Predicting Dataset Popularity for Improved Distributed Content Caching in Scientific Workflows

Malavikha Sudarshan[1], Alex Sim[2], Kesheng Wu[2]
[1]University of California at Berkeley, malavikhasudarshan@berkeley.edu
[2]Lawrence Berkeley National Laboratory, {asim, kwu}@lbl.gov

***Abstract*—The vast amounts of data generated by large High Energy Physics (HEP) experiments pose significant challenges for data management and analysis. To mitigate these challenges, distributed caching systems, such as XCache, are used as regional in-network caches to buffer recently accessed data files. By reducing the need for repeated file transfers, in-network caching decreases data access latency and enhances analysis efficiency.**

To better understand the impact of data file popularity on cache effectiveness, we conducted a study of operational logs from Southern California from June 2020 - April 2025, comprising approximately 35 million file requests. Our extensive exploratory data analysis revealed that a small subset of datasets accounts for a disproportionate number of access requests, suggesting that prioritizing these popular datasets in the cache could simplify cache replacement policies and improve access efficiency. However, our analysis also showed that dataset popularity exhibits significant temporal variability, making it challenging to predict future access patterns.

To address this, we developed a Long Short-Term Memory (LSTM) neural network model to forecast dataset access counts and volumes, training and testing this on the last year of data available (April 2024 - April 2025). Our model achieves a mean relative RMSE of 0.406 across the 20 most popular datasets, demonstrating its effectiveness in capturing general access trends. Although further validation is necessary, our results indicate the potential of LSTM-based predictions to enable the practical implementation of a "pinning" approach and improve data access efficiency for large-scale HEP analyses.

I. Introduction

High Energy Physics (HEP) experiments at the Large Hadron Collider (LHC) generate enormous amounts of data that need to be analyzed to extract meaningful insights. This analysis requires data to be transferred around the world, leading to significant network traffic and slowing down scientific discovery. To mitigate this issue, distributed caching systems such as XCache have been deployed [1], [2], [3], [4], [5]. XCache allows data to be shared among users and sites within the same geographical region, reducing redundant file transfers and improving efficiency of data analysis [6], [7], [8].

The importance of XCache in the HEP community is growing rapidly [9], but its effectiveness relies heavily on the cache management strategies employed. A well-designed strategy is crucial to ensure that the cache contains the most relevant and frequently accessed data, and that it is updated regularly to reflect changes in the underlying data and usage pattern. To achieve this, various algorithms and techniques can be used, such as least recently used (LRU) caching, which discards the least recently accessed data first, or predictive caching, which anticipates future data needs and caches accordingly. Through some initial exploration, we have identified a unit of data collection termed by the HEP scientists as 'dataset' to be a good unit for understanding popularity of data records for cache management [10]. This work expands on that exploration to more thoroughly study the prediction accuracy on popular datasets. Ultimately, we plan to "pin" some of the most popular datasets in storage cache to simplify the cache management process and improve the overall data analysis efficiency. This approach is of particular interest for HEP collaborations because it aligns well with the existing process of the cache management policy and has a higher likelihood being integrated into the routine operations of storage caches [11], [12], [13].

Accurate prediction of dataset popularity is crucial to achieving our ultimate goal. Building on earlier research, we have found that Long Short-Term Memory (LSTM) neural network models outperform other approaches in predicting various features [14]. This study investigates the effectiveness of LSTM in predicting dataset popularity, with a focus on assessing the feasibility of implementing a pinning strategy for XCache. To this end, we analyzed a comprehensive dataset of operational logs from an XCache deployment in Southern California, spanning 58 months and comprising of approximately 35 million data access requests. This large and representative sample provides a robust foundation for our exploration of LSTM's potential in predicting dataset popularity. We then focused in on the last year of recorded data, as temporal differences significantly affect the accesses of certain datasets and only using one year to train the model on provided more granularity in predictions.

