Profiling the Usage of an Extreme-Scale Archival Storage System

Hyogi Sim, Sudharshan S. Vazhkudai
Oak Ridge National Laboratory
{simh,vazhkudaiss}@ornl.gov

Abstract—**Profiling the archival storage system in scientific computing environments has received much less attention compared to the parallel file system, but is equally important since it stores the final data products safely, for a long duration. In this paper, we analyze eight years worth of data transfer logs for accessing the archival file system (HPSS) in the Oak Ridge Leadership Computing Facility (OLCF), which has been hosting the world's largest supercomputers and file systems. Our analysis encompasses about 135 million data transfer activities to the 80 PB High Performance Storage System (HPSS), between 2010 and 2017. We analyze the logs from several dimensions, including studying the workload characteristics (e.g., access patterns, frequency of accesses and temporal behavior), file system characteristics (e.g., directory depth, file system scaling trends, file types), and scientific user behavior (e.g., domain-specific usage and organization). Based on the analysis, we derive insights into the future evolution of the archive in terms of provisioning, desired features and functionality from the archive software, role and right sizing of the archive tiers, quota management, and the importance of smart and efficient metadata and storage management. We believe our study will prove useful for both operating current archival storage and the better provisioning of future systems.**

I. Introduction

The unprecedented advances in computing technologies, such as multi-core processors, high bandwidth memory, accelerators, and fast networking, have enormously escalated the computing capability of supercomputers [1]. Evidently, the average High Performance Linpack (HPL) score of the five fastest supercomputers has increased by almost five times in the last five years, i.e., from 17.5 Pflop/s in June 2014 to 84.2 Pflop/s in June 2019 [2]. This massive computing power has led supercomputers to produce ever more output data from large-scale scientific simulations [3] and data intensive applications [4], imposing higher performance and capacity demands for HPC storage systems. Consequently, HPC storage systems have not only grown in scale, but also become more complex in architecture by introducing newer storage tiers, in addition to the standard parallel file system (PFS) and the archival storage. These include tiers such as the burst buffer and the campaign storage system [5], [6].

However, despite such richness in the deep-storage hierarchy, the primary role of the archival storage system remains steady, i.e., the last resort to persist invaluable scientific outcomes at the bottom of the storage hierarchy. For example, at the Oak Ridge Leadership Computing Facility (OLCF) [7], which is home to the Summit system (No. 1 in the Top500 list [2] with 148 Pflop/s and deployed in 2018), the Titan system (No. 12 in Top500 with 17.59 Pflop/s and deployed in 2012) and several other analysis clusters, the High Performance Storage System (HPSS) [8] has been constantly supporting the data archive and backup requirements from diverse science projects for more than two decades. During this time, more than 10 supercomputers and PFSs have been deployed at OLCF. This demonstrates the indispensable role of the archival storage system. Therefore, we believe that understanding the workloads of the archival storage system, or *archival workload*[1] hereafter, is essential not only for operating the current archival storage system, but also for designing, developing, and deploying future HPC storage systems. Unfortunately, however, relatively less attention has been paid in understanding the archival workload compared to the PFS I/O workloads [9], [10] in HPC centers. A few prior reports on production HPC archival workloads do not provide comprehensive and general insights due to their insufficient sample periods [11], specialized system environments [12], or moderate system scales [13].

In this paper, we have analyzed eight years worth of data transfer log records, from 2010 to 2017, of the HPSS archival storage system in OLCF, one of world's largest HPC supercomputing centers. Specifically, we have analyzed more than 130 million data transfer activities that 1537 active system users from diverse scientific backgrounds triggered. Furthermore, during our sample period, the OLCF was operating two top supercomputers, i.e., Jaguar [14] (No. 1 in 2009 with 1.75 Pflop/s) and Titan [15] (No. 1 in 2012 with 17.59 Pflop/s), and had performed multiple major upgrades to its centralized PFS, Spider [16], [17]. From our analysis, we found that the HPC archival workload exhibits its own distinctive characteristics, while also sharing some characteristics with enterprise backup workloads [18]. For instance, almost 40% of the incoming requests to the OLCF archival storage system were read requests, significantly higher than the observation from the enterprise backup storage system [18].

Contributions

We summarize our contributions in this paper as follows.

- We have analyzed eight years worth of data transfer logs from the production archival storage system running in one of the world's largest supercomputing centers. This analysis

[1] This paper uses the term, *archival workload*, to indicate all incoming workloads to the archival storage system for backup and archival purposes.

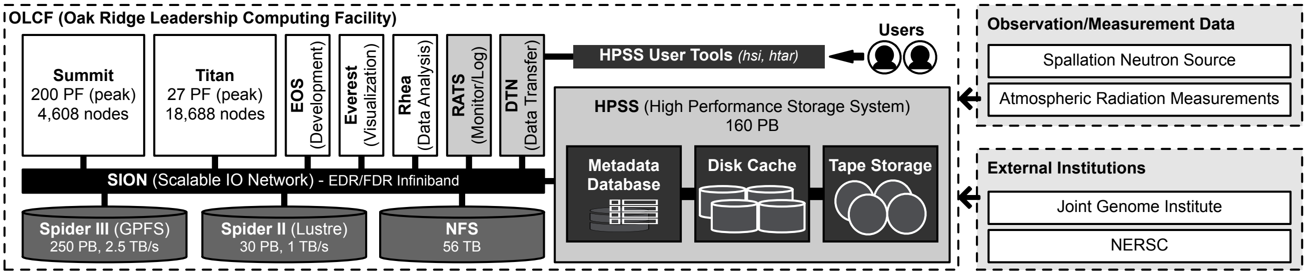


Fig. 1: An architectural overview of storage systems in Oak Ridge Leadership Computing Facility (OLCF) [7]. In addition to OLCF system users, external users from other institutions can also utilize HPSS for storing diverse scientific outcomes, such as large-scale simulation results and measurement data from scientific observatory devices.

Name	FS	Capacity	Backup	Retention	Purpose
Spider III	GPFS	250 PB	No	90 days	Scratch
Spider II	Lustre	30 PB	No	90 days	Scratch
Home	NFS	56 TB	Yes	No	User Data
HPSS	HPSS	160 PB	N/A	No	Backup

TABLE I: Different file systems in OLCF [7]. This table reports the file system status in early 2019. When a retention period (the *Retention* column) is associated with a file system, i.e., Spider III and Spider II, the system automatically purges files that have not been accessed for the retention period to keep enough free space.

period is the longest, to the best of our knowledge, in a scientific computing environment.

- In addition to analyzing incoming and outgoing data transfer activities, we have also inferred the file system characteristics, e.g., the number of files, directory depth, etc., from the data transfer logs.
- Besides immediate insights to the archival storage system itself, we also provide deeper insights from a whole system design perspective by associating our analysis with other system components, i.e., the PFS and system users.
- Lastly, our analysis provides valuable insights not only for better operating current archival storage systems but also for designing and provisioning future archival storage systems.

The rest of the paper is organized as follows. In § II, we explain the storage system architecture in OLCF, particularly focused on the HPSS archival storage system. We then provide an overview of our analysis methodology in § III, followed by the analysis results in § IV. We further summarize the insights from our analysis results in § V. After discussing related work in § VI, we conclude the paper in § VII.

II. Background

This section provides an architectural overview of our target system environment, focused on the storage system architecture.

A. OLCF Storage System Architecture

Figure 1 shows the overall architecture of the OLCF [7]. The computing resources in OLCF include the world's most powerful supercomputers (Summit [19] and Titan [15]) for running large-scale scientific simulations, and additional computing clusters for data analysis and visualization of the data emanating from the simulations. To facilitate a continuous workflow between the computing resources, OLCF provides centralized PFSs [20], i.e., GPFS-based Spider III and Lustre-based Spider II, and all clusters can access these file systems via consistent namespaces. Furthermore, a separate NFS file system hosts user home directories. Note that Summit and Spider III have been deployed in 2018, after the eight-year period of this study (from 2010 to 2017). Each of the aforementioned file systems implements its own quota and purge policies to assure sufficient performance and capacity [21], as shown in Table I. Therefore, for long-term data retention, users voluntarily need to move their data to High Performance Storage System (HPSS) [8], an archival storage system. This indicates that HPSS accumulates invaluable scientific data products, which users deem worthy, at the very bottom of the storage hierarchy. All data transfers between HPSS and other file systems are recorded by Resource Allocation and Tracking System (RATS) [22], along with other resource usage statistics. Table I summarizes notable characteristics of the file systems in OLCF.

COS	Small	Medium	Large	Huge
File Size	- 16 MB	16 MB - 8 GB	8 GB - 1 TB	1 TB -

TABLE II: Predefined Class of Service (COS) groups based on file size. HPSS internally implements differentiated data management policies, e.g., the number of copies, based on the COS values.

B. HPSS Archival Storage System

The HPSS archival storage system was first deployed at OLCF in 1997 and currently stores over 80 million files in its 160 PB available capacity. As depicted in Figure 1, HPSS internally consists of a disk cache tier and a tape tier. In 2017 (the last year of our sample period), the capacity of disk and tape tiers were about 20 PB and 160 PB, respectively.

