Access Patterns to Disk Cache for Large Scientific Archive

Yumeng Wang
University of California, Berkeley
Berkeley, CA, USA
wangyumeng2017@berkeley.edu

Kesheng Wu
Alex Sim
Lawrence Berkeley Nat'l Laboratory
Berkeley, CA, USA
{kwu,asim}@lbl.gov

Shinjae Yoo
Shigeki Misawa
Brookhaven Nat'l Laboratory
Brookhaven, NY, USA
{sjyoo,misawa}@bnl.gov

ABSTRACT

Large scientific projects are increasing relying on analyses of data for their new discoveries; and a number of different data management systems have been developed to serve this scientific projects. In the work-in-progress paper, we describe an effort on understanding the data access patterns of one of these data management systems, dCache. This particular deployment of dCache acts as a disk cache in front of a large tape storage system primarily containing high-energy physics data. Based on the 15-month dCache logs, the cache is only accessing the tape system once for over 50 file requests, which indicates that it is effective as a disk cache. The on-disk files are repeated used, more than three times a day. We have also identified a number of unusual access patterns that are worth further investigation.

CCS CONCEPTS

• **Information systems → Information storage technologies**; • **Computing methodologies → Model development and analysis**.

KEYWORDS

disk cache, hpss, tape, dcache

ACM Reference Format:
Yumeng Wang, Kesheng Wu, Alex Sim, Shinjae Yoo, and Shigeki Misawa. 2021. Access Patterns to Disk Cache for Large Scientific Archive. In *Proceedings of the 2021 Systems and Network Telemetry and Analytics (SNTA '21), June 21, 2021, Virtual Event, Sweden.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3452411.3464444

1 INTRODUCTION

The High Performance Storage System (HPSS) [6] at Brookhaven National Laboratory (BNL) is a large tape storage system for archiving large amounts of data for scientific experiments such as high energy physics and nuclear science [2]. HPSS stores data files on magnetic tapes managed by robotic arms. To read a file on a tape, the robotic arm must travel some distance to retrieve the tape, and place the tape in one of the available readers, then the reader has to spin the tape to locate the file on the tape before the actual reading could take place. Thus, there is a significant overhead to read a single file from a tape. Having a disk cache in front of the tape system would allow in cached files to be read directly without involving the tape. Additionally, the cache system might be able to reorganize the tape accesses to read multiple files from a single tape to amortize the tape access overhead.

One key motivation of this study is to understand whether the access patterns are conducive to some performance engineering. A number of earlier studies of high-energy physics data has found that the order of files specified in many of the analysis jobs are unimportant [2, 8], therefore, the file access requests could possibly be reordered without impacting the user analysis tasks. This exploratory analysis of the dCache log files could find out whether sufficient information is present to enable this reordering of file accesses. We are also interested in additional opportunities that might be revealed through this study.

In this study, we use a set of data access logs from the dCache system for A Toroidal LHC ApparatuS (ATLAS) experiment at the Large Hadron Collider (LHC) that manages a large disk cache in front of the HPSS at BNL [1, 5] [3, 4]. There is one HPSS "Class of Service" consisting of a disk and tape storage hierarchy dedicated exclusively to ATLAS. Disk cache can hold up to 600TB based on hardware RAID. ATLAS Tape drive distribution is 32 Linear Tape-Open 7 (LTO-7), 17 LTO-6, and 6 LTO-4, and there are two production Oracle SL8500 tape libraries. 4.74 PB of disk space is reserved in dCache for files coming out of HPSS (TAPE-stage area), and 350 TB of disk space is reserved for files going into HPSS (TAPE-write area). Files written into the TAPE-write area of dCache are immediately copied to HPSS. There are 80,492,618 files and roughly 76 PB of ATLAS data on tape. When a user data analysis job requests a file from HPSS, if the requested file is in the dCache disks, it will be served to the user, otherwise, dCache will send a retrieval request to HPSS to transfer the file to its disks. In this study of the log files, we will concentrate on understanding the common data access patterns and core statistics about the data accesses [9]. This study will allow us to evaluate the effectiveness of the disk cache system in reducing the traffic to HPSS and the possibility of reducing the overhead of tape mounting or pre-staging files from the tape [4, 7].