Prior to exploring the LSTM prediction model, this paper presents a statistical analysis that highlights the complexities of predicting dataset popularity. Our examination reveals that the distribution of dataset popularity is highly skewed across multiple dimensions. For instance, we observe that a small subset of datasets tends to dominate data access requests during any given time period, with the most popular dataset often accounting for a disproportionate share of requests. Furthermore, the popularity of datasets can change rapidly from day to day, with new datasets emerging as popular and others falling out of favor. This volatility in dataset popularity poses significant challenges for predicting which datasets will be in high demand. Despite these challenges, accurately predicting dataset popularity would have a profound impact on

the effectiveness of HEP storage caches, allowing researchers to optimize cache performance and improve access to the vast amounts of HEP data. Our investigation of the LSTM prediction model yields promising results, demonstrating that we can predict not only the number of requests for popular datasets but also the volume of data requested from these datasets. These findings suggest that the LSTM model has significant potential to become a valuable tool for predicting dataset popularity, and we are encouraged by the prospects of leveraging this technology to improve the management and utilization of HEP data.

The remainder of this paper is organized as follows: Section II provides a brief description of the 4-and-three-quarter-years worth of operational logs used as the input for our study. Section III contains a brief survey of statistics on access to datasets, underlying the challenges of making predictions about the popularity of datasets. Section V shows key results of the predictions made with a LSTM model. A concise summary is given in Section V along with a discussion of remaining work.

II. Operational Logs from SoCal

The SoCal Cache, also known as the Southern California Petabyte Scale Cache, is a storage cache that supports computing jobs for the US Compact Muon Solenoid (CMS) experiment, a High Energy Physics (HEP) collaboration. The cache consists of 23 nodes located at CalTech, UCSD, and ESnet at Sunnyvale, with a total storage capacity of approximately 2PB. The nodes are federated together to provide a unified caching layer for the CMS experiment.

The cache serves a variety of file types, including analysis object data (AOD), MiniAOD, and NanoAOD files [15], which differ in their information content and size. The majority of analysis jobs work with either MiniAOD or NanoAOD files, which are smaller and more efficient to transfer. However, some analysis jobs require access to larger file formats which can take longer to transfer and process.

The operational log of the SoCal Cache provides valuable insights into the usage patterns and performance of the cache. The log data was collected over a period of 4 years and 10 months, from June 2020 to April 2025, and consists of approximately 35 million records, including 28 million cache hits and 7 million cache misses. A cache hit occurs when the requested data file is found in one of the federated cache nodes, while a cache miss occurs when the file needs to be transferred from a remote storage location over the wide-area network.

The log data also provides information on the request sizes, file location, and network performance, including remote transfer throughput. This information is extracted from the XCache log files, which are processed and analyzed to derive cache hits, cache misses, and other performance metrics. The SoCal Cache runs on XCache software, which provides a scalable and efficient caching solution for the CMS experiment [5], [4], [16].

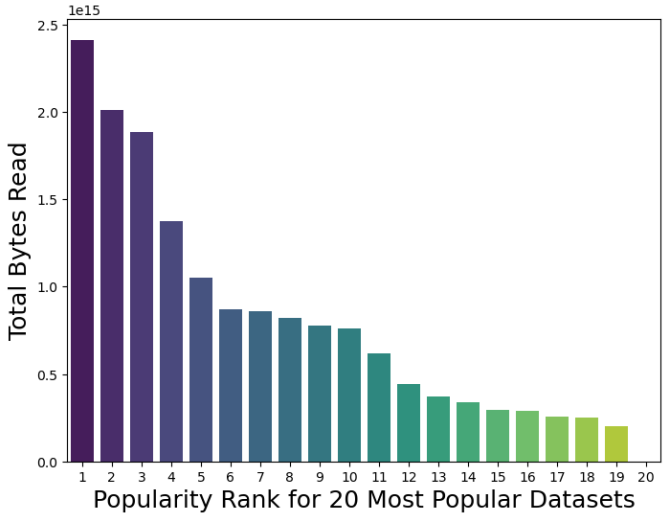

Fig. 1. Bytes Read in the top 20 popular datasets across the entire study period (June 2020 - April 2025).