When a file enters into HPSS, HPSS first stores the file in the disk cache and asynchronously copies down to the tape tier [2]. HPSS also assigns a Class of Service (COS) value to each file, which it references for differentiating management policies, such as the number of copies of a file and the victim selection in the disk cache tier. Currently, HPSS automatically assigns a COS based on the file size (Table II), but users can manually specify the COS for limited purposes, e.g., providing hints to HPSS. All file system and internal metadata are kept in a DB2 relational database [23] using dedicated

[2] We do not include the internal data migrations inside HPSS in our analysis.

SSDs. At OLCF, instead of a standard mountpoint, HPSS provides users with dedicated command line tools, *hsi* and *htar* [20], for migrating files. *htar* provides a familiar *tar*-like interface, and *hsi* features more comprehensive functionalities, e.g., controlling COS, parallel transfer, etc [24]. Most users prefer to access HPSS via a PUT/GET interface, similar to using object storage systems [25], although *hsi* also supports a POSIX interface. HPSS is accessible only from certain login nodes and dedicated data transfer nodes (DTN). For transferring Large and Huge files (Table II), HPSS exploits dedicated transfer agents (eight physical nodes) in parallel.

In addition to the output files from OLCF computing resources, HPSS also stores data from a number of experimental and observational facilities, as shown in Figure 1. In particular, HPSS at OLCF stores the Atmospheric Radiation Measurement (ARM) project [26] data, and continuously stores climate measurement data from atmospheric measurement observatories. Currently, the ARM data files occupy about 3% of the total used capacity in HPSS. For periodic migration needs, scientists often write scripts that automatically move data to HPSS [27]. However, we could not positively identify such automated workloads from our log records.

OLCF also allows limited accesses to HPSS from external institutions, such as JGI and NERSC [28] as shown in Figure 1, via Globus [29] and GridFTP [30]. However, such external data transfers accounted less than 1%, and we do not include them in this study.

C. Archival Data Transfer Logs

As discussed earlier (§ II-B), the most dominant method for OLCF users to access the archival storage system is through *hsi* and *htar* command-line utilities. OLCF has been recording all data transfer operations from those command-line utilities in a dedicated relational database [22]. From this database, we have exported all data transfer log records from/to HPSS, inclusively between 2011 and 2017, into a separate relational database for our analysis. After discarding irrelevant columns, each transfer record consists of 11 columns in our database, as shown in Table III. The final database contains 135 million records. In addition to the data transfer logs, we have also used other available system information, e.g., the UNIX user list, as necessary. For our analysis, we use MariaDB-5.5.56 running in a single server with eight cores (Intel Xeon E5-2609) and 256 GB RAM. We have created a dedicated XFS volume atop a 240 GB SSD to store the database. The final database size was 35 GB including all indices.

Column	Description	Example
DATE	Data and time of operation.	*2017-05-16 02:24:31*
HOST	Host where operation is triggered.	*host.ornl.gov*
UID	System user id.	*8951*
ACCNT	User account affiliation.	*NCCS*
TYPE	Operation type.	*PUT*
SPEED	Observed bandwidth in KB/s.	*200000*
AGENT	Access method.	*HSI*
SIZE	Data transfer size.	*150000*
COS	Class Of Service, internal policy.	*1*
SRC	External source pathname.	*/lustre/user/file.dat*
DST	Internal pathname in HPSS.	*/home/file.dat*

TABLE III: The data transfer record schema and example that we use for analysis.

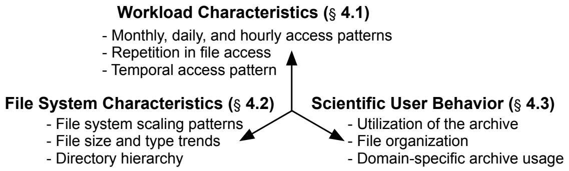

Workload Characteristics (§ 4.1)
- Monthly, daily, and hourly access patterns
- Repetition in file access
- Temporal access pattern
File System Characteristics (§ 4.2)
- File system scaling patterns
- File size and type trends
- Directory hierarchy
Scientific User Behavior (§ 4.3)
- Utilization of the archive
- File organization
- Domain-specific archive usage

Fig. 2: Overview of analysis dimensions in this paper.

III. Analysis Overview

In this section, we present the goals that guide our analysis of archival data transfer logs. By analyzing the archival data transfer logs, we specifically aim to gain insights on the following aspects.

Archival workload characteristics (§ IV-A): Compared to business and enterprise environments, we notice that relatively less attention has been given to the archival workload in scientific computing or HPC environments. Therefore, in this study, we aim to discover distinctive characteristics in the scientific archival workload by analyzing the data transfer log records from one of the world's largest HPC centers.

File system characteristics (§ IV-B): Archival data transfer logs describe data ingress or egress operations, but do not directly depict the file system status. However, they capture more than 99% of the activities within the archival file system, and therefore, allow us to confidently infer the incremental evolution of the file system status, e.g., number of files, file size, directory depth, etc., during our sample period. Such information, over a period of time, can help file system designers and developers pay attention to key metadata attributes.

Scientific user behavior (§ IV-C): It has been reported that the status of the PFS in the scientific computing centers is heavily influenced by the unique behavior of the scientific users [10]. We aim to complement such an observation by analyzing the scientific user behavior on HPSS and how it shapes the archive.

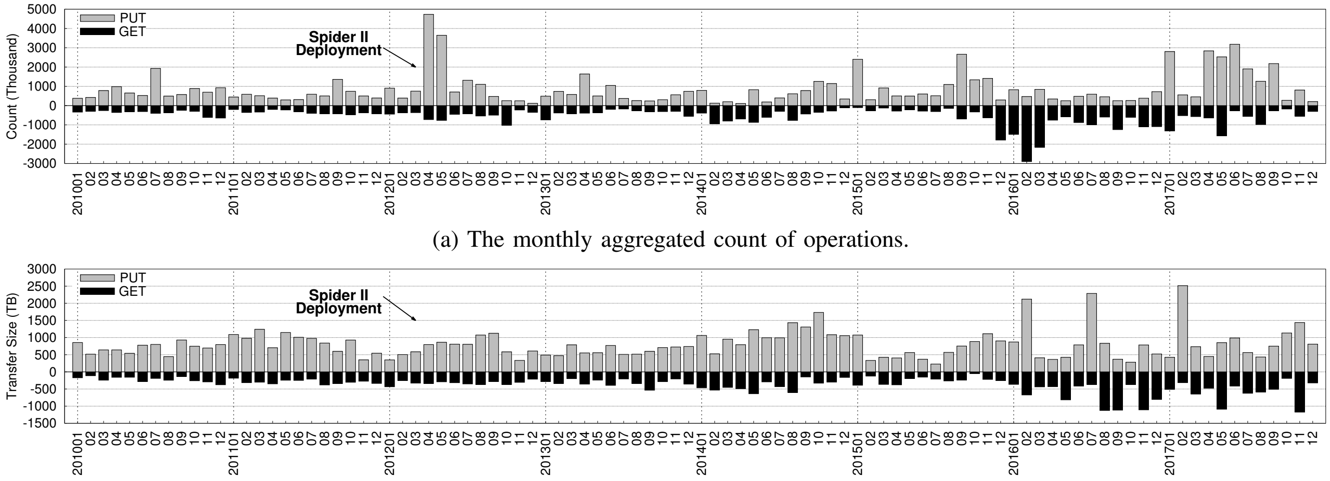
The storage system architecture in scientific computing environments is rapidly evolving by layering new storage tiers, such as burst buffers and object-based storage systems. An important objective of our analysis is to draw useful insights for provisioning future archival storage systems. Therefore, throughout the paper, we provide such insights as *observations* from our analysis results.

IV. Analysis Results

We now report our analysis results based on the goals (§ III), specifically, workload characteristics (§ IV-A), file system characteristics (§ IV-B), and scientific user behavior (§ IV-C).

A. Archival Workload Characteristics

1) Trends in Overall Access Patterns: Figure 3(a) and (b) depict the monthly-aggregated operation count and size, respectively, based on the operation type, i.e., PUT or GET,



(a) The monthly aggregated count of operations.

(b) The monthly aggregated size of operations.

Fig. 3: The monthly aggregated count and size of data transfer operations (PUT and GET) occurred in the OLCF archival storage system between 2010 and 2017. We observe a sudden spike in the operation count in 2010 April due to the deployment of the Spider II file system.

