forms of data analysis and analytics, including machine learning, to effectively process the massive amounts of data resulting from experimental sources, extreme-scale simulation, and uncertainty quantification. This aligns with the observation at NERSC that Temporary storage deliver balanced read and write performance.

All divisions within DOE SC anticipate that the dramatic increase in their computational requirements will drive similarly dramatic increases in their data storage and management requirements. Simply providing high capacity, high-bandwidth storage will no longer satisfy the broad range of requirements that arise from the aforementioned shift toward workflow-oriented processing and experimental analysis. Rather, future storage systems will have to deliver low latency (high IOPs), rich metadata facilities, and external connectivity, in addition to high parallel I/O bandwidth. These user requirements reinforce the need to treat storage infrastructure design as a multi-dimensional problem and support the approach described in Section 2.2.

3.4. Emerging Applications and Use Cases

A growing number of domain sciences need to leverage the capabilities of HPC systems, yet have data requirements that contrast with those of traditional HPC workloads. Many of these emerging data workloads are driven by machine learning and other data analytics techniques that rely on workflow frameworks (e.g. Apache Spark), analytics packages (e.g., Caffe and TensorFlow), and domain-specific libraries that traditionally have not been used in HPC. These analysis tools often exhibit I/O patterns that perform poorly on HPC systems as a result of their genesis in cloud environments, and while individual analytics tools can be refactored for use on HPC systems, the field of data science is evolving rapidly and independently of the HPC community. The next set of popular tools may exhibit the same deleterious I/O behavior and poor out-of-box performance, and they will need to be adapted to HPC environments because of their prioritization of productivity and their momentum in the larger data analytics community.

Many of these emerging applications areas are associated with observational and experimental facilities that are already generating large volumes of data, and, as highlighted in Section 3.3, their projected growth rates are staggering. For example, NERSC is collaborating with the Linear Coherent Light Source (LCLS) to enable real-time analysis of data generated by high-speed, high-resolution instruments. These instruments currently generate hundreds of megabytes per second of data but are projected to generate tens to hundreds of gigabytes per second of data with future upgrades. Instruments at the National Center for Electron Microscopy,[18] the Advanced Photon Source,[19] the Spallation Neutron Source,[20] and

[18] National Center for Electron Microscopy (NCEM). http://foundry.lbl.gov/facilities/ncem/. Accessed September 11, 2017.
[19] Advanced Photon Source. https://www1.aps.anl.gov/. Accessed September 11, 2017.
[20] Spallation Neutron Source. https://neutrons.ornl.gov/sns. Accessed September 11, 2017.

elsewhere project similar increases. These facilities also often run 24x7 for months at a time, so availability and reliability of the compute, storage, and network resources supporting these workflows is critical. Given the fact that researchers are often allocated very limited time on these instruments, providing continuity of storage and computing resources, even through system maintenance periods, is important.

Direct interactions between NERSC staff and the staff and users from a number of experimental facilities and projects have revealed several key storage requirements. There will be a need to transfer hundreds of GB/sec from the wide-area network directly to a durable storage resource such as Campaign or Forever storage in a reliable way. This translates to a need for high availability and accessibility of data on these tiers through maintenance, software upgrades, and storage expansion. Furthermore, predictable I/O performance for both data and metadata accesses is critical for co-scheduling experimental and computational resources, and providing quality of service controls is highly desirable across all storage tiers.

3.5. Operational Requirements

User requirements reviews and other surveys define many design criteria for the storage system architecture such as I/O performance and data manageability, but operational considerations and data lifecycle management needs give rise to additional requirements that are not directly user-facing. These operational requirements are especially critical for the Community and Forever storage resources, which will retain long-lived data. Data on these resources will routinely outlast the four- to five-year lifespan of individual compute platforms and must be available across all compute systems and accompanying edge services at the center.

As discussed in Section 2.1, the role of Community storage at NERSC is currently fulfilled by the project file system which has been in existence for more than 10 years. Forever storage is fulfilled by the HPSS-based archive and has been managed for more than 20 years. Dozens of NERSC staff have accumulated hundreds of years of direct experience managing long-lived HPC storage systems, contributing to community best practices and working with peers at other DOE HPC facilities. They have identified critical attributes needed to maintain and run these systems effectively. These operational requirements can be organized into three general categories, described in sections 3.5.1-3.5.3.

3.5.1. Reliability, Durability, Longevity, and Disaster Recovery

Because Community and Forever storage are expressly designed to store valuable data, ensuring that the data is highly resistant to corruption, available even in the presence of component failures, and can be quickly restored in the event of a disaster are paramount. Although virtually all mass storage systems make assurances about these features, it is important to note the effort required by storage system operators to exercise these features in practice. This effort has a direct effect on the staffing levels required to support the storage system as it increases in capacity and may be of critical importance to ensure the minimal downtime during outages required by the emerging applications and use cases discussed in Section 3.4.

Required features include:

- **Highly durable hardware and software**. For the archive, tape media has offered not only cost-effective capacity but additional durability assurance because the data is offline. This makes it far less prone to data corruption due to software error, as evidenced by a 2011 software-induced disaster at a leading hyperscale provider.[21]
- **High degree of reliability and integrity for data at rest and in motion**. This may be addressed by mechanisms like T10 DIF and data checksumming and is critical to preventing silent data corruption, as evidenced by a 2013 hardware-related data corruption issue within Internet2.[22]
- **Ability to shrink, grow, and migrate data "live," as capacity is increased or reconfigured**. This is an essential feature for repacking old data to new, higher capacity media. It also enables NERSC to allow large experiments and other data-intensive users to purchase additional storage to be co-located with their compute resources.
- **Ability to mount storage resources across different compute and login systems and over tens or hundreds of thousands of client nodes**. This is important for all tiers but particularly essential for the Campaign and Community storage tier, which must interface with a diversity of environments to ingest experimental data and share datasets.
- **Flexible support for a variety of high-performance networks**. This allows the storage to continue to be compatible as the center's network and compute technologies evolve with changing user requirements and emerging technologies.

3.5.2. Space management and curation features

Effectively managing storage resource utilization reduces storage costs and improves quality of service. While management features such as supporting user- and group-level quotas are supported by virtually all storage systems, it can be an inflexible and opaque approach if users do not have the ability to determine what data they have. Giving users and administrators the ability to determine which datasets are consuming the most space and where these large datasets are located simplifies their data management overhead. Required space management and curation features include:

- **Flexible methods to track usage and to specify and enforce limits** (e.g. user quotas, tree quotas, etc). This allows users and operators to make more informed decisions about which data can or should be deleted to ensure fair share of storage resources.
- **Methods to quickly walk the storage resource namespace**. In addition to helping inform space management decisions, understanding the distribution of file or object sizes, access frequencies, and other metadata informs policy decisions and system performance optimization.
- **Ability to manage hardware that has different characteristics** (bandwidth, capacity, IOPs) within the same system. This allows the storage system to grow along independent dimensions (e.g., performance and capacity) and is of increasing importance with emerging NAND and SCM media.

3.5.3. Availability

Maintaining the highest possible availability of storage resources is essential to operating a supercomputing center; an entire center can be rendered offline if its storage systems are offline. Furthermore, the need to maintain extreme availability and minimize maintenance outages only becomes greater as experimental facilities become coupled to HPC facilities; as described in Section 3.4, storage system downtime can severely impact the ability of a user of an experimental facility to do

---

[21] Treynor, B. 2011. Gmail back soon for everyone. https://gmail.googleblog.com/2011/02/gmail-back-soon-for-everyone.html. Accessed September 4, 2017.
[22] Foster, I. 2013. Globus Online ensures research data integrity. https://www.globus.org/blog/globus-online-ensures-research-data-integrity. Accessed September 4, 2017.

research. As such, we have identified the following operational requirements to ensure maximum availability:

- **Strong support for live updates, rolling upgrades, live configuration changes, etc**. This minimizes the need to take the system offline, especially for extended periods of time, and speaks directly to the requirement of maintaining high availability during maintenance.
- **Support for centralized management and monitoring**. This improves operational efficiency and reduces downtime by decreasing the amount of effort required for storage engineers to manage multiple tiers of highly distributed storage.
- **Ability to recover cleanly from faults or failures with minimal cleanup and manual intervention**. As with previous operational requirements, this is directly tied to reducing downtime and staffing requirements.