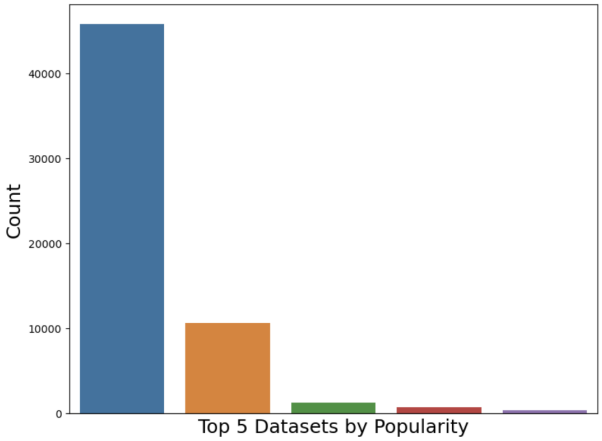

Fig. 2. Distribution of total accesses per dataset in 1 Day (13th June 2023). One dataset is predominantly accessed.

III. Statistics on Data Requests

Next we provide some statistics about the popularity of the datasets [10]. In this work, we measure the popularity of a dataset in two ways, the number of data files requested from the dataset, which we call the *access count*, and the number of bytes requested from the dataset, which we call the *access volume*. Because the different types of files, such as AOD, MiniAOD, and NanoAOD, have vastly different sizes, these two measures may lead different rankings of the datasets. In Figure 1, we show the bytes read (i.e., access volumes) of the top 20 most popular datasets in the whole period of observation, and in Figure 2, we focus on the popular dataset in a day. Despite the differences in time scale and the measure of popularity, one thing is clear: the most popular dataset is considerably more popular than the rest.

From Figure 1, we see that the most popular dataset has an access volume of about 2.4 PB, and the top 5 most popular

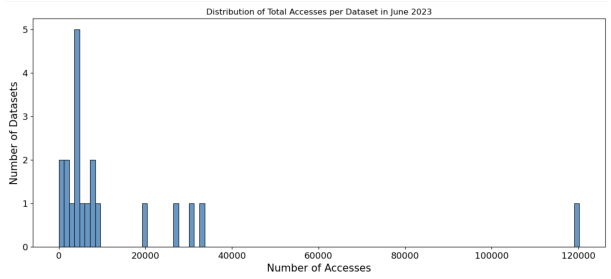

Fig. 3. Distribution of the number of accesses to datasets in June 2023. One dataset is accessed close to 120,000 times in one month, with the rest being accessed much less frequently.

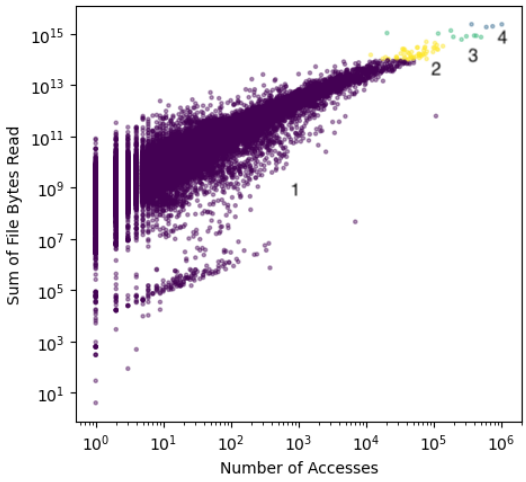

Fig. 4. Scatter plot of access volume against access count of all datasets over the full 56-month period. The coloring of the points are based on the cluster assignment from K-means cluster with $k = 4$. The points representing the most popular datasets form a narrower diagonal line indicating the average file sizes from these datasets are about same, suggesting the same type of file is used in popular data analysis tasks.

datasets have much larger access volumes in comparison to the next 15. Similarly, the most popular dataset is accessed about 45,000 times in a day, shown in Figure 2, which is more than the next 19 datasets combined.

Figure 3 shows a histogram of different access counts of all datasets in June 2023. This histogram indicates that there is a single dataset that is accessed 120,000 times in the month, while the next set of popular datasets are only accessed less than 40,000 times in the same month.