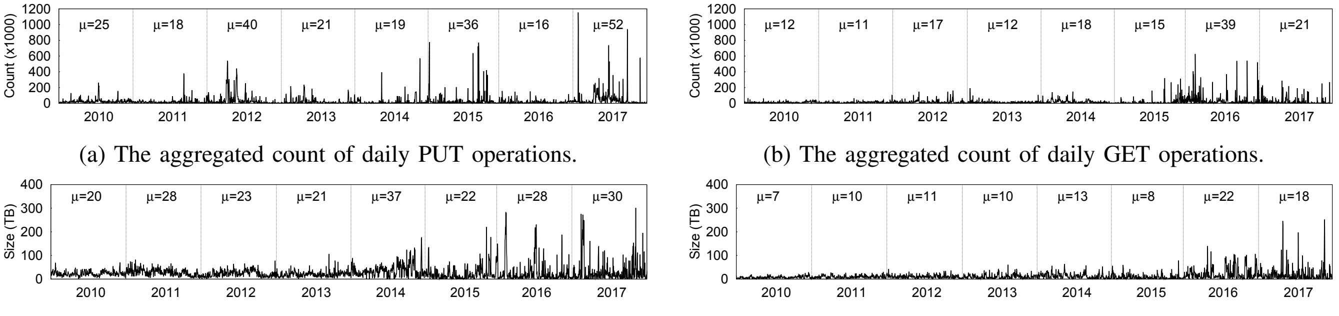
that occurred in OLCF HPSS between 2010 and 2017. In contrast to the PFS profiling result [10], where we could observe the positive increasing trend over time, the increasing trend in HPSS is relatively steady and less obvious. However, we notice sporadic spikes in the PUT count (Figure 3(a)) in 2012 and 2017. In particular, the monthly aggregated PUT counts between April (4.7 million) and May (3.6 million) in 2012 are noticeably higher (×5.5 and ×4.3, respectively) than the average of the monthly PUT count (826,446). This sudden increase was triggered by the deployment of a new supercomputer, Titan [15], and a PFS, Spider II [17]. Specifically, users transferred their files from the old PFS, Spider I [16], to HPSS for migrating to the new computing environment. However, we also note that the aggregated size of such files were not particularly huge enough to mark a similar spike in the operation size (Figure 3(b)). In contrast, the spikes of PUT counts in 2017 (Figure 3(a)) are not attributed to system changes, but triggered solely by users. Interestingly, such increases in the operation count do not always directly lead to a corresponding increase in the operation size. For instance, the maximum aggregate size of PUT is observed in February 2017, but the PUT count of the same month (519,676) is below the overall average (551,399). Not surprisingly, PUT operations surpass GET operations both in count and size, i.e., PUT exhibits 61% and 68% of overall operation count and size, respectively. While this PUT (or write)-dominant trend is similar to the previous observations from enterprise backup storage systems, we observe a significantly higher GET (or read) ratio in our system, i.e., 39% of all operations, almost ×4 more than enterprise systems [18]. Moreover, our GET ratio also far exceeds recent observations from other scientific institutions, e.g., 12% in National Center for Atmospheric Research (NCAR) [31].

Observation 1. *Besides sporadic workload spikes, the growth of system resources, i.e., supercomputers and PFSs, directly lead to an increasing use of the archival file system. Since archival storage investments are for a much longer duration than the typical lifecycle of a single supercomputer or a PFS, the archive needs to be provisioned (or designed to be easily upgradable) to accommodate such scenarios in order to guarantee long-term performance.*

Observation 2. *Despite sharing the same PUT (or write)-dominant trend, the OLCF HPSS exhibits about 30% higher GET (or read) requests, i.e., about 40%, compared to enterprise backup file systems, i.e., about 10%. This implies that the archive needs to optimize for both writes as well as reads.*

2) Daily Operations: Next, we investigate the daily data transfer trend. Figure 4 presents the daily aggregated operation counts ((a) PUT and (b) GET) and sizes ((c) PUT and (d) GET). Overall statistical summaries of both operations are also shown in Figure 4(e) and (f). The daily count and size of both operations remain steady except for sporadic spikes. The spikes occur more frequently in the last half of the sample period, i.e., between 2014 and 2017, particularly for PUT operation. For instance, during our eight-year sample period, there exist 15 days that have more than 500,000 PUT operations, and 13 of the 15 days occurred between 2014 and 2017. January 27th in 2017 recorded the highest 1.15 million PUT operations. However, we do not observe any significant increasing trend during our sample period. For instance, the Pearson correlation coefficient [3] values of PUT and GET counts are respectively ρ=0.06 and ρ=0.13, indicating that no consequential trend exists between the operation count and time. Similarly, PUT and GET sizes do not exhibit any correlations, i.e., ρ=0.06 and ρ=0.23, respectively. Interestingly, our observation from the archival file system is different from the PFS, where an apparent increasing trend was observed over time [10]. Further,

[3] The Pearson correlation coefficient, ρ, is defined as covariance of the variables (e.g., X and Y) divided by the product of their standard deviations, i.e., $\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$. A ρ value (ranging between -1 and 1) close to 0 indicates that no significant linear correlation is found.



(a) The aggregated count of daily PUT operations.

(b) The aggregated count of daily GET operations.

(c) The aggregated size of daily PUT operations.

(d) The aggregated size of daily GET operations.

PUT	Mean	50th	75th	95th	99th	Max.
Count (×1000)	28.4	10.2	23.8	105.1	309.8	1,154.7
Size (TB)	26.4	19.9	33.0	69.5	160.9	301.6

(e) Distribution of daily PUT operations.

GET	Mean	50th	75th	95th	99th	Max.
Count (×1000)	18.3	8.0	18.3	67.8	192.4	626.8
Size (TB)	12.6	8.3	15.1	39.0	80.8	251.9

(f) Distribution of daily GET operations.

Fig. 4: A summary of daily data transfer operations in the OLCF HPSS system. The μ values in (a)-(d) denote the annual average of the corresponding year. Although we do not observe any significant increasing trend in both operations, spikes in count and size occur more frequently between 2014 and 2017, i.e., the last half of the eight-year period.

from the summary tables (Figure 4(e) and (f)), we notice that the distributions of daily operations are heavily skewed. For instance, for eight years, the average count of the daily PUT operation is about 28,400, but the maximum daily PUT count was over a million, ×40 higher than the average. Daily GET operation exhibits a similar trend, i.e., the maximum count is about ×34 higher than the average.

Observation 3. *The daily operations in the OLCF HPSS exhibits a heavily skewed distribution both in count and size. The maximum daily requests of PUT and GET are about ×40 and ×34 higher, respectively, than the daily average requests. This also makes the case for some over-provisioning of the archive to cater to peaks (e.g., more data movers to satisfy the requests, more metadata servers, etc.) and not simply catering to perceived averages.*

3) Hourly Operations: Production systems oftentimes require downtimes for various maintenance purposes, such as software updates. It is important for system administrators to choose an appropriate downtime to minimize the impact on system users. To this end, we report the hourly access pattern of HPSS, as an operation heat map in Figure 5. As mentioned earlier (§ IV-A1), the relative higher activity in 2012 was caused by the system migration. Besides 2012, it is noticeable that the users have utilized HPSS more in recent years, i.e., between 2015 and 2017. Furthermore, between 2015 and 2017, it is perceptible that access to the HPSS becomes relatively less frequent in the early morning time, i.e., between 4am to 10am. This result suggests that the early morning time is most appropriate for administrators to perform maintenance tasks on HPSS. We have also performed a similar analysis to observe the system idleness based on the day of the week. Not surprisingly, HPSS activities were noticeably diminished on weekends.

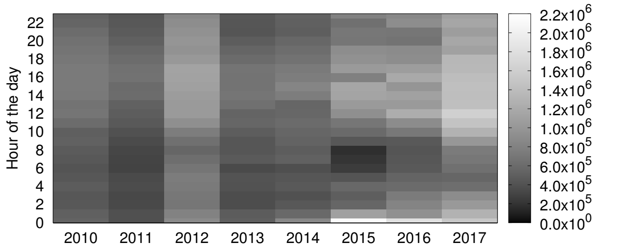


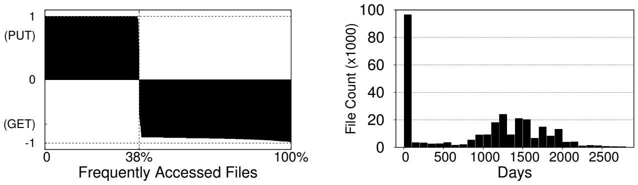
Fig. 5: Operation heat map with respect to the hours of day. The brighter cells indicates that relatively more user accesses took place in the corresponding timeslot.

Access	Mean	50th	75th	95th	99th	Max.
PUT	0.260	0	0	0	2	34,230
GET	0.775	0	1	4	8	16,057
ALL	**0.770**	**0**	**1**	**3**	**9**	**34,230**

TABLE IV: Per-file access frequency in the OLCF HPSS system between 2010 and 2017. Less than 1% of the 23.5 million files have been accessed more than 10 times, after the initial creation.

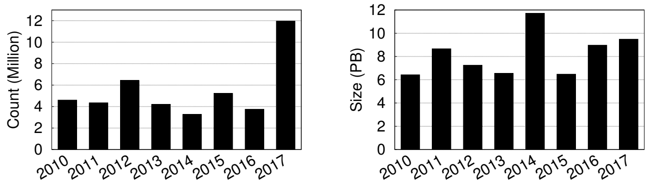
Observation 4. *The user access to HPSS is most idle in the early morning time, which can be an appropriate timeslot to perform maintenance tasks.*

4) Repetition in File Access: Next, we analyze how repeatedly files are accessed in HPSS. We perform the analysis with the destination pathnames in the transfer log. Specifically, we have created a separate database and populated it with transfer records that have a complete destination pathname. In addition, we exclude transfer records of accessing old files, i.e., files created before 2010, in this analysis. Consequently, our new database consists of 23.5 million distinct files and 48 million operations (about 30 million PUT operations) to them. Table IV summarizes how repeatedly the files are accessed after their creation. It is clearly noticeable that the access frequency is extremely skewed regardless of the operation type. Specifically, 67% of the files (about 16 million files) have never been accessed again after their initial creation, while less than 1% of the files (286,635 files) have been accessed more than ten times during the eight year period. Such a low re-referencing ratio is not surprising, considering that the primary



(a) Operation type. (b) File access duration.