3.6. Gaps and Challenges

While the current storage hierarchy described in Section 2 has served NERSC well, contrasting it with the requirements stated in this section reveal some shortcomings in its overall architecture, the deployed technology, and its ease of use. If these gaps are not addressed, they will be further aggravated by technology trends and emerging user needs.

3.6.1. Tiering

The number of layers in the hierarchy is driven by cost optimizations to provide fast, high-performance storage to support running simulations and analysis (Temporary storage); high capacity to support longer-term projects (Campaign/Community storage); and archiving data to support the scientific process (Community/Forever storage). Tiered storage adds complexity for users and staff, and the lack of automated data movement between tiers is a significant burden to NERSC users. Each layer of the storage hierarchy is a complex, independent system that requires expertise to manage, and collapsing tiers would simplify storage administration for NERSC and reduce data management complexity for users.

3.6.2. Data Movement

At present, moving data between NERSC's Temporary and Campaign/Community storage tiers is relatively frictionless, as they both provide a POSIX file system interface. Movement in and out of Forever storage is more challenging because it requires users to interact with custom client software similar to FTP or UNIX tar. The fact that data resides on tape—which introduces volume mount latencies that may span several minutes and linear read or write access restrictions, plus the fact that data may be scattered over many different tape cartridges—adds to the difficulty. Providing a common interface for all tiers, whether it be file-based or object-based, would streamline data movement and simplify the task of building more productive user interfaces to manage data movement.

3.6.3. Data Curation

Integrated search and discovery tools are lacking at all levels of the storage hierarchy today. This is more problematic for Community and Forever storage, where significant quantities of data are resident for years or decades. These tiers often serve as shared data repositories for multiple projects over a long period of time, and the individual owner or steward of a dataset may change over the course of a project.

To address these issues, large projects have built their own data catalogs that are completely external from the NERSC storage resources. Some, such as JAMO,[23] are focused narrowly on cataloging and data movement; while others, including those developed by the ATLAS[24] experiment at the Large Hadron Collider and by the Advanced Light Source[25] at Berkeley Lab, include web presentation and workflow features. Although we do not intend to define a metadata schema for all NERSC users, having a common set of metadata features across all tiers on which user communities can build their domain-specific cataloging systems would simplify data management and curation as NERSC's storage hierarchy continues to evolve over the next 10 years.

3.6.4. Workload Diversity

The span of NERSC user workloads is broad and, consequently, the scale and distribution of file characteristics and I/O patterns varies greatly. As discussed in Section 3.1, simulations running at scale often write very large checkpoints that stress the entire data path from interconnect to media. At the other end of the spectrum, many experimentally-driven projects run many low-concurrency jobs over large collections of smaller files. This can stress the metadata service and the storage system's ability to efficiently handle high volumes of small I/O operations that has knock-on effects on other users of the file system. Providing a means to distribute metadata over multiple storage servers is an essential requirement, and features that allow more intelligent partitioning of metadata on the basis of users, projects, or arbitrary data properties would benefit quality of service.

3.6.5. Storage System Software

Usability and manageability gaps exist across the storage system software used across all of NERSC's current storage tiers. For example, the Lustre-based scratch file system deployed as part of the Cori system's Temporary storage tier file system provides no straightforward way to add additional storage capacity or rebalance data across Lustre object storage targets. Lustre's management tools are also relatively immature; aside from Intel's now-unsupported Intel Manager for Lustre software,[26] there is no single-pane file system management interface for Lustre, and the majority of available tools are ad hoc scripts contributed by the community.

NERSC's Spectrum Scale-based project file system has its own set of challenges. Maintenance operations, such as file system integrity checks that require the file system to be taken offline for an extended period, work directly against the high availability requirements identified in Section 3.4. Furthermore, Spectrum Scale is a proprietary, closed-source system with annual licensing costs, and much recent development effort at IBM has gone into supporting requirements driven by enterprise, not HPC, needs.

The Forever tier, implemented using HPSS, is engineered to present a POSIX-compliant interface despite a simple put/get interface being sufficient for nearly all use cases. This POSIX-compliance adds

---

[23] New Metadata Organizer and Archive Streamlines JGI Data Management. http://www.nersc.gov/news-publications/nersc-news/nersc-center-news/2013/new-metadata-organizer-and-archive-streamlines-jgi-data-management. Accessed March 6, 2017.
[24] PDSF data disk summary. http://portal.nersc.gov/atlas_diskstat. Accessed March 6, 2017.
[25] Deslippe, J. et al. 2014. Workflow Management for Real-Time Analysis of Lightsource Experiments. *9th Workshop on Workflows in Support of Large-Scale Science*. (Nov. 2014), 31–40.
[26] Damkroger, T. 2017. A New Path with Lustre. http://intel.cmail20.com/t/ViewEmail/d/C316287F828160FA/5FC4DCCCE8C49BF9F6A1C87C670A6B9F. Accessed April 20, 2017.

significant complexity to the software, yet the user interface into this tier is through custom client software. A file system interface to the archive, either through integration with Spectrum Scale or FUSE, is possible, but the underlying tape storage can make operations that are unremarkable in a disk-based file system extremely inefficient and time consuming without careful planning.

3.6.6. Hardware Concerns

While all of the disk-based storage systems are architected for reliability with enterprise class RAID and redundancy, the demand for storage capacity is now being satisfied with more, not simply larger, disks. This has a significant effect on the overall reliability of a storage system and its characteristic mean time to data loss, and the extreme-scale storage industry is transitioning from block-based parity within each failure domain (e.g., RAID6) to highly distributed, object-level erasure coding across shelves, racks, and even data centers. File systems built upon block-based storage cannot make use of these advances in erasure coding despite the nature of magnetic disks effectively requiring it for resilience in the future, so moving the Campaign and Community storage tiers towards technologies that balance parity and resilience more effectively will be essential.

3.6.7. POSIX and Middleware

Over the last 50 years, the POSIX I/O standard[27] has stood the test of time as the canonical way to access storage devices. However, advances in software scalability and hardware performance have strained the appropriateness of the existing standard and its semantics. Either revisions to the standard or entirely new performance-optimized standards would be valuable for future applications to deal with emerging high-performance storage technologies.

Further, a great deal of I/O middleware, such as HDF5, PnetCDF, and ADIOS, are tuned to operating with the traditional memory-to-disk I/O endpoints. This middleware provides great value to application developers by isolating users from the vagaries of extracting peak performance from the underlying storage system, but it will need to be updated to handle the transition to a multi-tiered I/O configuration. Prefetching data from scratch or project into a burst buffer and migrating changes back again, support for asynchronous I/O operations, and other improvements to leverage new technologies are needed to continue supporting user requirements.

4. Technology Landscape and Trends

Having identified both the functional requirements of a future storage infrastructure at NERSC, as well as the requirements coming from users, experimental facilities, and operators, we now present hardware and software technologies that are or will be available to implement the Temporary, Campaign, Community, and Forever tiers over the next decade.

4.1. Hardware

Although the HPC industry has historically been a significant driver of mass storage hardware, the emergence of cloud and other hyperscale service providers has had a dramatic effect on the storage industry and its roadmaps for storage media. These economic forces, combined with the impending

[27] 2009. International Standard - Information Technology Portable Operating System Interface (POSIX) Base Specifications, Issue 7. ISO/IEC/IEEE 9945:2009(E). (Sep. 2009), 1–3880.

scaling limits of some physical media and the emergence of entirely new forms of others, are causing rapid and significant changes in the future landscape of storage hardware.