Currently we have daily records from January 2019 to March 2020. On each day, there are multiple gigabytes of access logs from the dCache system. The total log file size is about 75.4 GB of compressed (gzipped) CSV file. Our investigation shows that out of 50 requests to this dCache system, only one needs to touch the HPSS system, which indicates that the particular access patterns are well-served by the disk cache installation. Among the accesses to the tape system, about the 70% of the files retrieved are used multiple times, while other 30% are used only once. Potentially, many files from the tapes that are only used once might be due to the unusual access patterns that we have found with a small number of files, where they are repeatedly retrieved from the tape system, up to

SNTA '21, June 21, 2021, Virtual Event, Sweden

ACM ISBN 978-1-4503-8386-8/21/06…$15.00
https://doi.org/10.1145/3452411.3464444

Table 1: Median daily requests of various types to dCache

Store	Restore	Transfer
20,158	3,470	1,164,898

250 times in a day. Such repeated accesses to the same file on the tape is definitely unexpected, and we plan to investigate further.

2 OVERALL DATA ACCESS STATISTICS

There are three types of dCache accesses that are of interest to our study at the moment: store, restore, and transfer. The store operation takes a file generated from somewhere else and store it to HPSS, the restore operation retrieves a file from HPSS and places it on the disk cache, and the transfer request is to pass the data file from the disk cache to a client program.

Table 1 shows the median number of requests per day for these three types. From this table, we see that the total number of the in-cache accesses is more than 1 million a day, while less than 25,000 accesses involve the tape system. Overall, we see that out of 50 data accesses, only one involves the tape system. This indicates that the disk cache is working very well for the time period.

Figure 1 shows the number of files accessed each day during 2019. In both plots, the vertical axis is in the log scale. We see that the total number of files accessed per day is about 3.5 million (in the left plot) and the number of unique files accessed per day is about a million (in the right plot). The number of times each file is repeatedly accessed varies in a fairly narrow range from day to day, where the first quartile of this daily repetition rate is at 3.02, the median is at 3.35, and the third quartile is at 3.75.

3 ACCESSES TO MAIN TAPE ARCHIVE

The actual tape system is divided into a number of tape silos. In this study, we focus on the main tape silo known as LakeTape. Access records involving LakeTape begin to appear in May 2019. In May and September, there were two big inflows of files through the store operations.

The Figure 2a shows a histogram the transfer rates (bytes per second = file size / transfer time) for each file transfer and the Figure 2b shows how the transfer rates vary over time. From the histogram plot of the transfer rates in 2019, we see that the transfer rates span a very wide range. There are a few instances where the transfer rates are only a few kilobytes/second, while some other instances reach 1 GB/s. The first quartile is about 5.4 MB/s (5.4×10^6 bytes/sec); the median is about 19 MB/s (1.9×10^7 bytes/sec); and the third quartile is about 83 MB/s (8.3×10^7 bytes/sec). Note also that the average (mean) of the transfer rates is 76 MB/s (7.6×10^7 bytes/sec), which is close to the third quartile instead of the median. Thus, we observe the distribution of transfer rate is highly skewed.

The Figure 2b shows the average transfer rate per day, along with the standard deviations. From this plot, we see that not only the individual data transfers vary dramatically in speed, the daily average transfer rates also vary quite a bit over time. We observe that there are a number of days with slow transfer rates at the end of May and beginning of September, which forms a trough in the middle of the plot.

Next, we look into these days in more detail on the days with store operations to LakeTape in May and September. Figure 3a shows transfer rate distribution for store operations on LakeTape for 8 days in May 2019, where the black bar indicates the median transfer rate for the day, the widest bar indicates the range of one standard deviation from the median value, the outliers are shown as dots, and the top of the y-axis is 1.0 GB/s. We see the transfer rate distributions have similar ranges but the median transfer rates on the 14th, 16th, 17th and 28th are clearly lower than on other days. Figure 3b shows more transfer rate distributions from the days in September 2019. We notice that the median transfer rates on the 18th, 19th and 20th are noticeably smaller than the rest of the month. Similar to what happened in May, we again see that median transfer rates are much lower than the fastest ones. Potentially, the store operations that require a new tape to write file would like to take more time and have lower transfer rates. However, at this time, the dCache log files does not contain such information for us to verify such a conjecture.

4 DISCUSSION

During the normal operations, once a file enters the disk cache, either through store or restore operations, dCache would serve the disk copy of the file to satisfy the user requests. However, from our investigation, we found a number of files that are accessed in surprising ways.