Fig. 6: The temporal access pattern of the 286,635 frequently accessed files in the OLCF HPSS storage system. Users tend to access the frequently accessed files through a single operation type, i.e., either PUT or GET, within 100 days of the initial file creation.



(a) Number of new files. (b) Aggregated size of new files.

Fig. 7: Increase of the number and size of the files in OLCF HPSS. The number of files has been increasing consistently, i.e., ×1.36 each year on average. This result does not include directories.

role of the archival storage system is to recover data from a disastrous loss. However, around 33% of the files have been read at least once after creation, which indicates the need to optimize the read path in addition to the write path as mentioned earlier.

Observation 5. *Files that are repeatedly accessed are extremely rare, i.e., less than 1% of all files, in the OLCF HPSS. This provides a reasonable estimation to determine the disk cache size. It is particularly important to base the disk cache on such measured metrics instead of ad hoc provisioning, which is the current state of practice. Overall, the OLCF HPSS had a 33% recall rate.*

5) Operation Type for Frequently Accessed Files: To investigate the operation type for the 286,645 frequently accessed files (IV-A4), we now define the *operation purity* as a ratio of a dominant operation type, i.e., PUT or GET. We calculate the operation purity for each of the frequently accessed files, then negate the purity value if the dominant operation type is PUT. Figure 6(a) depicts the operation purity values of the frequently-accessed files, sorted in descending order. We observe that a single operation type dominates accesses to these files. For instance, the most frequently accessed file (34,230 times) has been accessed only through repetitive PUT operations. We suspect that certain users repeatedly archive their data files using the same name. From Figure 6(a), we also notice that the 72% of overall accesses to the frequently accessed files is through the GET operation, indicating that, in HPSS, there exist a small number of files that users keep accessing. Furthermore, this observation also provides a useful hint to design a file management policy. For instance, if a file is accessed via an operation type, e.g., GET or PUT, after its creation, the subsequent accesses to the file are likely to repeat the original operation type.

Observation 6. *When accessing frequently accessed files, users tend to repeatedly use a single operation type, i.e., either PUT or GET, suggesting that the system can predict the future operation type of a file based on its access history. This knowledge can help with pre-staging optimizations of the archive, i.e., the HPSS software can employ sophisticated prefetching techniques.*

6) Temporal Access Patterns of Frequently Accessed Files: We further explore the temporal access characteristics of the frequently accessed files (IV-A4). The histogram in Figure 6(b) shows the number of days these files are accessed after initial creation. The histogram bin size is 100 days, e.g., the right-most bar demonstrates that there exist 428 files whose last access were between 2,700 and 2,800 days, i.e., more than seven years, after their creation. We observe that the last accesses to 34% of these files (96,455 files) occurred within 100 days after creation (the left-most bar in the histogram). Also, 64% of these files (61,967 files) were only accessed for a month, i.e., 29 days, after creation. This observation can provide a credible guideline to design an internal data management policy in HPSS. For instance, it could be reasonable to keep newly ingested files in the fast tier, e.g., the disk tier in HPSS (§ II-B), for a month. However, an increase of the file access duration does not lead to a corresponding increase of the access frequency. Even after excluding the outliers, i.e., files that have been accessed more than a thousand times, we could not observe a strong correlation between the access duration and frequency (ρ=-0.01). In addition, we have analyzed the access interval, i.e., the amount of time between two consecutive accesses to a file. Specifically, we have calculated the access interval from each access request to the frequently accessed files (total 8.2 million access requests). We see that 84% of the file accesses to frequently accessed files occur within five weeks from the previous access request.

Observation 7. *Accesses to 64% of frequently accessed files occur within a month after the file creation. In addition, when a file is accessed again after its creation, the following access is likely to happen within five weeks, with over 80% of probability. Such temporal access characteristics can help to improve the internal migration policy in HPSS, e.g., a default retention period of a file in the disk cache, which can help right size the cache.*

B. File System Characteristics

As previously mentioned (§ II), our transfer log records capture more than 99% of data admissions to the OLCF HPSS storage system and also contain the pathnames. In this section, we analyze and report the file system characteristics based on the pathnames.

1) Number of Files: Figure 7(a) shows the number of newly created files, excluding directories, between 2010 and 2017. It clearly shows the increase of new files in 2012 and 2017, and this result is consistent with the result from the monthly aggregated PUT operation trend (§ IV-A1). In particular, the number of new files in 2017 (12 million) is ×2.6 higher than the average of prior years (4.5 million). Note that this increase in 2017 is purely triggered by system users (§ IV-A1). Considering that users rarely remove files from the archival storage [12], we can infer that the total number of files has been steadily increasing over time. This observation is aligned with our prior observation from the Spider II file system (the scratch file system in OLCF) [10]. We have also analyzed the number of parent directories under which new files were created. The ratio of such directories to files is only about 1% until 2016, and drops down further to 0.5% in 2017. For instance, over 13 million files were created under 60,386 directories, 222 files per directory on average, in 2017.

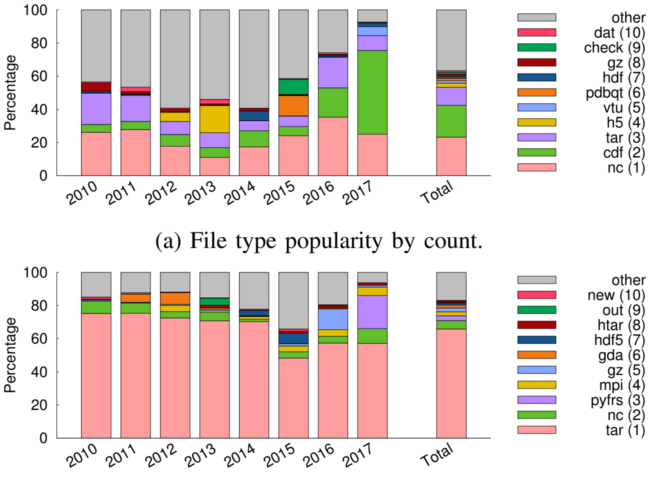
Observation 8. *The total number of files in HPSS has been growing consistently, i.e., on average ×1.36 each year. In particular, the number of new files has sharply increased recently since 2017. Knowing the growth rate is particularly important for future projections and provisioning of the storage.*

Observation 9. *As the number of files grows, users are likely to create more number of files under a single directory. Therefore, it is crucial for the HPSS software (metadata layers) to support huge directories as the file system scale grows.*

2) File Size: Next, we analyze the spatial increase of HPSS utilization. Figure 7(b) shows the storage space occupied by newly created files between 2010 and 2017, and it is difficult to find a clear increasing or decreasing trend. On average, about 8 PB storage space has been incrementally consumed by new files in each year between 2011 and 2017. However, we observe that a sudden increase occurs in 2014, when 46% higher space (11.7 PB) was consumed than the overall average. In Table V, we further report the size distribution of individual new files for each year. Notice that the average file size in 2014 is more than double of the overall average. In addition, we observe that most files are rather small, i.e., more than 75% of the files are smaller than 1 GB, although there exist a few large files, i.e., 0.02% of files that are larger than 1 TB. Lastly, we found 320,401 zero-sized files (about 0.7%) that were created between 2010 and 2017. We infer that most zero files have been generated by applications and then moved with other files by users, because we rarely observe descriptive file names from them.

	Mean(G)	50th(GB)	75th(GB)	95th(GB)	99th(GB)	Max.(T)
2010	1.46	0.03	0.20	1.83	18.51	11.57
2011	2.09	0.01	0.20	1.85	35.64	10.93
2012	1.18	0.00	0.03	1.13	14.20	15.09
2013	1.63	0.01	0.13	1.90	11.10	94.07
2014	3.77	0.00	0.06	1.63	53.03	60.12
2015	1.30	0.00	0.03	1.24	9.43	26.73
2016	2.51	0.04	0.15	1.97	22.78	44.30
2017	0.83	0.00	0.02	0.24	2.00	102.28
Total	**1.57**	**0.00**	**0.07**	**1.17**	**12.42**	**102.28**

TABLE V: The size distribution of new files in OLCF HPSS. The distribution is extremely skewed that 75% of the files are less than 1 GB, but there exists a few huge files that are larger than 100 TB.



(a) File type popularity by count.

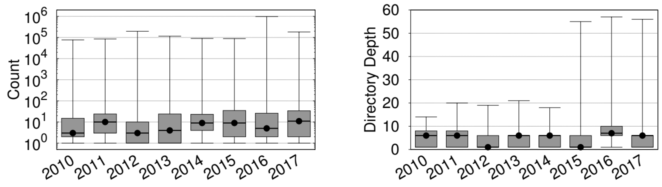
(b) File type popularity by size.