4.1.1. Magnetic Disk

Magnetic disk is transitioning from a medium designed for both capacity and bandwidth into one solely for capacity as a result of two factors:

- Magnetic storage media is reaching a physical limit on how small individual magnetic domains on the disk surface can be.
- High-performance NAND is proliferating, satisfying storage performance requirements and disincentivizing innovation towards better magnetic disk performance.

Combined with the scaling of I/O performance with the square root of the bit density on rotating media, the disparity between disk capacity and performance is only expected to widen.

That said, there are a number of capacity-focused improvements on the magnetic disk roadmaps of vendors and industry consortia. As shown in Figure 8, there are technology improvements that are projected to deliver a 10x increase in areal density over the next 10 years.

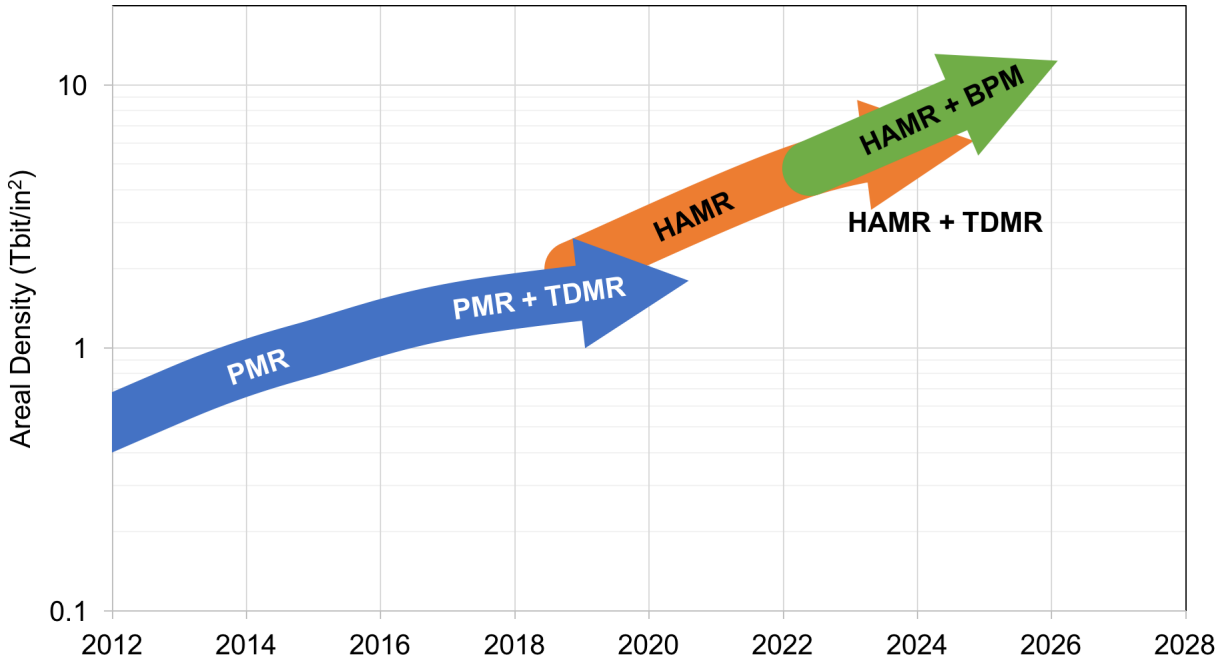

FIGURE 8. PROJECTED AREAL DENSITY IMPROVEMENTS FOR MAGNETIC DISK STORAGE TECHNOLOGY. BASED ON PROJECTIONS FROM SEAGATE[28] AND ATSC.[29] PARALLEL MAGNETIC RECORDING (PMR) IS THE STANDARD TECHNOLOGY OF TODAY.

The modest 10% areal density (AD) improvement from two-dimensional magnetic recording (TDMR)[30] is likely to reach the enterprise market in the near term, and heat-assisted magnetic recording (HAMR)

[28] Anderson, D. 2016. Whither Hard Disk Archives? *32nd International Conference on Massive Storage Systems and Technology*. (May 2016).
[29] 2016 ATSC Technology Roadmap. http://idema.org/?page_id=5868. Accessed September 3, 2017.

and bit-patterned magnetic recording (BPM)[31] promise to deliver more aggressive increases in bit density in the longer term. However, both HAMR and BPM represent largely new recording techniques rather than small refinements to existing approaches, and there is a nontrivial risk that HAMR will not be a commercially or economically viable option in 2020.

Thus, it is more likely that vendors will continue increasing the per-drive storage capacity by relying on refinements to shingling (e.g., via TDMR) and increasing platter counts. These two approaches will result in high-capacity drives with reduced write performance, flat read performance, and slightly increased power consumption. While suitable for the WORM workloads prolific in enterprise applications and content distribution networks, the evolution of spinning disk media is moving away from the balanced read-write workloads described in Section 3.1 and common to scientific computing in general.

4.1.2. Solid-State Storage

NAND-based solid-state storage devices (flash) have become a growing presence in HPC in the form of node-local scratch storage[32] and centralized burst buffers[33] designed to reach a better performance-per-bit than magnetic disk media. As demand for flash media continues to increase, driven by both mobile electronics and hyperscale markets, the lower power consumption and high performance of flash are expected to continue to push magnetic disk into lower-performance roles.

The low power consumption and high bit density of flash make it an attractive archival media. Although the cost-per-bit of flash is still significantly higher than that of magnetic disk and tape, the cost-per-bit of flash storage can be reduced by sacrificing performance and endurance. Hyperscale consumers (e.g., Facebook[34]) are driving the development of quad-level cell (QLC) flash as a low-power, high-density medium for WORM- and archival storage, and the first QLC NAND products have recently been announced by vendors including Samsung[35] and Toshiba.[36] By the 2020 timeframe, it is entirely conceivable that QLC flash may find a role alongside higher performance, higher endurance MLC and TLC NAND in tiered, all-flash storage systems.

The cost-per-bit of flash is also expected to drop precipitously before 2020 as the global NAND manufacturing industry completes the process of converting 2D (planar) NAND fabrication plants to 3D NAND. This will likely push prices for performance flash below $0.10 per GB, encroaching on a market traditionally held by magnetic disk.[37] Advances in 3D NAND fabrication technology, driven by healthy

---

[30] Victora, R.H. et al. 2012. Two-Dimensional Magnetic Recording at 10 Tbits/in^2. *IEEE Transactions on Magnetics*. 48, 5 (May 2012), 1697–1703.
[31] Albrecht, T.R. et al. 2015. Bit-Patterned Magnetic Recording: Theory, Media Fabrication, and Recording Performance. *IEEE Transactions on Magnetics*. 51, 5 (May 2015), 1–42.
[32] Strande, S.M. et al. 2012. Gordon: design, performance, and experiences deploying and supporting a data intensive supercomputer. *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment - XSEDE (Chicago,* 2012), 1.
[33] Bhimji, W. et al. 2016. Accelerating Science with the NERSC Burst Buffer Early User Program. *Proceedings of the 2016 Cray User Group* (London, 2016).
[34] Rao, V. 2016. "How We Use Flash at Facebook: Tiered Solid State Storage." *Flash Memory Summit 2016*. (August 2016).
[35] Elliot, J. 2017. "Advancements in SSD and 3D NAND Reshaping Storage Market." *Flash Memory Summit 2017*. (August 2017).
[36] Toshiba Develops World's First QLC BiCS FLASH 3D Memory with 4-Bit-Per-Cell Technology. https://toshiba.semicon-storage.com/us/company/taec/news/2017/06/memory-20170627-1.html. Accessed September 9, 2017.
[37] Handy, J. Flash Market Current & Future. *Flash Memory Summit 2017*. (August 2017).

competition in the marketplace, will allow 3D NAND to scale well beyond 2020 as well; approaches such as string stacking are expected to allow areal densities of flash to scale to at least 5-10x the state of the art today.

