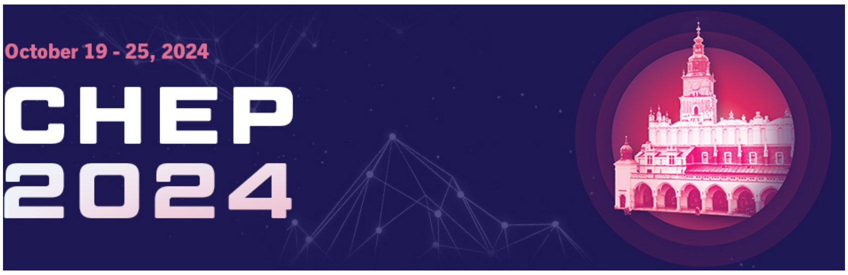
Conference on Computing in High Energy and Nuclear Physics



Contribution ID: **67**

Type: **Poster**

Exploring Data Caching Policy with Data Access Patterns from dCache Logs

Monday 21 October 2024 16:00 (15 minutes)

The dCache storage management system at Brookhaven National Lab plays a vital role as a disk cache, storing extensive datasets from high-energy physics experiments, mainly the ATLAS experiment. Given that dCache' s storage is significantly smaller than the total ATLAS data, it's crucial to have an efficient cache management policy. A common approach is to keep files that are accessed often, ready for future use. In our research, we analyze both recent and past patterns of file usage to predict the chances of them being needed again. Although dCache considers each file separately, we've observed that files within a dataset tend to be used together. Therefore, the system manager often gets requests to retain entire datasets in the cache, especially if they're expected to be in high demand soon. Our main focus is to determine if we could accurately forecast a dataset's future demand to automate the process of deciding which datasets to prioritize in the cache.

Our approach's cornerstone is a dynamic learning mechanism that regularly analyzes recent access logs. This process updates our machine learning models, enabling them to forecast the popularity of various datasets shortly. Specifically, our predictive model estimates the expected number of accesses for each dataset in the upcoming days. We then synchronize these predictions with the cache space allocated for monitoring sought-after datasets. This allows us to proactively load the most in-demand datasets into the disk cache. This strategic reservation method operates in conjunction with the current file removal policy, collectively improving the overall efficiency of the system.

To develop a predictive model for our caching system, we assessed several techniques and metrics to distinguish popular datasets from less popular ones effectively. Employing k-means clustering, we categorized datasets based on their popularity and explored diverse methods to precisely measure dataset usage. Given our constrained disk space, our aim was to optimize the selection of retained datasets, thereby improving cache efficiency.

Prior study [1] has demonstrated the feasibility of detecting popular datasets using a machine learning approach. In this study, we compare the predictive efficacy of two distinct models: a neural network model and a gradient-boosted trees regression model (XGBoost). The models, configured with 17 input variables, are trained on 127 million data points, collected over a span of three years from our data processing pipeline. Additionally, both models underwent hyperparameter tuning via Optuna, conducted on Perlmutter at NERSC.

<img src=”https://sdm.lbl.gov/students/chep24/xgboost$_{d}ecember_{2}023.png”width = ”350” >$
$Fig.1.December2023comparisonofpredictedandactualdatasetaccessesusingXGBoost.Axesrepresentnext-$
$dayactual(x)vs.predicted(y)accesses.Pointsarecoloredbasedontherecencyoflastaccess,withlighterpointsindicatingpredict$

Despite the inherent difficulty in forecasting future dataset accesses, our models showed promising performance. Notably, the XGBoost model displayed a lower root mean squared error (RMSE) for testing datasets compared to the neural network. Specifically, the relative ratios of testing RMSE to standard deviation were 0.28 for XGBoost and 0.84 for the neural network models.

Our research confirms that predicting dataset popularity is feasible through careful analysis of data features and the application of well-designed models. While the real-world application of these models in live caching policies requires further testing, our study underscores the potential of machine learning in improving dCache systems. Future endeavors will concentrate on implementing, benchmarking, and validating the efficacy of these proposed methods.

REFERENCES
[1] J. Bellavita, C. Sim, K. Wu, A. Sim, S. Yoo, H. Ito, V. Garonne, and E. Lancon, “Understanding data access patterns for dcache system,” in 26th International Conference on Computing in High Energy & Nuclear Physics (CHEP2023), 2023.

Primary author: ALDRICH, Jacob (University of California, Berkeley)

Co-authors: SIM, Alex; LANCON, Eric Christian (Brookhaven National Laboratory (US)); ITO, Hironori (Brookhaven National Laboratory (US)); WU, John (LAWRENCE BERKELEY NATIONAL LABORATORY); YOO, Shinjae; GARONNE, Vincent (Brookhaven National Laboratory (US))

Presenter: WU, John (LAWRENCE BERKELEY NATIONAL LABORATORY)

Session Classification: Poster session

Track Classification: Track 1 - Data and Metadata Organization, Management and Access