Fig. 8: File type popularity in OLCF HPSS. Archival file extensions (e.g., *tar*, *htar*, etc.) and scientific file extensions (e.g., *nc*, *hdf*, etc.) dominate the popularity.

Observation 10. *Similar to other file systems, most files in HPSS are rather small, i.e., 75% of files are smaller than 1 GB. However, there exists a few huge files, larger than 100 TB. That is, a small set of large files contribute to the overall capacity, while a large number of small files contribute to the total number of files.*

3) File Types: Figure 8(a) and (b) shows the top ten popular file types categorized by the file count and size, respectively. Not surprisingly, we see many scientific data formats, such as CDF (*cdf*), netCDF (*nc*), HDF5 (*h5 and hdf5*), and PyFR (*pyfrs*), and archival data formats, such as *tar*, *htar*, and *gz*, from both results. For the file count (Figure 8(a)), CDF and netCDF files are dominant, claiming 42.5% of all files (about 19 million files) between 2010 and 2017. Both formats are widely adopted for storing measurements data in Atmostpheric and Climate sciences [26], and, particularly, *cdf* files have suddenly increased in 2017. Furthermore, compared to the result from the PFS [10], we observe a higher and steadier ratio of well-known scientific files throughout our eight-year period. For instance, around 20% of files have steadily been *nc* files without a noticeable fluctuation. We expect that this is because users tend to migrate only resulting data files from the PFS to HPSS, and also the measurement data files that are directly migrated to HPSS, i.e., instead of being moved from the PFS. For the space consumption (Figure 8(b)), 66% of the space between 2010 and 2017 was consumed by *tar* files, because users oftentimes prefer to combine multiple files into a single file when moving files to HPSS. We have also analyzed how many well-known document files reside in HPSS using the file name extension. In particular, we accounted the number of files that can be opened by MS office products, i.e., 68 file extensions that MS office products support, and found that only 3% of files (1.4 million files) had such extensions.

Observation 11. *The OLCF HPSS stores a higher ratio of*



(a) Number of child entries. (b) Directory depth.

Fig. 9: The directory hierarchy in OLCF HPSS. Both results are acquired by analyzing the pathnames of newly created files in each year. Each whisker bar shows the maximum, 75% quantile, median, 25% quantile, and the minimum values, from top to bottom.

well-known scientific files, such as netCDF (.nc) and HDF (.h5) files, than their ratio that was observed in the PFS. Since many of these file formats are often self-describing, this suggests an opportunity to provide advanced services, e.g., metadata extraction, for well-known file formats within HPSS.

4) Directory Hierarchy: Now, we analyze the directory hierarchy from the pathnames of newly created files. First, Figure 9(a) shows a heavily skewed distribution of the number of child entries under a single directory, or directory size. While 75% of directories have less than 35 child entries, e.g., 35 and 34 entries respectively in 2015 and 2017, we can also find directories with more than 100,000 child entries. For instance, we found a directory with 976,726 child entries in 2016, which was populated with output files from a Plasmoid simulation. In total, we have identified ten directories that have more than 100,000 child entries, and five such directories were found in 2017, for storing atmospheric measurement files. Overall, despite a few huge directories, the directory size does not exhibit any significant increasing or decreasing trend during the eight-year period. For the directory depth, we observe a sudden increase of the maximum directory depth starting from 2015, as depicted in Figure 9(b). For instance, the average of maximum directory depth between 2015 and 2017 is more than ×2 larger than the average between 2010 and 2014. However, this skewness of directory depth is relatively less severe compared to the skewness previously observed in the PFS [10], i.e., the maximum directory depth in the Spider II file system was greater than 2000 while more than 95% directories had a depth smaller than 15.

Observation 12. *Although the directory hierarchy has not significantly changed between 2010 and 2017, extreme cases, both in the directory size and depth, are frequently observed, emphasizing the need for efficient metadata management within HPSS.*

5) Top-Level Directories: For OLCF HPSS, users are advised to organize their files under a few top-level directories, instead of populating the file system root ('/'). For instance, */home* and */proj* are directories, under which users are supposed to organize user-oriented and project-oriented files, respectively. In addition, HPSS has dedicated directories for archiving measurement data files from Atmospheric Radiation Measurement facilities [26], or ARM hereafter, and a few more directories for archiving system log files. Figure 10 shows the occupancy of each of these top-level directories. */other* occupying less than 1% of overall file count and capacity, denotes a few special directories for storing system log files, and ordinary users do not have access to them. Noticeably, ARM measurement files (*/ARM*) are dominant in the file count, populating 70% of all files. The space occupancy of */ARM*, however, is only 3%, while */home* (62%) and */proj* (35%) files occupy most of the file system space.

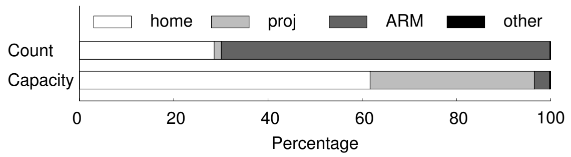


Fig. 10: The aggregated file count and size of top-level directories in OLCF HPSS between 2010 and 2017. The files under the */home* directory occupy 62% of the overall capacity utilization.

Observation 13. *The regular archive of scientific measurement files accounts for 70% of newly created files in HPSS. However, ordinary system users utilize 97% of the file system space with their home and project files. While such utilization is subjective to different deployments, it does suggest that differentiated and dedicated management policies can be associated to each top-level directory for efficient storage management.*

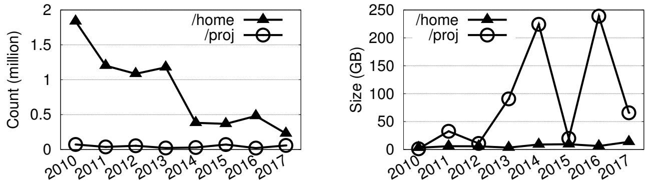
C. Scientific User Behavior

Here, we identify behavioral characteristics of OLCF scientific users in utilizing the archive, focusing on */home* and */proj* directories.

1) Utilization of the Archive: We first analyze how actively scientific users utilize the archival storage system. There were 1537 system users in OLCF between 2010 and 2017. Table VI summarizes the number of operations that the users performed between 2010 and 2017. First, it is noticeable that the number of active users who accessed HPSS at least once in the corresponding year has decreased during our sample period. For instance, only 234 users accessed HPSS in 2017, only about 35% of the number of active users in 2010. On

	Active	**Per-User**	**Operation**	**Summary**		**Total**
Year	**Users**	**Q_1**	**Q_2**	**Q_3**	**Maximum**	**Operations**
2010	651	11	96	2,431	1,744,285	13,742,816
2011	672	8	61	1,125	1,825,637	10,789,079
2012	672	9	63	1,128	3,289,832	20,862,767
2013	640	7	55	1,077	3,053,717	12,033,986
2014	515	9	88	1,733	4,506,538	13,769,468
2015	312	6	53	2,083	4,415,338	18,370,385
2016	298	10	97	2,755	6,838,528	20,341,751
2017	234	7	77	2,289	13,247,520	27,313,737

TABLE VI: Per-user operation counts. Q_1, Q_2, and Q_3 show the 25%, 50%, and 75% quantile values, respectively. Despite of the decrease in the number of active users, the utilization of individual active users is becoming higher.



(a) Aggregated new file count.

(b) Average new file size.

Fig. 11: The utilization of */home* and */proj* directories in HPSS. For each year, (a) depicts the aggregated new file count, and (b) depicts the aggregated new file size.

the contrary, for eight years, the total operation count has almost doubled. This suggests that, although the number of active users is decreasing, the utilization of individual active users is becoming higher. In fact, the active users in 2017 triggered 116,726 operations on average, ×5.5 higher than the average operation count in 2010, i.e., 21,111. Next, Table VI also demonstrates that the distribution of utilization degree is heavily skewed among the active users. For instance, from the quantile values (Q_1, Q_2, and Q_3), we observe that 50% of active users accessed HPSS less than 100 times a year. However, there also exist users who have accessed HPSS more than a million times a year. In 2017, for example, a single user triggered more than 13 million operations, generating almost half of the total workload in the year. This clearly demonstrates that the distribution of HPSS utilization among the scientific users is heavily skewed.