The NVMe over Fabrics (NVMeoF) protocol is a rapidly evolving standard that enables block-level access to NVMe devices over any network fabrics that support remote direct memory access (RDMA), including InfiniBand and Intel OmniPath. In combination with RDMA fabrics whose bandwidth and performance align with the performance of NAND, NVMeoF is expected to enable fabric-attached NVMe devices as a viable, high-performance, disaggregated storage architecture.

Furthermore, it is technologically feasible to use NVMeoF to transfer block-based data to remote targets without CPU intervention *and* without copying blocks through host memory. Although such a feature requires extensive hardware support and driver compatibility between NVMe devices and RDMA-enabled network interfaces, it has the potential to enable hyperconverged node designs for HPC that do not suffer from I/O-induced jitter. Although such zero-jitter architectures are in key vendors' roadmaps, it is important to stress that these solutions remain unproven in production environments. Furthermore, block-level data transfer will still require storage system software to run on top of NVMeoF which is *not* jitter-free.

A complementary technology is the Storage Performance Development Kit (SPDK),[38] which is an emerging set of libraries that provide a mechanism for applications to perform I/O to NVMe and NVMeoF devices entirely in user space. This significantly reduces the I/O latency of interacting with flash media by completely removing the need for data to transit the system kernel, and it is one of several efforts to provide a completely new interface to storage media that exposes the full capabilities of the hardware. SPDK is not widely used in production storage systems at present, but it is an instrumental component in future products, including DAOS.[39]

4.1.3. Storage Class Memory and Nonvolatile RAM

Storage class memory (SCM) technologies, which include Intel/Micron's 3D XPoint, are on the horizon and promise to deliver nonvolatile and byte-addressable storage whose performance lies somewhere between today's DRAM and NAND. Although such technologies deliver higher performance and durability than NAND, the significantly higher cost per bit (and therefore lower capacity) render SCM a pure performance technology that is likely to be integrated into larger, flash-based storage systems to remediate the software overheads incurred by processes such as data journaling. While SCM will undeniably play a role in storage systems in the 2020 timeframe, it is likely to first appear as highly integrated components within a larger storage system. This is analogous to how flash was first integrated into enterprise storage as extensions of traditional RAM-based cache tiers such as in ZFS's ZIL/L2ARC.[40]

There is opportunity for SCM to be directly used by users and applications in the form of byte-addressable nonvolatile storage with extremely low latency, but the consistency semantics of reading and writing data from a global storage resource with a load/store interface present a number of new

---

[38] Storage Performance Development Kit. http://www.spdk.io/. Accessed September 10, 2017.
[39] Paciucci, G. HPC Storage Trends. *HPC Advisory Council Swiss Conference*. (April 2017).
[40] Leventhal, A. 2008. Flash storage memory. *Communications of the ACM*. 51, 7 (Jul. 2008), 47.

challenges that remain a subject of intense research.[41] Of note, the NVM Library[42] is an emerging interface for persistent memory that preserves most of the low-latency benefits of SCM and flash by enabling user-space I/O directly to such devices through key-value, block, and other semantics. Although the NVM Library is currently being used to develop experimental storage services on SCM,[43] libraries and applications that can make direct use of the byte-addressability of SCM are unlikely to be production-ready by 2020.

4.1.4. Magnetic Tape

LTO and enterprise magnetic tape media have a comfortable technological runway because they capitalize on the investments made toward improving magnetic disk media. Furthermore, state-of-the-art magnetic tape technology typically comes to market five or more years after the same technology reached the magnetic disk market, giving the tape industry a healthy lead time in the event that magnetic disk reaches any fundamental barriers to improvement.

As a consequence of tape technology trailing disk technology, though, the roadmap for magnetic tape is driven by economics, not technology. Taking LTO tape (which holds a vast majority share of the magnetic tape market) as an example, tape revenue has been steadily decreasing despite steadily increasing volumes of capacity shipped, as shown in Figure 9.

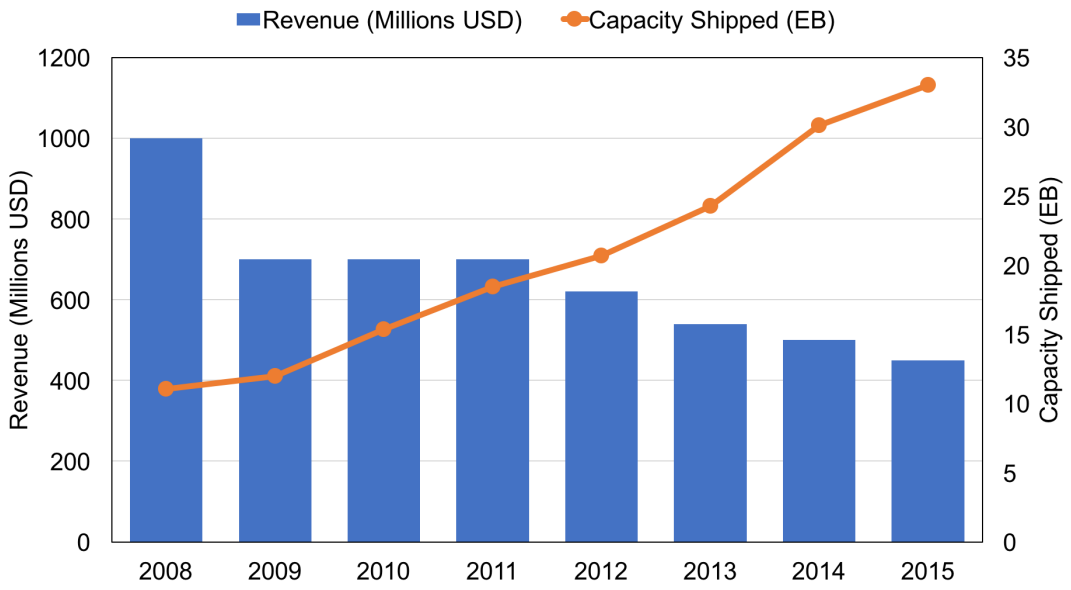

FIGURE 9. ANNUAL REVENUE AND EXABYTES SHIPPED OF LTO TAPE MEDIA. DATA FROM FONTANA AND DECAD.[44]

[41] Chowdhury, M. and Rangaswami, R. 2017. Native OS Support for Persistent Memory with Regions. *Proceedings of the 2017 International Conference on Massive Storage Systems and Technologies*. (2017).
[42] pmem.io: NVM Library. http://pmem.io/nvml/. Accessed September 9, 2017.
[43] Carns, P. et al. 2016. Enabling NVM for Data-Intensive Scientific Services. *4th Workshop on Interactions of NVM/Flash with Operating Systems and Workloads (INFLOW'16)* (Savannah, GA, 2016).
[44] Fontana, R., Decad, G. 2016. Storage Media Overview: Historic Perspectives. *32nd International Conference on Massive Storage Systems and Technology*. (May 2016).

In addition, the diversity of the tape manufacturing market has shrunk dramatically over the last decade: as of 2014, only Sony and Fujifilm continue to manufacture magnetic tape media, and as of 2017, IBM remains the only vendor to develop tape drives and cartridges. As a direct consequence of the steady decline of tape revenue and market competition, it is likely that the rate of innovation in magnetic tape will decelerate relative to magnetic disk. The perceptible effects of this decline are less certain though, and it is not clear if the cost advantages of tape for archival storage will be surpassed by another media in the next five to ten years.

If one assumes that data generation rates are ultimately bounded by the available capacity being produced, and the majority of storage capacity is provided by magnetic disk as evidenced in Figure 10, the deceleration of magnetic tape capacity shipments relative to magnetic disk presents a significant risk because it follows that a constant investment in disk-based storage will require *increasing* investment in tape-based storage to provide a constant ratio of disk to tape. Thus, while tape remains cost-effective for archival in the near term, it is unlikely to be the optimal long-term solution. However, the cross-over point is not imminent, and it is not clear that this point will occur before 2025.

