Predicting Resource Usage Trends with Southern California Petabyte Scale Cache

Caitlin Sim\(^1\), Kesheng Wu\(^2\), Alex Sim\(^3\), Inder Monga\(^3\), Chin Guo\(^3\)
Frank Würthwein\(^4\), Diego Davila\(^4\), Harvey Newman\(^5\), Justas Balca\(^5\)
University of California, Berkeley\(^3\), Lawrence Berkeley National Laboratory\(^2\), Energy Sciences Network\(^3\)
University of California at San Diego\(^4\), California Institute of Technology\(^5\)
caitlinsim@berkeley.edu\(^1\), {kwu,asim}@lbl.gov\(^2\), {imonga,chin}@es.net\(^3\)
{fkw,davila}@ucsd.edu\(^4\), {newman,jbalca}@hep.caltech.edu\(^5\)

There has been a significant increase in data volume from various large scientific projects, including the Large Hadron Collider (LHC) experiment. The High Energy Physics (HEP) community requires increased data volume on the network, as the community expects to produce almost thirty times annual data volume between 2018 and 2028 [1]. To mitigate the repetitive data access issue and network overloading, regional data caching mechanisms [2], [3], or in-network cache has been deployed in Southern California for the US CMS, and its effectiveness has been studied [4], [5]. With the number of redundant data transfers over the wide-area network decreasing, the caching approach improves overall application performance as well as network traffic savings.

In this work, we examined the trends in data volume and data throughput performance from the Southern California Petabyte Scale Cache (SoCal Repo) [6], which includes 24 federated caching nodes with approximately 2.5PB of total storage. From the trends, we also determined how much a machine learning model can predict the network access patterns for the regional data cache. The fluctuation in the daily cache utilization, as shown in Figure 1, is high, and it is challenging to build a learning model to follow the trends.

The daily and hourly study also modeled the cache utilization and data throughput performance, with 80% of the training data and 20% of the testing data. Figure 2 shows the samples of our hourly study results. The root-mean-square error (RMSE) is measured and compared to the standard deviation of the input data values to provide a reference to determine how large the errors of predictions are. The relative error, ratio of testing RMSE and standard deviation, is less than 0.5, indicating the predictions are pretty accurate.

The study results can be used to optimize the cache utilization, network resources, and application workflow performance, and become the base for exploring characteristics of other data lakes as well as examining longer term network requirements for the data caches.

**REFERENCES**