Figure 4 shows one example involving a file named "AOD.-05536542._000001.pool.root.1.CSV". From Figure 4, we see that from May 29 to June 20, 2019, dCache repeatedly accessed this file from LakeTape instead of the disk copy of the file, while we expect it to be in the disk cache already. The log entries do not contain any indication that this file has been removed nor indications that the file retrieval commands have failed, which suggests that the file should be remained on disk. The number of times this file is accessed from the tape each day is shown as red dots in this figure. On some days, this file is accessed 250 times from the tape, which is very surprising. The reason for these tape accesses could not be determined from the logs.

One speculation about the access pattern to "AOD.05536542.-_000001.pool.root.1.CSV" is that it might not be a normal data file because its name ends with the extension "CSV," while most of the files for high-energy physics ends with the extension "root" indicating that they are ROOT files [3, 8]. In Figure 5, we see another example involving what appears to be a ROOT file named "AOD.-11063346._00957.pool.root.1". In this case, the file is retrieved a few times, three times on May 28, once on September 12, and twice on September 26. Similar to the previous example, we are not able to locate requests to remove this file from disk or indications that the file retrieval commands have failed somehow. We plan to explore this issue further.

5 SUMMARY & NEXT STEPS

So far, in this work, we mainly studied data access statistics such as the number of files accessed per day, transfer rates to the tape system, and so on. Overall, we see that out of 50 file requests only one involves the tape system, which means the disk cache is working very well in reducing the need to access the tapes. Once a file

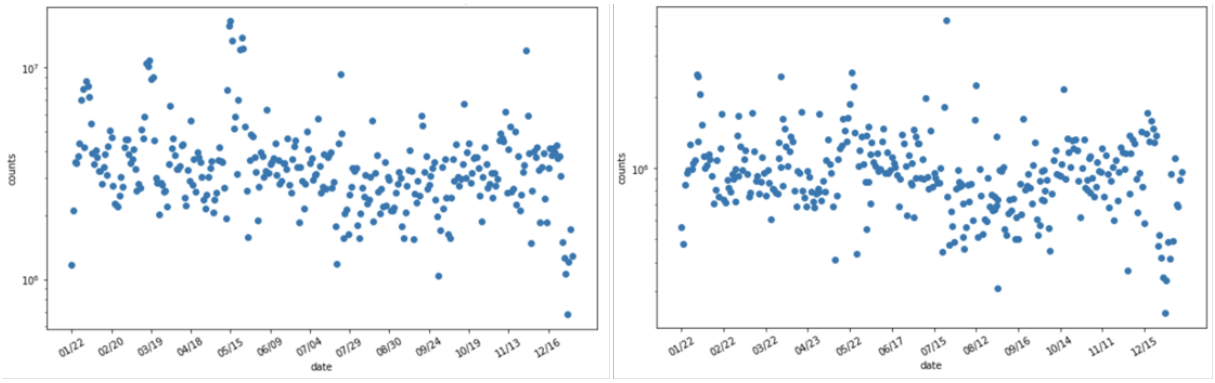
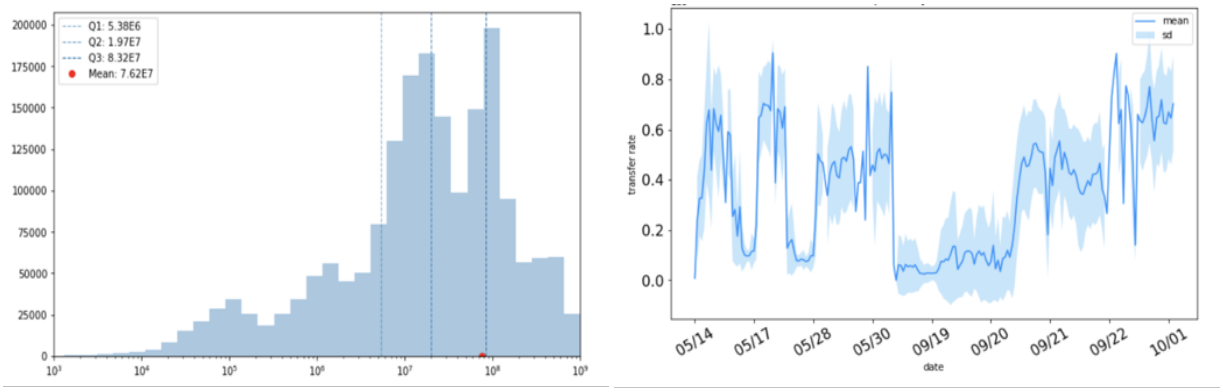
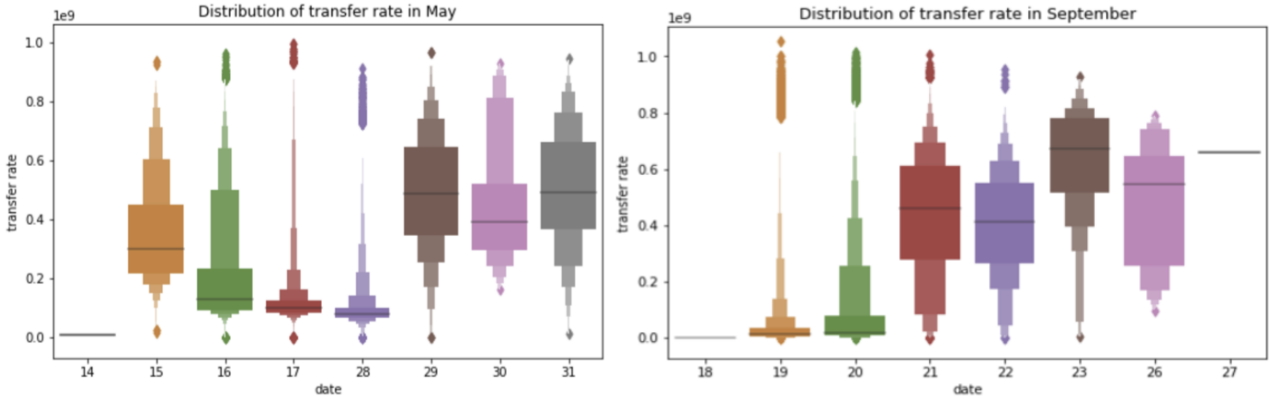