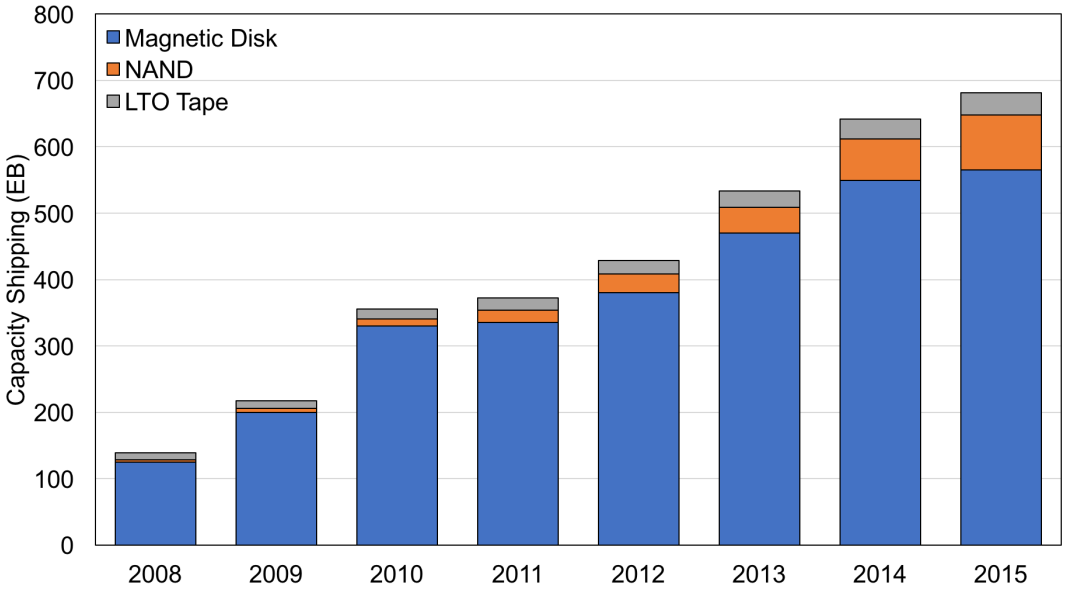

FIGURE 10. ANNUAL EXABYTES OF STORAGE MEDIA SHIPPED. DATA FROM FONTANA AND DECAD.[45]

The low cost-per-bit of tape, combined with its minimal power consumption as an offline storage medium, continues to make it an attractive archival storage technology in the short term. Given the uncertainties outlined above, though, tracking the economics of the tape market and following vendor roadmaps are essential for longer-term planning.

4.1.5. Storage System Design

Storage system architectures in 2020 will be shaped by the technological developments outlined in this section in several key ways:

[45] Fontana, R., Decad, G. 2016. Storage Media Overview: Historic Perspectives. *32nd International Conference on Massive Storage Systems and Technology*. (May 2016).

1. NAND devices will stratify into performance-oriented, high-endurance MLC/TLC and low-performance, high-capacity QLC, both of which consume less power and possess higher bit density than magnetic disk.
2. Magnetic disk media will disappear from performance-critical data paths and become a capacity-only medium.
3. Magnetic tape, which has historically been a capacity-only medium, has an uncertain future as its revenues drop. However, dramatic shifts in the economics of tape are unlikely to manifest before 2020-2025.

Based on these technological and economic trends, the role of these different media will also evolve:

1. MLC/TLC NAND will replace magnetic disk in all performance-critical applications, and QLC NAND will begin to supplant magnetic disk in many WORM application areas.
2. Magnetic disk will begin to eat away at the most performance-sensitive applications of magnetic tape, including hot archive and replicated tape.
3. Magnetic tape's role in the data center will continue to shrink toward deep archive applications as QLC NAND and magnetic disk approach it in cost.

4.2. Software

Beyond the changes coming in the hardware realm, there are many improvements and additions needed in extreme-scale storage and I/O software as well. The increasing difficulty in scaling POSIX-based parallel file systems to extreme scales is becoming a significant impediment, and, as discussed in Section 4.1.3, new software interfaces are a requirement to make optimal use of emerging low-latency storage hardware. Because these new non-POSIX interfaces are optimized for performance over usability, though, I/O middleware will become more important to bridge the semantic gap between the I/O operations that scientific applications demand and the I/O operations supported by the underlying storage system.

4.2.1. Non-POSIX Storage System Software

The stateful file-based nature of POSIX I/O, combined with its prescriptive metadata schema and strong consistency semantics, make it difficult to scale POSIX-based file systems to the extreme levels of parallelism anticipated for exascale systems. Object stores, initially driven by the extreme-scale I/O needs of cloud providers, eschew POSIX I/O semantics in favor of stateless put/get operations and immutable data objects. By exposing these I/O primitives directly to applications, they provide a much more scalable foundation on which more feature-rich storage services and systems can be built.

As a result, we expect to see scalable object-based storage systems, such as DAOS[46] or Ceph,[47] take on a more prominent role in HPC systems in the near future. POSIX file-based interaction will still be an option for users' source code, configuration files, and input decks, but this POSIX interface will be implemented as middleware atop a native object interface rather than being the lowest-level user interface to storage. As POSIX moves from a native interface to a middleware layer, we anticipate the hardware advances described in Section 4.1 to drive a gradual replacement of parallel file systems with object stores for both performance and capacity without requiring immediate, disruptive changes to user applications.

---

[46] Gorda, B. 2015. DAOS: An Architecture for Exascale Storage. *31st International Conference on Massive Storage Systems and Technology*. (May 2015).
[47] Weil, S. A. et al. 2006. Ceph: A Scalable, High-Performance Distributed File System. *Proceedings of USENIX Symposium on Operating Systems Design and Implementation*. (2006).

4.2.2. Application Interfaces and Middleware

As POSIX evolves into middleware, we also see a greater percentage of the application community moving to use other I/O middleware packages like HDF5[48] and ADIOS.[49] This shift allows application teams to use more semantically meaningful APIs (e.g., store a whole array rather than manually serialize data structures) and benefit from the effort and experience of the middleware package developers. The increasing adoption of I/O middleware packages will also insulate applications from the underlying shift away from current POSIX consistency semantics, allowing them to automatically gain the benefits of new hardware without having to directly interact with the storage system's native API.

Increased storage of observational data and a push toward improved reproducibility of science results also leads to a need for storing provenance information on all data, as identified in Section 3.3. Enhancing I/O middleware to automatically add provenance to application data will go a long way toward improving the current wild-west conditions of data curation by providing always-available, queryable information on the storage system. These data curation improvements will add to the momentum for a long-lived Community/Forever storage that is independent of Temporary/Campaign storage.

5. Next Steps

As discussed in previous sections, the diversity of NERSC's workload will continue to drive NERSC's storage requirements in several different dimensions. File system performance must be measured not only in bandwidth but metadata performance, latency, and variability as well. Partnerships with experimental facilities and the continued growth of data science workloads will also add new data retention requirements in terms of both durability and manageability. In addition, the size of NERSC-9 will demand new levels of scalability and resilience. These requirements drive our vision for the future and our strategy in getting there.

5.1. Vision for the Future

While every HPC user desires a single, high performance, high capacity, and highly durable storage system, cost will continue to require tiered storage at HPC centers. As has been the case for the past two decades, HPC will continue to deploy storage systems built from enterprise components whose economics are now being driven largely by consumer and cloud markets. In the 2020-2025 timeframe, the most notable shift will be the move in platform storage away from HDDs and toward higher-performance but economical nonvolatile memory technologies.

The massive disk-based parallel file system, which has served the HPC community for more than two decades, will see its role diminished. It will no longer be the high-bandwidth resource used for all job I/O, as emerging storage technologies expressly built for NVM—such as Intel DAOS[50], IBM's burst buffer[51], and Cray DataWarp[52]—become the principal interface to on-platform storage. For off-platform

---

[48] Folk, M. et al. 2011. An overview of the HDF5 technology suite and its applications. *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*. (2011).
[49] Lofstead, J. et al. 2009. Adaptable, metadata rich IO methods for portable high performance IO. *2009 IEEE International Symposium on Parallel & Distributed Processing* (May 2009), 1–10.
[50] Gorda, B. 2015. DAOS: An Architecture for Exascale Storage. *31st International Conference on Massive Storage Systems and Technology*. (May 2015).
[51] Goldstone, R. 2016. The Road to Coral…and Beyond. *HPC Advisory Council Stanford Conference*. (February 2016).

storage, cost-effective and scalable solutions such as object stores will begin to replace it. On-platform Temporary/Campaign storage will almost certainly be built entirely out of performance NAND and SCM, while off-platform Community/Forever storage will be a mix of QLC NAND, magnetic disk, and tape in a combination dictated by cost, technological evolution, and performance/capacity balance.