Observation 14. *Despite a decrease of active users, the total workload of HPSS has doubled between 2010 and 2017, due to a sharp increase in the individual users' workload. This implies that more intuitive end-user tools can potentially improve the utilization of the archive from scientific users.*

Observation 15. *In OLCF HPSS, while 50% of the users perform less than 100 operations in a year, a few active users perform more than a million operations in a year. This indicates that the current static, per-user quota policy may deter higher utilization of HPSS. More dynamic and workload-aware quota management is needed.*

2) File Organization: As mentioned above (§ IV-B5), OLCF users organize files in HPSS under */home* and *proj* directories. We now analyze how the users populate these top-level directories. Figure 11 depicts (a) the aggregated file count and (b) the average file size inside */home* and */proj* directories between 2010 and 2017. Note that these results do not include files that were created before 2010. From Figure 11(a), we observe that the number of files under */home* exhibits a diminishing trend, i.e., from about 1.8 million in 2010 to 229,461 in 2017. In contrast, for eight years, the number of */proj* files steadily remains under 100,000, indicating that */home* files clearly dominates */proj* files in the file count. However, as shown in Figure 11(b), the average size of */proj* files is noticeably larger than the average file size in */home*. For instance, the overall average of */proj* file size is 56 GB, more than ×10 larger than average size of files under */home* (5 GB), as presented in Table VII. Table VII also suggests that the ratio of the large files is greater in */proj* than the ratio in */home*, i.e., the third quantile value (Q_3) of */proj* is more than ×30 larger. Specifically, the ratio of large or huge files, i.e., files greater than 8 GB (Table II), in */proj* is 28%, about ×5 greater than the ratio in */home* (5%). This is because most data products, which can potentially grow large, are produced on a project basis and thus stored within */proj*, while the */home* space is used for storing program code and configuration files, which are relatively smaller than the data products. In addition, from Figure 11(b), we observe that average size of */proj* files heavily fluctuates, while it remains rather steady for */home* files.

	Mean(GB)	Min.(GB)	Q^1(GB)	Q^2(GB)	Q^3(GB)	Max.(TB)
/home	5.36	0.00	0.01	0.06	0.31	102.28
/proj	56.37	0.00	0.07	0.42	9.80	94.07

TABLE VII: The aggregated summary of file size inside */home* and */proj* directories in OLCF HPSS between 2010 and 2017. */proj* files tend to be larger than */home* files.

Observation 16. *Overall, /home files surpass /proj files in count and size. However, on average, the files in /proj are significantly larger, i.e., more than ×10, than the files in /home. Therefore, different data management policies can potentially be implemented and associated to distinct file characteristics of these directories (similar to the observation in § IV-B5).*

3) Domain-Specific Archive Usage: We characterize the user behavior associated to their science domains. Table VIII summarizes the major characteristics of files under */proj* and */home*, categorized by science domains. The first (***/proj* files**) and the second (***/home* files**) column groups respectively present the characteristics of */proj* and */home* files of the corresponding science domain. The last column group (**Comparison**) shows *dominance ratio*, r, which we define as $r = sgn(h-p)\frac{max(h,p)}{h+p}$, where sgn is a sign function, and h and p respectively refers to */home* and */proj* values. For instance, $r_{filecount}$ of *Accelerator Physics* is 1.00, indicating that almost 100% of *Accelerator Physics* files have been created in */home* instead of */proj*. However, $r_{filesize}$ ($r_{filesize}$=-0.58) shows that 58% of the space consumption from *Accelerator Physics* is accounted from */proj* files. Therefore, we can infer that the *Accelerator Physics* scientists seldom create files under */proj*, but such */proj* files tend to be huge. Overall, $r_{filecount}$ values suggest that many science domains, i.e., 14 positive $r_{filecount}$ values out of 21, prefer to utilize the */home* directory. Furthermore, for 12 science domains, both $|r_{filecount}|$ and $|r_{filesize}|$, i.e., the absolute dominance ratio values, are greater than 0.9, meaning that either */proj* or */home* directory has been almost exclusively utilized. For instance, *Aerodynamics* and *Atmospheric Science* have exclusively utilized */home* and */proj*, respectively. Particularly for the file count, except for five science domains, i.e., *Climate Science*, *Combustion*, *Fusion Energy*, *Materials Science*, and *Physics*, all science domains exclusively created 90% of their files either in */proj* or */home*.

For the aggregated file count under */proj*, *Climate Science* records the highest, i.e., 52% of all */proj* files. Interestingly,

	***/proj* files**				***/home* files**					**Comparison**	
	Aggregated File		**Popular**	**Max. Dir.**	**Active**	**Aggregated File**		**Popular**	**Max. Dir.**	**Dominance Ratio** $(-1 \leq r \leq 1)$	
Science Domain	**Count**	**Size (TB)**	**File Type**	**Depth**	**Users**	**Count**	**Size (TB)**	**File Type**	**Depth**	$r_{filecount}$	$r_{filesize}$
Accelerator Physics	1,642	**1,763.93**	tar, h5	7	26	**1,646,896**	**1,279.86**	tar,hdf	**15**	1.00	-0.58
Aerodynamics	9	0.29	pyfrs, pyfrm	7	**64**	**517,123**	**1,206.19**	nc,f	**18**	1.00	1.00
Astrophysics	91,302	**7,287.53**	tar, h5	8	1	45	0.13	h5	3	-1.00	-1.00
Atmospheric Science	1,915	157.33	tar, cpio	8	**54**	**5,260,650**	**1,832.40**	nc,tar	**18**	1.00	0.92
Bioinformatics	4	0.90	tar	6	10	7336	226.62	bp,tar	**13**	1.00	1.00
Biology	1,014	5.65	dcd, old	6	4	31,667	171.67	tar,lime	7	0.97	0.97
Biophysics	105	153.23	trr, tar	7	40	**190,253**	523.04	xyz,out	**12**	1.00	0.77
Physical Chemistry	82	0.51	gz	5	13	80,113	394.82	tar,lime	7	1.00	1.00
Climate Science	**189,952**	**1,781.07**	nc, tar	**13**	11	**156,359**	344.54	nc,grb2	8	-0.55	-0.84
Combustion	956	75.97	tar, mpi	**16**	7	3,016	133.54	tar,trr	8	0.76	0.64
Computer Science	1,155	**5,424.09**	tar, mpi	8	1	8	46.87	tar,bin	6	-0.99	-0.99
Life Science	301	1.68	tar, dat	4	15	86,609	**1,481.35**	mpi,tar	**10**	1.00	1.00
Fusion Energy	39,486	**1,212.54**	gda, tar	7	9	34,712	**1,392.23**	png,out	**18**	-0.53	0.53
Geosciences	9,255	72.41	tar, htar	5	12	**305,249**	336.67	f,o	**10**	0.97	0.82
High Energy Physics	86	93.26	tar, hdf5	5	0	0	0.00	N/A	0	-1.00	-1.00
Lattice Gauge Theory	9,219	**1,133.09**	lime, gz	**10**	0	0	0.00	N/A	0	-1.00	-1.00
Materials Science	9,709	564.36	tar, dat	8	5	3,021	92.22	gz,trr	8	-0.76	-0.86
Nuclear Physics	6,827	238.01	lime, gz	7	49	**232,808**	398.17	enc,tgz	**13**	0.97	0.63
Physics	22	11.14	tar	6	1	67	3.50	tar,gz	5	0.75	-0.76
Staff	199	0.84	tar, mysqldump	5	12	**181,692**	467.91	tar,lime	**12**	1.00	1.00
Turbulence	20	0.08	gz, tar	5	8	10,467	650.13	tar,out	**10**	1.00	1.00

TABLE VIII: The comparison of file characteristics between */proj* and */home* files, categorized by 21 science domains. The Comparison column group shows *dominance ratio*, or r, values of aggregated file count and size. The positive r indicates the dominance of */home* files (blue cells), while negative r indicates dominance of */proj* (red cells). The color intensity represents the absolute value of r, i.e., the degree of dominance.

this result is quite different from our previous profiling result from the parallel file system [10]. For instance, none of the top five science domains in the */proj* file count, i.e., *Astrophysics*, *Computer Science*, *Climate Science*, *Accelerator Physics*, and *Fusion Energy*, appeared in the top five list during the PFS profiling[4]. Similarly, except for *Biophysics*, none of top five science domains in the */home* file count, i.e., *Atmospheric Science*, *Accelerator Physics*, *Aerodynamics*, *Geosciences*, appeared in the top five list from the PFS profiling[5]. This suggests that a heavy PFS utilization does not always lead to a heavy archival file system utilization. This is because some science projects, e.g., climate science projects, do not often require to perform scientific simulations but rely more on the long-term measurement data archive, resulting in a heavier utilization of the archival storage system.

The maximum directory depth of a science domain is more frequently found under */home*, but the difference between */proj* and */home* is miniscule, i.e., the maximum directory depths in */proj* and */home* are 16 (*Combustion*) and 18 (*Aerodynamics*, *Atmospheric Science*, and *Fusion Energy*), respectively. In addition, the maximum directory depths in */proj* are rather moderate, compared to the result from the PFS, i.e., the median value (Q_2) of maximum directory depths of all science domains is 7 in Table VIII, about one third of the median value observed in the PFS (Q_2=22) [10].

For the file types, archival file extensions, such as *tar* and *htar*, are popular across all science domains both in */proj* and */home*. In addition, each science domain utilizes its own appropriate file type, e.g., netCDF (*nc*) files in *Climate Science*.

Observation 17. *Domain scientists tend to exclusively utilize one of /proj and /home directories. This usage pattern suggests that a combined quota policy could be more effective than the current quota policy, which is applied independently to /proj and /home.*

Observation 18. *A higher PFS utilization of a science domain does not lead to a higher archival file system utilization of the science domain. This could be because users wish to retain/curate only the final products of scientific simulations. Consequently, an HPC center might want to encourage such behavior by offering users tools to better curate data products within HPSS instead of indiscriminately moving files to the archive, e.g., with tools to obtain digital object identifiers (DOIs) for curated data.*

V. Discussion

In this section, we summarize our important findings that will guide development and operation of the future archival storage systems in scientific computing environments.