Taken together, Figures 1, 2, and 3 illustrate the skewness of the popularity distribution using two different popularity metrics and three different timescales. From the later figures, we also see that the accesses to datasets are also bursty in that requests come in at irregular time points and different subsets of a dataset are requested each time so that both access counts and access volume are different at each request. All of these features contribute to the challenge of making accurate predictions on the future popularity of a dataset.

The next set of figures (4 and 5) shows scatter plots of access volume against access count of different datasets, where

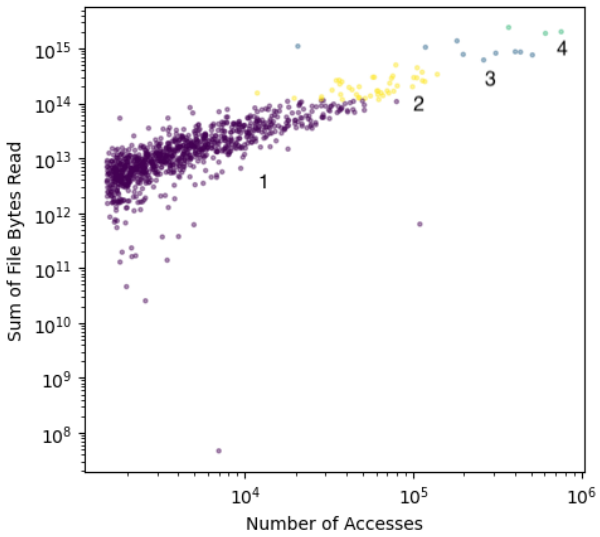

Fig. 5. A zoomed-in view of Figure 4 focusing on the top 1000 most popular datasets.

Figure 4 includes all datasets from the whole study period while Figure 5 only includes the 1000 most popular datasets. The coloring of the points is determined by labels produced from a K-means clustering with $k = 4$. From Figure 4, we see that the datasets with low access counts show very large spread in their access volumes. For example, the datasets are accessed only once have a range from bytes (10^0) to gigabytes (10^9) for their access volumes. In comparison, the range of access volumes are in much narrower range for the access counts near 100,000 (10^5). This suggests that popular analyses typically involve files of a single type, implying that accesses to widely used datasets may be more predictable than those to less popular ones. The next study aims to leverage this observation to explore the predictability of popular datasets.

IV. Predictions with LSTM

Previous studies [11][17] have looked at various approaches to predicting dataset popularity, including classification models, frequent pattern mining, and decision trees. We used a Long Short-Term Memory (LSTM) model to predict dataset behavior [18], as access patterns are sequential and exhibit temporal dependencies, such as irregular spikes. LSTMs are well-suited to capture such long-term dependencies and non-linear relationships across multiple features.

The LSTM model consisted of two LSTM layers with 256 and 128 units, respectively, each using dropout (0.3) and recurrent dropout (0.2) for regularization. A dense output layer was added, and the model was compiled with a custom quantile loss function (tau=0.75) designed to better capture spike patterns, using the Adam optimizer with a learning rate of 0.0005. To handle the highly skewed nature of access patterns with rare extreme spikes, we implemented logarithmic transformation of target values and applied adaptive sample weighting, where the top 1% of spikes received 15X higher

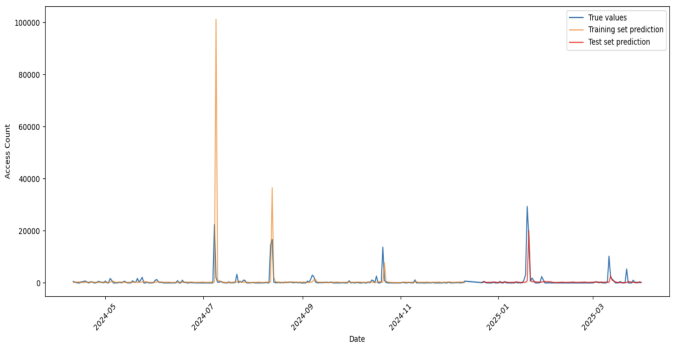

Fig. 6. Daily access count predictions for the most popular dataset from April 2024 to April 2025, RunIISummer20UL18MiniAODv2/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8, using an LSTM model to predict accesses