An increasing number of scientific applications will interact with storage through an I/O middleware layer, allowing highly scalable storage (which provides POSIX compliance as an option, not a default) to transparently serve as the backing store. Nonvolatile memory will make inroads throughout the storage hierarchy, and as it does, storage software will be reengineered to wring out performance bottlenecks that appear when latencies are no longer dominated by the physical characteristics of disk drives. We are beginning to see this in the form of low-latency, user-space I/O libraries such as Mercury[53] and the NVM Library,[54] and this trend toward optimizing software for low latency will become a requirement to match the low latency of emerging nonvolatile memory technologies.

Archival storage software, one of the last vestiges of purpose-built system software for HPC, will be radically impacted by software innovations from cloud providers. The same put/get interfaces used to store data in cloud services such as Amazon S3 also suffice for storage in the onsite archive, and the archive will provide access via these standard object APIs, including S3 and Swift. For long-term storage, the lines may well be blurred between data that resides within the local facility and data that resides offsite, either in a commercial cloud or at another open science center. Data replication currently offered by commercial object stores and cloud providers, including attributes to guarantee geographical separation, will become part of the archival software suite.

Throughout the HPC storage stack, there will be an emphasis on ease of movement between storage tiers. A new set of standards-based APIs to interact with the performance, capacity, and archival tiers will help with adoption and portability, and efforts are already underway within DOE and amongst vendors to develop these APIs. Job-scheduling software will be able to move data between all tiers as part of a run, with resource managers including Slurm, Torque, and PBSpro already beginning to support this. The combination of standard APIs and scheduler-moderated data motion will enable users to steer jobs and marshal data between tiers more expressively. This rich, procedural interface will ensure that data is in the correct place as different workflow stages ingest, manipulate, and store data in different ways.

The hierarchical file system of today will only be one of a number of views through which users can interact with their data. Alternate views of data, searchable by user-defined attributes associated with data, are a feature of today's cloud-based storage that will find their way into the HPC space. There are a handful of efforts to provide rich metadata capabilities atop existing parallel file systems, but they are implemented as an external software layer and have seen limited adoption in production HPC. We anticipate that search and discovery based on user-defined metadata will be better integrated directly into the storage system, and this will catalyze broader user adoption and provide a more stable foundation on which domain-specific metadata catalogs can be developed.

---

[52] Henseler, D. et al. 2016. Architecture and Design of Cray DataWarp. *Proceedings of the 2016 Cray User Group* (London, 2016).
[53] Soumagne, J. et al. 2013. Mercury: Enabling remote procedure call for high-performance computing. *2013 IEEE International Conference on Cluster Computing (CLUSTER)* (Sep. 2013), 1–8.
[54] pmem.io: NVM Library. http://pmem.io/nvml/. Accessed September 9, 2017.

Although the high-bandwidth Temporary tier will continue to be purchased with the supercomputer, Community and Forever storage will be best managed as separate resources owing to the longevity of the data they will store. By decoupling these longer-term tiers' refresh cadences from the compute systems' procurement cycles, we will be able to deploy the most feature-rich storage resources the market offers, integrate new technology over time, and realize the cost benefits of purchasing storage only when it needs to be deployed.

5.2. Strategy

The changes required to realize this vision for the future of storage in HPC will require innovations that involve hardware vendors, software and middleware developers, and the larger research community. The following strategy, divided into near-term (present day through 2020) and long-term (2020-2025) targets, strives to ensure a smooth transition for NERSC users and to identify areas where NERSC leadership and community engagement would be most beneficial. The evolution of the storage hierarchy during this period is summarized in Figure 11.

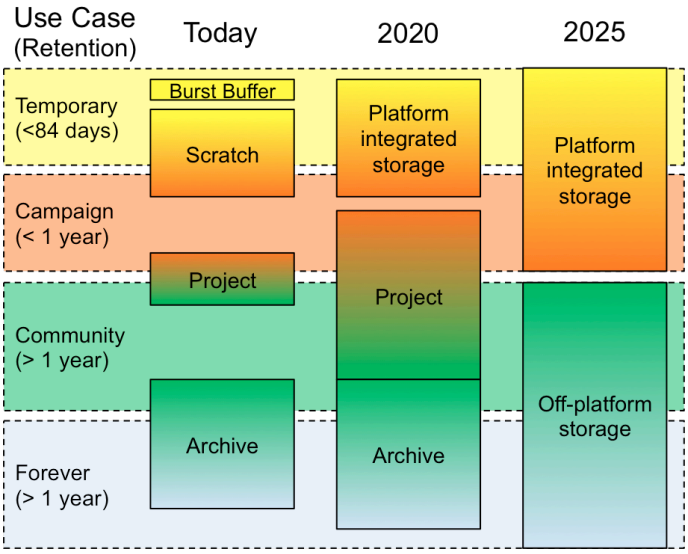

FIGURE 11. EVOLUTION OF THE NERSC STORAGE HIERARCHY BETWEEN TODAY AND 2025.

In the following sections, we detail the actions required to realize this evolution.

5.2.1. Near Term (2017 – 2020)

The most significant change to the storage hierarchy in the 2020 timeframe will be a collapse of the burst buffer and disk-based scratch file system back into a single, high-performance, modest-capacity tier. Through the highly successful Burst Buffer Early User Program at NERSC[55] and ongoing production use of the burst buffer on Cori, solid-state media has demonstrated its viability for Temporary storage, and a single-tier, all-flash platform storage system would simplify data management for users without sacrificing substantial functionality. Given the trends of the NAND industry discussed in Section 4.1, this should be economically viable as well.

[55] Bhimji, W. et al. 2016. Accelerating Science with the NERSC Burst Buffer Early User Program. *Proceedings of the 2016 Cray User Group* (London, 2016).

In addition to this all-flash platform-integrated tier, a disk-based, POSIX-compatible storage system will also need to exist during this time period to satisfy the needs of the colder portions of the Campaign tier and the hotter portions of the Community tier. Unlike the NERSC project file system of today, though, this tier will be optimized for capacity and manageability, not performance. It will meet the needs of data that must be retained beyond the design of NERSC-9's temporary tier, such as high-value experimental observations, community-curated datasets, and other emerging use cases outlined in Sections 3.3 and 3.4. This capacity-optimized tier will present a familiar file system interface to support existing data management and transfer tools, but it will also provide access via more future-looking, object-based APIs to allow users to begin transitioning applications to put/get semantics.

The 2020 Campaign/Community storage will also satisfy many of the operational requirements discussed in Section 3.5. NERSC presently relies on key storage manageability features, including metadata replication, dynamic storage resizing, snapshotting, and enforcing project-based quotas. The 2020 Campaign/Community storage system will expand upon these manageability features and provide a foundation to begin developing additional system monitoring and management tools for the future. It will also serve as the basis for future data curation tools and interfaces that NERSC will provide to users and support features to facilitate object or file metadata searches and queries.

Due to the different performance, capacity, and feature requirements of this 2020 Campaign/Community tier, it will be acquired and managed as a resource that is independent of system platform storage through the 2020 timeframe. Unlike compute, storage is not a resource that is fully utilized as soon as it arrives, and incremental growth guided by user needs and center policy will take advantage of the expected 10%-30% annual reduction in cost-per-bit and allow economical resale of extra storage to projects that need it. This planned growth allows us to adopt new storage and network technologies incrementally, deploy novel solutions earlier, and increase NERSC's agility to innovate on the new techniques and technologies in storage described in Section 4.