First, observations from our workload analysis (§ IV-A) indicate that the system provisioning should consider the maximum workload size instead of relying on the long-term averages. Specifically, abrupt workload spikes in a day, which cannot be precisely reflected by monthly aggregated statistics, have occurred more frequently in recent years. Furthermore, read requests accounted for 39% of the total requests, suggesting that the archival storage system needs to treat the read requests equally important to the write requests. In addition, our analysis on the temporal access patterns provides a reasonable guideline to establish the adequate size and

[4] The top five science domains that created the largest number of files in the PFS was *Staff*, *Biophysics*, *Computer Science*, *Physical Chemistry*, and *Turbulence* [10].

[5] As shown in § II, the parallel file system only provides */proj* areas, while NFS provides separate user */home* areas.

eviction policies of the internal disk cache tier. Traditionally, such design decisions have followed rules of thumb, e.g., a certain ratio to the overall capacity.

Second, our file system analysis (§ IV-B) reveals the importance of metadata management in the archival storage system, particularly at large-scale. Not only is the overall file count continuously increasing but individual directories are also becoming larger, e.g., million files under a single directory. Although we cannot precisely quantity the amount of data from data hoarding, i.e., indiscriminate data migration for merely avoiding the system-wide purge [32], our observation of the 33% recall rate (§ IV-A4) suggests that there exist valuable scientific data in the archive. Therefore, it will be crucial to provide additional services that can facilitate users with their data management tasks [33], [34], [35], [36]. Our file type analysis further suggests that integrating the advanced data management services within the archival storage system, e.g., metadata indexing, automatic metadata annotation and extraction, etc., looks feasible.

Lastly, distinct scientific user behavior (§ IV-C) unveils a potential limitation of enforcing monolithic storage policies, e.g., a static quota allowance, to all system users. For instance, most scientific users exhibit a strong tendency to exclusively utilize a single storage space, i.e., either */home* or */proj*. Likewise, each science domain demonstrates distinct usage characteristics, suggesting the need for more sophisticated storage policies, such as dynamic quota implementation and differentiated storage area management. Particularly, a recently reported migration of a 2.9 PB dataset from a single science project [37] in the OLCF HPSS further signifies that traditional homogeneous storage policies may not be sustainable.

VI. Related Work

Workload characteristics of large-scale networked file systems have been extensively explored in diverse system environments. Earlier studies analyzed the file system traces from academic file servers [38], [39] and I/O traffic from the CIFS file system in enterprise environments [40]. In addition, enterprise file systems were also explored particularly for studying the efficacy of deduplication in workstation file systems [41] and backup file systems [18]. Recent studies also encompass the workloads from cloud storage environments, e.g., analyzing the similarity in the virtual machine images [42] or usage patterns in the personal cloud backend [43], [44]. However, the insights from such studies cannot directly benefit scientific computing centers due to the fundamental dissimilarities in the system purpose and architecture.

In HPC environments, the performance of the PFS has been considered to be lagging the computing performance [45], and thus I/O traces and snapshots of PFS have been extensively explored [46], [47], [48], [49], [50], [51]. Recent studies have also explored the scientific user behavior from PFS snapshots [10] and the deduplication efficacy in the scientific data centers [52]. Our analysis of archival file system workload is complementary to such studies and potentially provides a deeper insight into understanding the complete storage stack in large-scale HPC environments.

Compared to the PFS, relatively less attention has been paid to the archival file systems in scientific computing centers. A few examples include a study of characterizing various file systems in different HPC centers [11], which studied archival file systems from Pacific Northwest National Laboratory and Arctic Region Supercomputing Center. Another study found that the access pattern of archival file systems had become more write-intensive and less frequent, over two decades between early 1990s and 2010s [53], [13], [54]. However, the study analyzed the file system snapshot data, which could not precisely expose the temporal workload characteristics. In addition, the target file systems in such studies were substantially smaller, e.g., 1.3 PB at most, than our archival file system, i.e., 80 PB of used capacity. More importantly, we have analyzed data transfer logs of eight consecutive years in one of the world's largest scientific computing centers.

Lastly, a recent study of the storage system in European Centre for Medium-Range Weather Forecasts (ECMWF) analyzed the three-year workloads from the 14.8 PB HPSS file systems [12]. Despite the thorough analysis, the storage architecture in the study is deeply customized for the specific purpose of the institution, e.g., object storage database. Another study of the 30 PB HPSS file system in National Center for Atmospheric Research (NCAR) revealed distinctive user behavior, e.g., a substantial ratio of delete operations (15%) [31], [27]. Similar to ECMWF, NCAR is specialized to a single science domain, i.e., atmospheric research, and thus the observations do not comprehensively reflect general scientific computing centers. In contrast, we analyze the data transfer activities for a longer period, i.e., eight years, from a larger archival file system, in a scientific computing center that facilitates a more diverse range of scientific disciplines [55].

VII. Conclusion

In this paper, we have analyzed eight years worth of data transfer activities in the OLCF HPSS, one of the world's largest HPSS deployments. Specifically, we have analyzed the workload characteristics, file system characteristics, and scientific user behavior. Our analysis indicates that the archival storage system in OLCF exhibits unique characteristics including the substantial read request ratio and science domain-specific user access patterns.

We believe our study will offer useful guidelines for operating and designing archival storage systems in large-scale scientific computing environments.

Acknowledgment

We would like to thank Mitchell R. Griffith and Vicky White in the OLCF HPSS team for their assistance. This work was supported by, and used the resources of, the Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at ORNL, which is managed by UT Battelle, LLC for the U.S. DOE (under the contract No. DE-AC05-00OR22725).

References

[1] S. S. Vazhkudai, B. R. de Supinski, A. S. Bland, A. Geist, J. Sexton, J. Kahle, C. J. Zimmer, S. Atchley, S. Oral, D. E. Maxwell, V. G. V. Larrea, A. Bertsch, R. Goldstone, W. Joubert, C. Chambreau, D. Appelhans, R. Blackmore, B. Casses, G. Chochia, G. Davison, M. A. Ezell, T. Gooding, E. Gonsiorowski, L. Grinberg, B. Hanson, B. Hartner, I. Karlin, M. L. Leininger, D. Leverman, C. Marroquin, A. Moody, M. Ohmacht, R. Pankajakshan, F. Pizzano, J. H. Rogers, B. Rosenburg, D. Schmidt, M. Shankar, F. Wang, P. Watson, B. Walkup, L. D. Weems, and J. Yin, "The Design, Deployment, and Evaluation of the CORAL Pre-exascale Systems," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, ser. SC '18, 2018.

[2] *TOP500 Lists*, http://www.top500.org/lists/.

[3] S. R. Alam, J. A. Kuehn, R. F. Barrett, J. M. Larkin, M. R. Fahey, R. Sankaran, and P. H. Worley, "Cray XT4: An Early Evaluation for Petascale Scientific Simulation," in *SC '07: Proceedings of the 2007 ACM/IEEE Conference on Supercomputing*, 2007.

[4] A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, and N. Andrew, "Deep Learning with COTS HPC Systems," in *International Conference on Machine Learning*, 2013.

[5] N. Liu, J. Cope, P. Carns, C. Carothers, R. Ross, G. Grider, A. Crume, and C. Maltzahn, "On the role of burst buffers in leadership-class storage systems," in *IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*, 2012.

[6] H. Tang, S. Byna, F. Tessier, T. Wang, B. Dong, J. Mu, Q. Koziol, J. Soumagne, V. Vishwanath, J. Liu *et al.*, "Toward Scalable and Asynchronous Object-Centric Data Management for HPC," in *Proceedings of the 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 2018.

[7] *Oak Ridge Leadership Computing Facility - The OLCF was established at Oak Ridge National Laboratory in 2004 with the mission of standing up a supercomputer 100 times more powerful than the leading systems of the day.*, https://www.olcf.ornl.gov.

[8] *HPSS Collaboration*, http://www.hpss-collaboration.org.

[9] Y. Liu, R. Gunasekaran, X. Ma, and S. S. Vazhkudai, "Server-side Log Data Analytics for I/O Workload Characterization and Coordination on Large Shared Storage Systems," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '16, 2016.

[10] S.-H. Lim, H. Sim, R. Gunasekaran, and S. S. Vazhkudai, "Scientific User Behavior and Data-sharing Trends in a Petascale File System," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '17, 2017.

[11] S. Dayal, "Characterizing HEC Storage Systems at Rest," *Parallel Data Lab, CMU*, 2008.

[12] M. Grawinkel, L. Nagel, M. Mäsker, F. Padua, A. Brinkmann, and L. Sorth, "Analysis of the ECMWF Storage Landscape," in *13th USENIX Conference on File and Storage Technologies (FAST 15)*, 2015.

[13] I. F. Adams, M. W. Storer, and E. L. Miller, "Analysis of Workload Behavior in Scientific and Historical Long-Term Data Repositories," *ACM Transactions on Storage (TOS)*, vol. 8, no. 2, 2012.