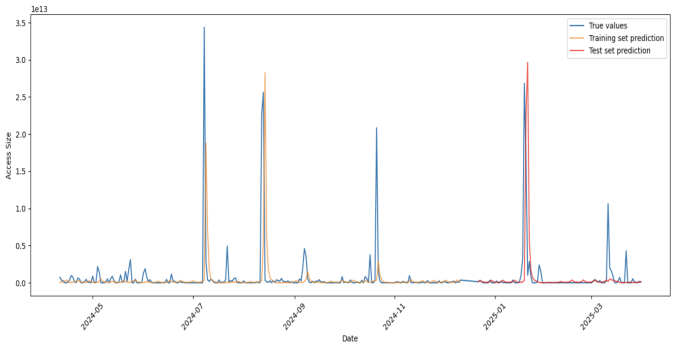

Fig. 7. Daily access size predictions for the most popular dataset from April 2024 to April 2025, RunIISummer20UL18MiniAODv2/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8, using an LSTM model to predict accesses

weights and values above the 90th percentile received 5X weights. To evaluate its performance, we compared the predictions of the most popular dataset with those of the 20th most popular dataset. This comparison allowed us to assess the model's ability to predict access patterns across datasets with varying access frequencies.

The variables used in the model included the file name, timestamp of data access, and data size. These features were chosen to help the model understand both the temporal and quantitative aspects of file access behavior. The input data to the LSTM model was preprocessed for a few additional derived features, including the access count (total number of accesses per file), access size (total bytes read per file), cache hit count and hit size, and cache miss count and miss size, as well as lagged access count for up to seven days to give temporal context. These features were standardized using StandardScaler, while access count (the prediction target) was log-transformed and scaled with a MinMaxScaler to handle skewness and extreme spikes. To handle missing data in the input, we filled in the gaps during the study period with rows containing zero values for all columns. Any NaN or infinite values were replaced with zeros to prevent issues during training. We then split the data into training and validation sets, using 70% of the data for training and the remaining 30% for validation.

Figs. 6 and 7 display predictions for the daily access count and access size of the most popular dataset, "RunIISummer20UL18MiniAODv2/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8," and Figs. 8 and 9 show access predictions for the 6th most popular dataset, "RunIISummer20UL16MiniAODv2/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8". The second dataset was chosen due to the clear visualization of its "spikiness," demonstrating periods with many accesses on a few dates interspersed with intervals of no access during the study period. In these four figures, the vertical axis are daily access metrics and the horizontal axis is time. The blue lines are for actual observations, orange lines are for training results, and red lines are predictions. We see that only some days have non-zero accesses, which causes the lines to appear as a set of sharp spikes. The test

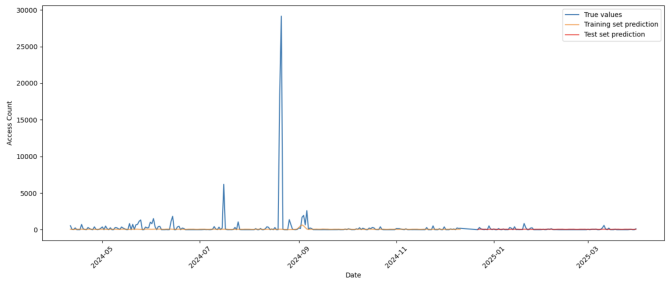

Fig. 8. Daily access count predictions for the 6th most popular dataset, RunIISummer20UL16MiniAODv2/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8, using an LSTM model to predict accesses from April 2024 to April 2025

results could follow these spikes reasonably well, while there are noticeable discrepancies between the orange training lines and the corresponding blue lines. Next, we provide a quantitative measure of these discrepancies.