The 2020 Forever storage will remain predominantly tape-based due to tape's economic advantages. Tape technology will continue to be more cost effective than disk through 2020, and transitioning an exabyte of data (or more) to a new storage medium would require significant capital investment and time. There may be opportunity to explore alternative archive media, but there are no truly compelling options in the near term. Other key technologies that may become technologically viable for archive, such as low durability NAND[56] or hyperscale disk-based object stores, will still not be cost-competitive versus tape by 2020.

NERSC will undoubtedly continue to deploy tape-based storage beyond 2020, but it is unlikely that tape's economic scaling rates will continue. Although NERSC's Forever storage has been treated as a limitless data store for users in the past, the economics of the tape market are making this an unsustainable policy. We have already begun to take steps to sharpen the focus of the NERSC archive, resulting in a 10% reduction in size, and further refinements will be made based on close monitoring of the tape market.

The sum of these findings drives us toward the storage hierarchy for NERSC in 2020 shown in Figure 12.

---

[56] Peglar, R. 2016. Innovations in Non-Volatile Memory: 3D NAND and its Implications. *32nd International Conference on Massive Storage Systems and Technologies* (2016).

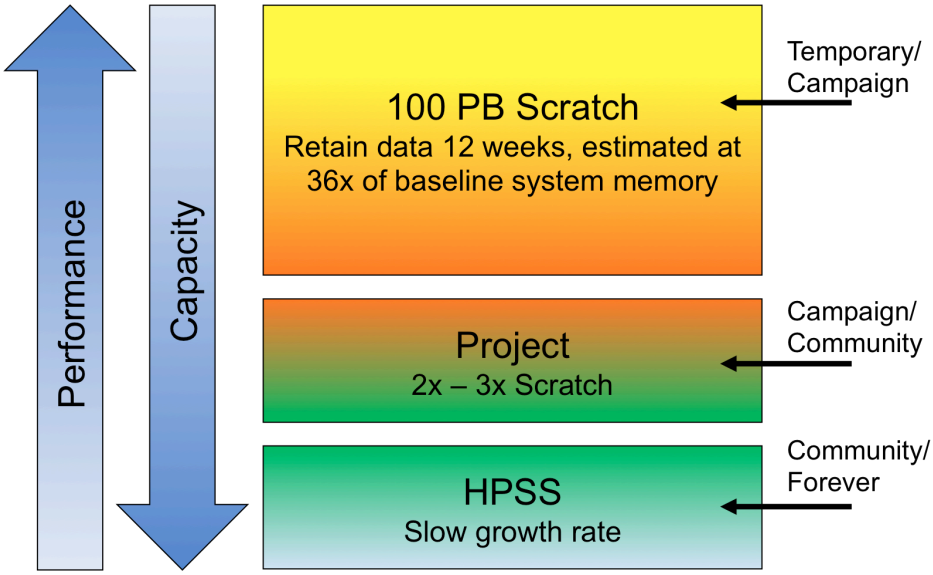

FIGURE 12. TARGET THREE-TIER STORAGE HIERARCHY FOR NERSC IN 2020.

To meet these near-term requirements and evolve the storage hierarchy toward this design, several critical actions must be taken before 2020:

1. The present NERSC project file system must be expanded significantly to reflect its role relative to the platform-integrated Temporary/Campaign tiers on Cori and NERSC-9. Because this storage system is optimized for manageability, accessibility, and usability, its capacity should reflect the desire of users to store the bulk of their working data on it, and the aim is for a size of 2-3x the performance tier. This is in contrast with today's hierarchy, where users store data for as long as possible on the performance tier (before data gets purged), and then move data to the forever tier.
2. Investments must be made toward fully utilizing the data management features present in NERSC's project file system and archive. Building new data management tools that unify these tiers will be essential; this includes improving accessibility (via new interfaces such as industry-standard object APIs) and introspection (via expanded indexing, monitoring, and characterization capabilities).
3. Given that the project file system will hold the Community tier, we expect decelerated growth for the tape-based archive. Policies and stricter quotas may be necessary to ensure that maintaining Forever storage is economically sustainable.

The result of these efforts will be a single, high-performance, platform-integrated storage system that satisfies the role of Temporary storage and some very hot Campaign storage; a high-capacity but scalable and manageable storage system that satisfies the role of Campaign and Community storage; and a closely integrated, high-capacity, high-durability storage system that satisfies the role of very cold Community storage and Forever storage.

5.2.2. Long Term (2020 – 2025)

The next evolutionary step beyond the 2020 Storage architecture will aim to transform the closely integrated Community and Forever storage systems into a single Community/Forever tier for long-term data retention, curation, and sharing. This results in a two-tier storage hierarchy, as shown in Figure 13.

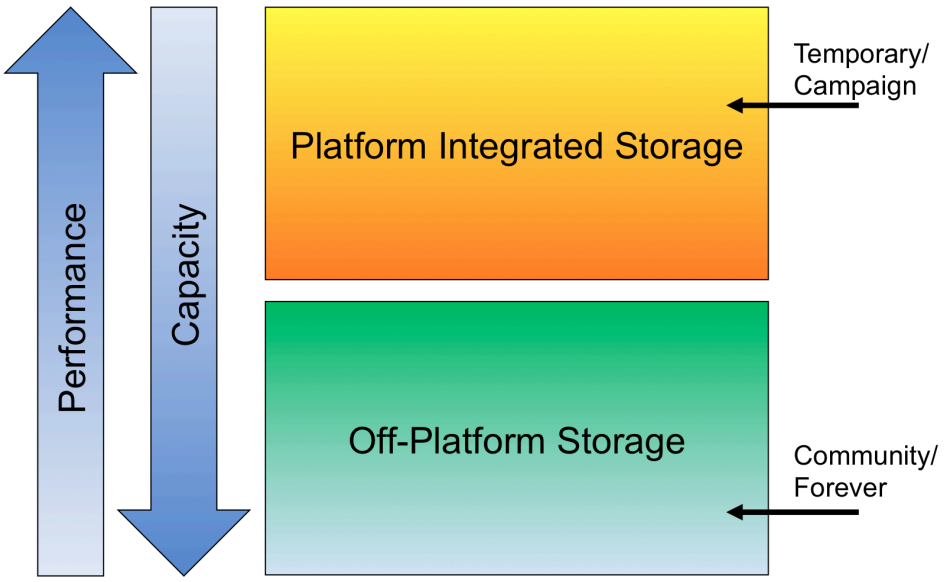

FIGURE 13. TARGET TWO-TIER STORAGE HIERARCHY FOR NERSC IN 2025.

As with the 2020 storage infrastructure, the platform-integrated tier will emphasize performance first. It will provide a native interface that delivers extreme performance through asynchronous I/O, relaxed consistency semantics, and a user-space client implementation.[57] Users will still be able to access this tier through a familiar POSIX interface implemented as middleware, but this file-based API will not deliver the full performance of the underlying NAND- and SCM-based hardware. Rather, applications that require extreme performance will have to either use I/O middleware that supports the native interface or restructure their I/O to use the native interface directly. Given the disruptive nature of such a change, the semantics of this new API should be well defined by 2020, and experimental systems must be available to allow users to begin testing and modernizing their application I/O.

At the Community and Forever tiers, preparing for a transition away from established solutions like tape-based storage and HPSS toward object-storage solutions backed by shingled disk or archival NAND will require a careful assessment of the potential replacement technologies and production hardening. As a point of reference, DOE has invested decades in the development of HPSS to meet its mission needs, but adopting off-the-shelf technologies (such as open-source or commercial object-storage solutions) will pay future dividends by aligning our approach to mass storage with those of the cloud and hyperscale communities. Moving users to an object-based interface for the archive will allow us to transparently migrate away from tape-based media should tape continue to decline. However, building these bridges requires connecting users with these technologies, and ensuring they meet user and operational requirements will require investment on the part of NERSC and the HPC community.