Figure 1: The total number of files accessed (left) and the number of unique files accessed per day during 2019



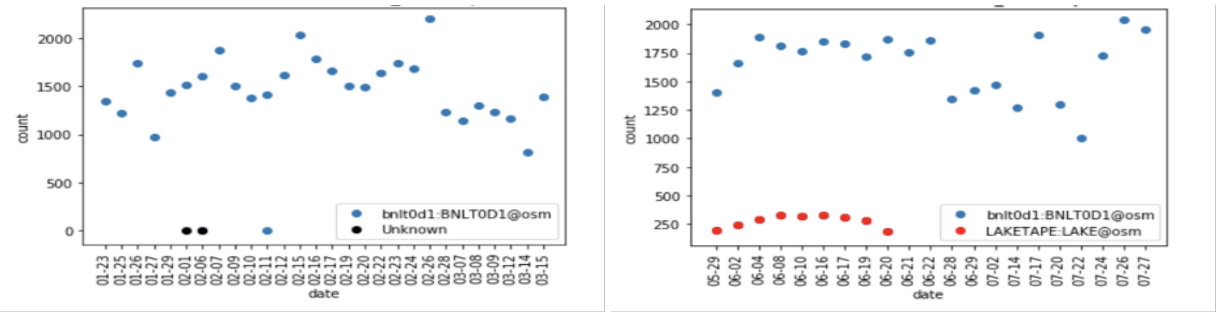
(a) **(b)**

Figure 2: (a) histogram of transfer rates to LakeTape and (b) the variations of transfer rates during the second half of 2019