For measuring the discrepancies between predictions and actual measurements, we use relative Root Mean Squared Error (RMSE), where we compare the RMSE of the predictions against the standard deviation of the actual measurements. When this relative RMSE is less than 1, it indicates that the prediction is somewhat following the general trend of the actual measurements. This relative RMSE error for daily access count is shown in Fig. 10. The mean relative error for

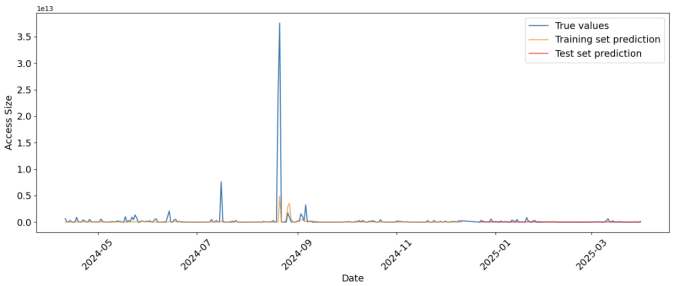

Fig. 9. Daily access size predictions for the 6th most popular dataset, RunIISummer20UL16MiniAODv2/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8, using an LSTM model to predict accesses from April 2024 to April 2025

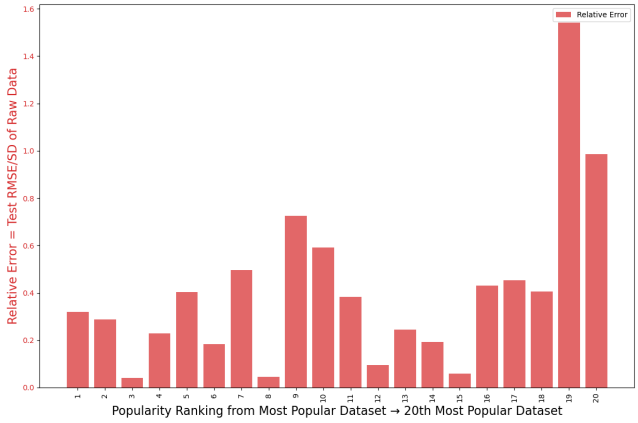

Fig. 10. Relative error values for the top 20 most popular datasets. ie 90% of the 20 most popular datasets have RMSE < 1

the top 20 datasets was 0.406, and increased as the popularity of the dataset decreased, which makes sense as the less popular datasets have fewer access events and are therefore more difficult to predict accurately. Relative Error values > 1 indicate challenges in predicting certain datasets, particularly those with irregular access patterns.

V. Conclusion

Efficient data management is critical for large distributed scientific workflows, where accurate prediction of dataset popularity can significantly enhance caching strategies and resource utilization. This work introduces a novel machine learning approach in a HEP context leveraging Long Short-Term Memory (LSTM) models to forecast dataset popularity and anticipate future access patterns. Our LSTM-based model demonstrates robust performance in capturing general access trends, achieving a mean relative error of 0.406 across the top 20 popular datasets. These preliminary results underscore the potential of machine learning models to predict access patterns, particularly for popular datasets, enabling optimized caching policies and more efficient use of storage systems. Further comparison against other machine learning techniques will be prioritized in future work. Future research will focus on validating this approach with storage cache system operators, exploring transferability to other cache deployments, and integrating anomaly detection techniques to improve the robustness of the model in handling irregular access patterns, ultimately improving its reliability for real-world applications.

Acknowledgment

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] E. Fajardo, D. Weitzel, M. Rynge, M. Zvada, J. Hicks, M. Selmeci, B. Lin, P. Paschos, B. Bockelman, A. Hanushevsky, F. Würthwein, and I. Sfiligoi, "Creating a content delivery network for general science on the internet backbone using XCaches," *EPJ Web of Conferences*, vol. 245, p. 04041, 2020.

[2] X. Espinal, S. Jezequel, M. Schulz, A. Sciabà, I. Vukotic, and F. Wuerthwein, "The quest to solve the hl-lhc data access puzzle," *EPJ Web of Conferences*, vol. 245, p. 04027, 2020.