Preparing the NERSC storage hierarchy to transition into this long-term vision by 2025 requires additional actions within the next five years:

1. The NERSC Data Archive mission must be redefined to align its growth trajectory with the long-term target capacities and investments so that the transition to 2025 is seamless. This will

[57] See discussion of Mercury, NVML, DAOS, and other software interfaces discussed in Sections 4.1.3, 4.2.1, and 5.1.

involve user engagement with those users whose data needs will exceed storage capacity projections, and it will involve developing software and infrastructure to assist users in managing and migrating their data.

2. Test platforms must be fielded to explore new I/O paradigms, including performance-oriented object stores and software systems capable of effectively utilizing next-generation nonvolatile memory technologies. This will allow NERSC to establish a credible understanding of how difficult a future transition to such systems would be for our users, and also allow us to develop tools that address those difficulties. Such a system would also inform the return on investments users can expect from this effort and maintain our understanding of these technologies' maturity.
3. We must develop the tools and infrastructure that allow the performance/projects tiers and campaign/archive tiers to collapse. For example, many components of DAOS would glue together the performance aspects of DAOS' asynchronous object interface with a lower-performance but higher-durability flash layer. Similarly, a software technology such as IBM's GHI would have to be proven out to integrate a GPFS-based campaign tier with an HPSS-based archive tier.

5.2.3. Opportunities to Innovate and Contribute

NERSC is uniquely positioned to lead a transition to this storage architecture because of its broad user base, deep understanding of user requirements, and proven ability to partner with application developers in code modernization efforts. As such, our role in leading a transition to future storage technologies is centered around two key areas:

1. Driving requirements that will steer emerging software, middleware, and hardware technologies in a direction that will be broadly accessible and useful across all segments of HPC and scientific computing markets.
2. Demonstrating and hardening emerging software, middleware, and hardware technologies in extreme-scale but highly diverse workload environments that span traditional high-performance simulation, high-throughput experimental data processing and synthesis, and machine learning-driven data analytics at scale.

Ultimately, leading the ground-up design of novel storage systems or defining new storage paradigms at the bleeding edge of computational science is not within the NERSC mission. Rather, our expertise lies in understanding how such radical changes will affect each of the scientific domain areas' workflows at all scales, and this is where NERSC's leadership will be essential to ensure that emerging I/O technologies will be viable and sustainable as they mature into the broader HPC ecosystem. This contribution is essential to help new storage systems and APIs meet their full potential by broadening user adoption. Opportunities to drive requirements are manyfold, and we categorize these opportunities as being in software, middleware, and hardware.

At the **software level**, NERSC's broad user base serves as a unique sounding board for emerging I/O APIs and software technologies. The NERSC Burst Buffer Early User Program has been an exemplar of how well NERSC is suited to proving out new storage systems, new modes of user-defined configuration, and new mechanisms of data access. The program provided the vendor with continuous feedback about how different user communities wanted to interact with flash storage and both drove its design and demonstrated its viability to the greater HPC community. Not only did this work strengthen the burst buffer software (much to the benefit of the user community *and* the vendor), it demonstrated that software-defined storage and flash-based file systems are viable technologies for the future. This effort is augmented now by the Tiered Storage Working Group, a partnership of DOE labs and burst buffer vendors, to define standards-based APIs for interacting with future multi-tier storage platforms.

It is critical that NERSC continue to make investments in partnering with storage software providers to ensure that our users' needs are represented in designs. The strategic importance of this cannot be overstated as the HPC industry begins to explore radically new alternatives to the traditional parallel file system and as the enterprise industry drives object-based archival solutions into the HPC space. Failing to engage both software vendors and users to explore new storage paradigms presents a significant risk that these storage solutions will evolve in directions not suitable for the broad user community and that compute- and data-intensive computing will bifurcate at the storage layer.

The **middleware level** represents an ideal area where NERSC should lead in bridging the gap between rapidly changing storage hardware and the diversity of user applications that change much more slowly. A case in point was a recent demonstration of using the HDF5 middleware to interface directly with DAOS[58]; because a significant number of NERSC data is stored as HDF5, a substantial amount of the work required to port applications to entirely new I/O APIs and paradigms can be done in the middleware layer, effectively enabling broad adoption at only a modest investment from NERSC. Given the broad and increasing use of I/O middleware in HPC, this investment would be of significant benefit to the greater HPC community as well.

It is therefore essential that we continue to engage with the broad user community to transition applications to use I/O middleware where appropriate. Furthermore, we must continue close engagement with middleware developers to ensure that the essential features of users, including metadata, provenance tracking, and ease of use, guide the development of these middleware. Failure to invest in this will hold open a gap between today's applications and the native interfaces of non-POSIX storage systems, reducing the performance and scalability benefits offered by new, nonvolatile hardware.

At the **hardware level**, NERSC has begun an effort to integrate the monitoring of the storage tiers into a holistic understanding of emerging I/O demands, and continuing this work will provide critical feedback to vendors. For example, monitoring the workloads and wear rates on Cori's burst buffer has identified that HPC workloads would benefit greatly from multi-stream support in SSD firmware,[59] and ongoing vendor engagement and sharing of endurance data has found that HPC workloads would be better served by trading high write endurance for added capacity on enterprise SSDs. Furthermore, these monitoring efforts are improving the performance, reliability, and usability of NERSC's storage systems by establishing detailed baseline behavior and maintaining relationships with vendors that facilitate rapid diagnosis, resolution, and improvements when aberrations arise.

Tracking NERSC production workload telemetry, curating and contextualizing it, sharing it with the larger vendor and research community, and actively maintaining productive engagements with vendors and researchers have provided significant returns for NERSC and the larger HPC community. In the absence of NERSC investment, the evolution of new storage technologies may be shaped by boutique workloads and the enterprise market. This would result in overall loss of value in future generations of NVM, network technologies, and SCM.

---

[58] Breitenfeld, M.S. et al. 2016. Use of a new I/O stack for extreme-scale systems in scientific applications. *Proceedings of the 1st Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems* (2016).
[59] Han, J. et al. 2017. Accelerating a Burst Buffer via User-Level I/O Isolation. *2017 IEEE International Conference on Cluster Computing (CLUSTER)* (2017), 245–255.

6. Conclusion

The increased amount of data generated at experimental facilities and the prevalence of high-speed network connections between their instruments and centers such as NERSC point to an explosive increase in the volume of experimental data stored at computing sites. This, combined with the massive increase of data produced by exascale computations, requires rethinking the HPC storage hierarchy to maintain acceptable performance and cost. We have established four logical tiers of data storage based on required performance, capacity, shareability, and manageability and mapped these logical tiers to physical storage systems based on the prevalent trends in storage technologies.

In the short term, collapsing platform-integrated, high-performance, flash-based storage systems into a single tier that satisfies the requirements of Temporary and hot Campaign storage is feasible and desirable to simplify I/O for scientific workflows and data management. Moving the colder, disk-based Campaign/Community and tape-based Forever storage tiers into a more closely integrated group of systems is also tractable by 2020 and positions NERSC for a two-tier storage hierarchy in 2025.

This two-tiered 2025 storage system establishes a converged Temporary/Campaign storage system and a Community/Forever storage system, allowing NERSC to separately optimize extreme I/O performance from the orthogonal needs of long-lived, high-value community datasets. This transition will be critical to meeting the needs of NERSC users using the best available storage technologies in both 2020 and 2025, and immediate investments in software, middleware, and hardware technologies are necessary to achieve the benefits foreseen by that transition.

As the principal provider of HPC services to the DOE Office of Science, NERSC will deploy these new storage technologies while continuing to provide fast and reliable storage resources that meet the needs of our broad spectrum of users. The diversity of workflows and unique datasets that rely on NERSC's computational and storage resources put NERSC in a strong position to understand how the changing storage landscape will affect the scientific domain areas' workflows at all scales. Executing the strategy presented in this document will ensure that emerging I/O technologies will be viable and sustainable solutions to meeting the needs of the DOE Office of Science as well as the broader HPC community.