[14] *Jaguar - Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090 | TOP500 Supercomputer Sites*, http://www.top500.org/system/176544.

[15] *Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x | TOP500 Supercomputer Sites*, http://www.top500.org/system/177975.

[16] Shipman, Galen and Dillow, David and Oral, Sarp and Wang, Feiyi and Fuller, Douglas and Hill, Jason and Zhang, Zhe, "Lessons Learned in Deploying the World's Largest Scale Lustre File System," in *The 52nd Cray user group conference*, ser. CUG '10, 2010.

[17] S. Oral, J. Simmons, J. Hill, D. Leverman, F. Wang, M. Ezell, R. Miller, D. Fuller, R. Gunasekaran, Y. Kim, S. Gupta, D. Tiwari, S. S. Vazhkudai, J. H. Rogers, D. Dillow, G. M. Shipman, and A. S. Bland, "Best Practices and Lessons Learned from Deploying and Operating Large-scale Data-centric Parallel File Systems," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '14, 2014.

[18] G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, "Characteristics of Backup Workloads in Production Systems," in *Proceedings of the 10th USENIX Conference on File and Storage Technologies*, ser. FAST'12, 2012.

[19] *Summit – Oak Ridge Leadership Computing Facility*, https://www.olcf.ornl.gov/summit/.

[20] *File Systems: Data Storage & Transfer - Oak Ridge Leadership Computing Facility*, https://www.olcf.ornl.gov/for-users/system-user-guides/summit/summit-user-guide/.

[21] *OLCF Policy Guide – Oak Ridge Leadership Computing Facility*, https://www.olcf.ornl.gov/for-users/olcf-policy-guide/.

[22] R. M. Whitten, T. E. Barron, D. L. Hulse, P. Pfeiffer IV, and S. L. Scott, "Resource Allocation and Tracking System (RATS) Deployment on the Cray X1E and Cray XT3 Platforms," in *Cray User Group*, ser. CUG'06, 2006.

[23] *IBM Db2 – Data management software – IBM Analytics*, https://www.ibm.com/analytics/us/en/db2/.

[24] *File Systems: Data Storage - Transfers - Oak Ridge Leadership Computing Facility*, https://www.olcf.ornl.gov/for-users/system-user-guides/summit/file-systems/.

[25] M. Mesnier, G. R. Ganger, and E. Riedel, "Object-based storage," *IEEE Communications Magazine*, vol. 41, no. 8, pp. 84–90, Aug 2003.

[26] *ARM Climate Research Facility*, http://www.arm.gov/.

[27] I. F. Adams, B. A. Madden, J. C. Frank, M. W. Storer, E. L. Miller, and G. Harano, "Usage Behavior of a Large-scale Scientific Archive," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '12, 2012.

[28] *National Energy Research Scientific Computing Center*, https://www.nersc.gov.

[29] *Research data management simplified. | globus*, https://www.globus.org.

[30] *GridFTP - Globus Toolkit - Globus.org*, http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/.

[31] J. C. Frank, E. L. Miller, I. F. Adams, and D. C. Rosenthal, "Evolutionary Trends in a Supercomputing Tertiary Storage Environment," in *2012 IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, 2012.

[32] A. Holloway, "The Purge Threat: Scientists' Thoughts on Peta-scale Usability," in *Proceedings of the Sixth Workshop on Parallel Data Storage*, ser. PDSW '11, 2011.

[33] C. Johnson, K. Keeton, C. B. Morrey III, C. A. Soules, A. Veitch, S. Bacon, O. Batuner, M. Condotta, H. Coutinho, P. J. Doyle *et al.*, "From Research to Practice: Experiences Engineering a Production Metadata Database for a Scale out File System," in *Proceedings of the 12th USENIX Conference on File and Storage Technologies*, ser. FAST '14, 2014.

[34] S. S. Vazhkudai, J. Harney, R. Gunasekaran, D. Stansberry, S. Lim, T. Barron, A. Nash, and A. Ramanathan, "Constellation: A science graph network for scalable data and knowledge discovery in extreme-scale scientific collaborations," in *2016 IEEE International Conference on Big Data (Big Data)*, ser. BigData '16, 2016.

[35] H. Sim, Y. Kim, S. S. Vazhkudai, G. R. Vallée, S.-H. Lim, and A. R. Butt, "Tagit: An Integrated Indexing and Search Service for File Systems," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '17, 2017.

[36] M. Lawson, C. Ulmer, S. Mukherjee, G. Templet, J. Lofstead, S. Levy, P. Widener, and T. Kordenbrock, "Empress: Extensible Metadata Provider for Extreme-scale Scientific Simulations," in *Proceedings of the 2Nd Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems*, ser. PDSW-DISCS '17, 2017.

[37] *Argonne Team Breaks Record for Globus Data Movement*, https://www.prnewswire.com/news-releases/argonne-team-breaks-record-for-globus-data-movement-300880969.html.

[38] T. J. Gibson and E. L. Miller, "Long-Term File Activity Patterns in a UNIX Workstation Environment," in *Proceedings of the 15th IEEE Symposium on Mass Storage Systems and Technologies (MSST)*, 1998.

[39] D. Ellard, J. Ledlie, P. Malkani, and M. Seltzer, "Passive NFS Tracing of Email and Research Workloads," in *Proceedings of the 2Nd USENIX Conference on File and Storage Technologies*, ser. FAST '03, 2003.

[40] A. W. Leung, S. Pasupathy, G. R. Goodson, and E. L. Miller, "Measurement and Analysis of Large-Scale Network File System Workloads," in *USENIX annual technical conference*, ser. ATC '08, 2008.

[41] D. T. Meyer and W. J. Bolosky, "A Study of Practical Deduplication," *ACM Transactions on Storage (TOS)*, vol. 7, no. 4, 2012.

[42] K. R. Jayaram, C. Peng, Z. Zhang, M. Kim, H. Chen, and H. Lei, "An Empirical Analysis of Similarity in Virtual Machine Images," in

Proceedings of the Middleware 2011 Industry Track Workshop, ser. Middleware '11, 2011.

[43] I. Drago, M. Mellia, M. M Munafo, A. Sperotto, R. Sadre, and A. Pras, "Inside DropBox: Understanding Personal Cloud Storage Services," in *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*. ACM, 2012, pp. 481–494.

[44] R. Gracia-Tinedo, Y. Tian, J. Sampé, H. Harkous, J. Lenton, P. García-López, M. Sánchez-Artigas, and M. Vukolic, "Dissecting UbuntuOne: Autopsy of a Global-Scale Personal Cloud Back-End," in *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, ser. IMC '15. ACM, 2015.

[45] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller *et al.*, "Exascale Computing Study: Technology Challenges in Achieving Exascale Systems," *Defense Advanced Research Projects Agency Information Processing Techniques Office (DARPA IPTO), Tech. Rep*, vol. 15, 2008.

[46] K. K. Ramakrishnan, P. Biswas, and R. Karedla, "Analysis of File I/O Traces in Commercial Computing Environments," in *Proceedings of the 1992 ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '92/PERFORMANCE '92, 1992.

[47] Y. Kim, R. Gunasekaran, G. M. Shipman, D. A. Dillow, Z. Zhang, and B. W. Settlemyer, "Workload Characterization of a Leadership Class Storage Cluster," in *Proceedings of the 5th Petascale Data Storage Workshop*, ser. PDSW '10, 2010.

[48] P. Carns, R. Latham, R. Ross, K. Iskra, S. Lang, and K. Riley, "24/7 Characterization of Petascale I/O Workloads," in *Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on*, ser. CLUSTER '09, 2009.

[49] R. Miller, J. Hill, D. A. Dillow, R. Gunasekaran, G. M. Shipman, and D. Maxwell, "Monitoring Tools for Large Scale Systems," in *Proceedings of Cray User Group Conference (CUG 2010)*, ser. CUG '10, 2010.

[50] H. Luu, M. Winslett, W. Gropp, R. Ross, P. Carns, K. Harms, M. Prabhat, S. Byna, and Y. Yao, "A Multiplatform Study of I/O Behavior on Petascale Supercomputers," in *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing*, ser. HPDC '15, 2015.

[51] H. Shan, K. Antypas, and J. Shalf, "Characterizing and Predicting the I/O Performance of HPC Applications Using a Parameterized Synthetic Benchmark," in *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, ser. SC '08, 2008.

[52] D. Meister, J. Kaiser, A. Brinkmann, T. Cortes, M. Kuhn, and J. Kunkel, "A Study on Data Deduplication in HPC Storage Systems," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '12, 2012.

[53] E. L. Miller and R. H. Katz, "An Analysis of File Migration in a UNIX Supercomputing Environment," in *USENIX Winter*, 1992.

[54] B. Madden, I. F. Adams, J. Frank, and E. L. Miller, "Analyzing User Behavior: A Trace Analysis of the NCAR Archival Storage System," *University of California, Santa Cruz, Tech. Rep. UCSC-SSRC-ssrctr-12-02*, 2012.

[55] *Leadership Science - Oak Ridge Leadership Computing Facility*, https://www.olcf.ornl.gov/leadership-science/.