[3] L. Bauerdick, D. Benjamin, K. Bloom, B. Bockelman, D. Bradley, S. Dasu, M. Ernst, R. Gardner, A. Hanushevsky, H. Ito, D. Lesny, P. McGuigan, S. McKee, O. Rind, H. Severini, I. Sfiligoi, M. Tadel, I. Vukotic, S. Williams, F. Würthwein, A. Yagil, and W. Yang, "Using xrootd to federate regional storage," *Journal of Physics: Conference Series*, vol. 396, no. 4, p. 042009, 2012.

[4] L. Bauerdick, K. Bloom, B. Bockelman, D. Bradley, S. Dasu, J. Dost, I. Sfiligoi, A. Tadel, M. Tadel, F. Wuerthwein, A. Yafil, and the CMS collaboration, "Xrootd, disk-based, caching proxy for optimization of data access, data placement and data replication," *Journal of Physics: Conference Series*, vol. 513, no. 4, 2014.

[5] D. Weitzel, M. Zvada, I. Vukotic, R. Gardner, B. Bockelman, M. Rynge, E. Hernandez, B. Lin, and M. Selmeci, "Stashcache: A distributed caching federation for the open science grid," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*, 2019.

[6] C. Sim, K. Wu, A. Sim, I. Monga, C. Guok, F. Wurthwein, D. Davila, H. Newman, and J. Balcas, "Effectiveness and predictability of in-network storage cache for scientific workflows," in *IEEE International Conference on Computing, Networking and Communication (ICNC 2023)*, 2023.

[7] R. Han, A. Sim, K. Wu, I. Monga, C. Guok, F. Wurthwein, D. Davila, J. Balcas, and H. Newman, "Access trends of in-network cache for scientific data," in *5th ACM International Workshop on System and Network Telemetry and Analysis (SNTA 2022)*, 2022.

[8] E. Copps, H. Zhang, A. Sim, K. Wu, I. Monga, C. Guok, F. Wurthwein, D. Davila, and E. Fajardo, "Analyzing scientific data sharing patterns with in-network data caching," in *4th ACM International Workshop on System and Network Telemetry and Analysis (SNTA 2021)*, 2021.

[9] J. Zurawski, D. Carder, E. Colby, E. Dart, C. Hawk, K. Miller, A. Patwa, K. Robinson, and A. Wiedlea, "High energy physics network requirements review: Two-year update," Energy Sciences Network, Tech. Rep. OSTI ID:2405935, 2023. [Online]. Available: https://doi.org/10.2172/2405935

[10] Bellavita, Julian, Sim, Caitlin, Wu, Kesheng, Sim, Alex, Yoo, Shinjae, Ito, Hiro, Garonne, Vincent, and Lancon, Eric, "Understanding data access patterns for dcache system," *EPJ Web of Conf.*, vol. 295, p. 01053, 2024.

[11] E. Meoni, M. Ananya, F. Fanzago, F. Aversa, N. Tonellotto, and N. Magini, "Exploiting data popularity to improve transfer performance in cms," in *Proceedings of the 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID 2018)*, 2018.

[12] F. Jiang, C. Castillo, and S. Ahalt, "Cachalot: A network-aware, cooperative cache network for geo-distributed, data-intensive applications," in *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*, 2018.

[13] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: Moving from cloud to edge," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 36–42, 2016.

[14] S. M. P. K. Patra, M. Sahu and R. K. Samantray, "File access prediction using neural networks," *IEEE Transactions on Neural Networks*, vol. 21, no. 6, pp. 869–882, 2010.

[15] Rizzi, Andrea, Petrucciani, Giovanni, and Peruzzi, Marco, "A further reduction in cms event data for analysis: the nanoaod format," *EPJ Web Conf.*, vol. 214, p. 06021, 2019.

[16] E. Fajardo, A. Tadel, M. Tadel, B. Steer, T. Martin, and F. Würthwein, "A federated xrootd cache," *Journal of Physics: Conference Series*, vol. 1085, p. 032025, 2018.

[17] S. Chung and J. Lee, "Efficient caching of internet data," *Computers & Operations Research*, vol. 28, no. 10, pp. 1023–1033, 2001.

[18] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.