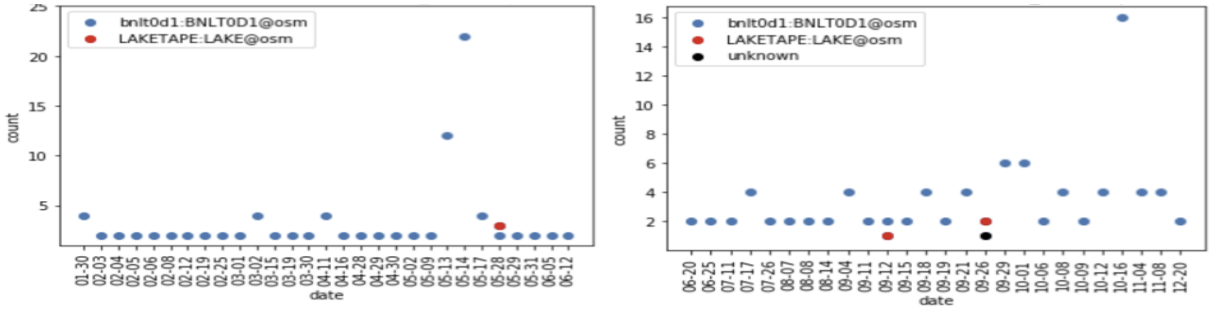
(a) **(b)**

Figure 3: Transfer rates distributions (a) in May 2019 and (b) in September 2019



(a) (b)

Figure 4: Number of accesses during selected periods of 2019 for a file named AOD.05536542._000001.pool.root.1.CSV



(a) (b)

Figure 5: Number of accesses during selected periods of 2019 for a file named AOD.11063346._00957.pool.root.1

arrives at the disk cache, we observe that 70% of them are accessed more than once and the other 30% of the files are accessed only once. On an average day, there are about 3.5 million file accesses to about 1 million unique files.

The disk cache system is meant to serve the disk copy when it is available. However, we see occasionally that a file is retrieved from the tape many times (up to 250 times in a day). Additional information is needed to understand this unexpected access pattern. There are a few days where the transfer rates are consistently lower than other days. It would be useful to study these days further.

We are interested in studying the possibility of reordering the tape accesses, however, this would require additional information such as the tape ID. With such information, we may be able to build a clustering model to predict which files are usually being accessed together or more efficient HPSS file staging policy by pre-staging files before they are requested by users which can reduce the overhead cost from the tape mounting.

ACKNOWLEDGMENTS

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and also used resources of the National Energy Research Scientific Computing Center (NERSC).

REFERENCES

[1] G Behrmann, P Fuhrmann, M Grønager, and J Kleist. 2008. A distributed storage system with dCache. *Journal of Physics: Conference Series* 119, 6 (jul 2008), 062014. https://doi.org/10.1088/1742-6596/119/6/062014

[2] L Bernardo, H Nordberg, D Olson, A Shoshani, A Sim, A Vaniachine, D Zimmerman, B Gibbard, R Porter, T Wenaus, and D Malon. 2001. New capabilities in the HENP Grand Challenge Storage Access System and its application at RHIC. *Computer physics communications* 140 (Oct. 2001), 179–188.

[3] R. Brun and F. Rademakers. 1997. ROOT : AN OBJECT ORIENTED DATA ANALYSIS FRAMEWORK. *Nuclear instruments & methods in physics research, Section A* 289, 1-2 (1997), 81–86.

[4] D. Düllmann. 2008. Improvement Options for LHC Mass Storage and Data Management. https://indico.cern.ch/event/33835/contributions/796554/attachments/662741/910925/dirkd-dm-hepix.pdf

[5] M Ernst, P Fuhrmann, M Gasthuber, T Mkrtchyan, and C Waldman. 2001. dCache, a distributed storage data caching system. *Journal of Physics: Conference Series* (2001).

[6] R. W. Watson and R. A. Coyne. 1995. The Parallel I/O Architecture of the High-Performance Storage System (HPSS). In *Proceedings of the 14th IEEE Symposium on Mass Storage Systems (MSS '95)*. IEEE Computer Society, USA, 27.

[7] Kesheng Wu, Surendra Byna, Doron Rotem, and Arie Shoshani. 2011. Scientific Data Services: A High-Performance I/O System with Array Semantics. In *Proceedings of the First Annual Workshop on High Performance Computing Meets Databases* (Seattle, Washington, USA) *(HPCDB '11)*. Association for Computing Machinery, New York, NY, USA, 9–12. https://doi.org/10.1145/2125636.2125640

[8] Kesheng Wu, Wei-Ming Zhang, Alexander Sim, Junmin Gu, and Arie Shoshani. 2003. Grid Collector: An Event Catalog With Automated File Management. In *Proceedings of IEEE Nuclear Science Symposium 2003.* https://doi.org/10.1109/NSSMIC.2003.1351830

[9] X. Zhang, D. He, D. H. c. Du, and Y. Lu. 2006. Object Placement in Parallel Tape Storage Systems. In *2006 International Conference on Parallel Processing (ICPP'06)*. 101–108. https://doi.org/10.1109/ICPP.2006.55