Analyzing Dataset Popularity for Optimizing In-network Storage

Gunwoo Kim
University of California, Davis
gnwkim@ucdavis.edu

Alex Sim (advisor), Kesheng Wu (advisor)
Lawrence Berkeley National Laboratory
Berkeley, California, USA

I. Introduction

In High Energy Physics (HEP), large-scale experiments generate massive amounts of data that are stored in globally distributed storage systems. To reduce redundant data transfers and improve analysis efficiency, a disk caching system named XCache is used to manage data accesses. While XCache has shown promise in reducing network traffic, its utility is highly dependent on the cache management strategy. This work studies a specific approach known as "pinning." Instead of considering pinning individual files, we consider sets of files known as *datasets* in the HEP community. If one could predict future popularity, the cache could pin the most popular datasets. Various techniques have been explored to predict dataset popularity [1]. In general, popular dataset accesses come in random bursts. This kind of event is commonly modeled as Hawkes process and we explore the possibility of incorporating it to improving predictability of popularity. Furthermore, we compare the performance of Hawkes to a more common approach, LSTM.

II. Methods

The data is from Southern California Petabyte Scale Cache (SoCal Cache), which consists of 23 nodes [1]. The study period was from 2022-08 to 2025-03 and approximately contained 10.9 million records. Each record includes the names of the requested files, the operation's start and end time, whether it was a cache hit or miss, and additional details.

The file accesses are aggregated into "datasets" for this work. The *access times* of each dataset were modeled using a statistical model known as Hawkes process.

Hawkes process models events that have self-exciting behavior [2]. This means that if an event occurs, the next events are more likely to occur. This behavior is apparent in the access logs of datasets, as shown in Fig. 1. In this case, no request to this data occurs at the beginning, a burst of accesses occur, and finally the access count dies down. Furthermore, the distribution of *intra-arrival times* of a Hawkes tends to be heavy-tailed. Looking at the intra-arrival time in Fig. 2, it is clear that they form a heavy-tailed distribution. This is generally true for any datasets.

In this paper, a Hawkes process is modeled with three parameters: μ, α, and β. The μ models the ambient arrivals. The α and β controls how past arrivals affect the current rate of arrivals. We specify the intensity function used below.

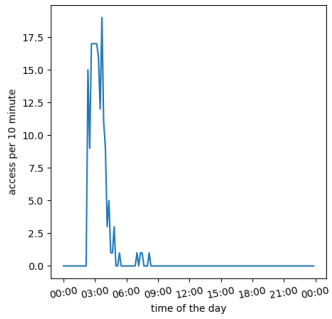

Fig. 1. Access pattern of a single dataset during a single day.

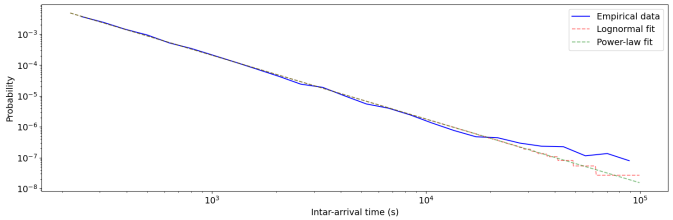

Fig. 2. Plot showing the result of fitting two heavy-tailed distributions, lognormal and power-law, on the intra-arrival times of a dataset. The plot is in log-log scale.

$$\lambda(t|H_t) = \mu + \sum_{t_i<t} \alpha e^{-\beta(t-t_i)}$$

We chose an exponential kernel under the assumption that past accesses that has occurred long time ago has no effect on the current rate of arrivals. The parameters of the model are determined using maximum likelihood estimation (MLE). The fitted model is further validated using the KS test. Conventionally, we reject the model if $p < 0.05$. Hawkes is fitted on the top 30 most accessed dataset to see if long-term behavior of dataset popularity follows the Hawkes process.

We further examine the potential for the model by making *monthly access count* prediction. Hawkes is trained on 60 days of true values, predicts the next 30 days, then slides forward 30 days and repeats. The prediction is done by computing the expectation. We compute the expectation of the next 30 days, given the first 60 days, i.e. $\mathbb{E}[N(90) - N(60)|H_{60}] \approx \mathbb{E}[N(30)]$. Furthermore, using a closed form formula, we can

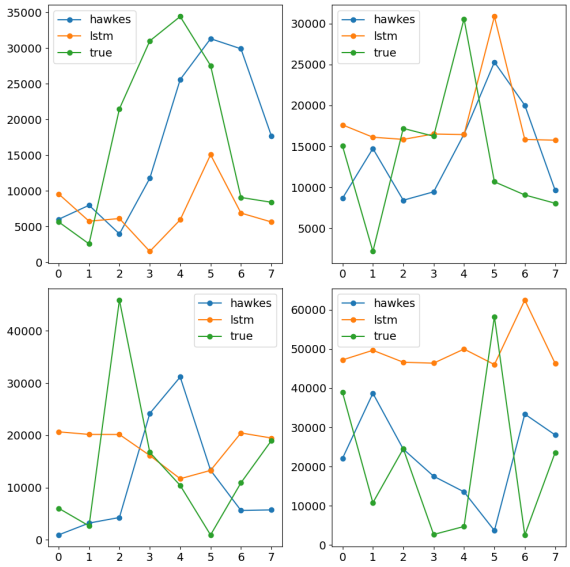

Fig. 3. Result of monthly prediction for LSTM and Hawkes. Showing 4 out of 200 datasets. There were 8, 30-day windows, hence, there are 8 points in the x-axis. The y-axis is the monthly access count.

easily compute the expectation [2].

$$\mathbb{E}[N(t)] = \frac{\mu t}{1-\alpha/\beta} + \frac{\alpha\mu}{(\alpha-\beta)^2}\left(e^{-t(\beta-\alpha)} - 1\right)$$

Finally, we compare Hawkes to LSTM. We train LSTM to predict the monthly access count. To do this, we first train it to predict the daily access count and perform *recursive multi-step prediction* to predict monthly accesses. The model uses a hidden size of 2, 2 layers, with FC layer for output. The training consisted of 1000 epochs, with 0.01 learning rate, using MSE loss, and Adam optimizer. The training data was preprocessed using min-max scaling. For each dataset, LSTM was trained on dates before 2024-08-01. Predictions were made for 30 day intervals, beyond that date. For each monthly count, all true values were given to the trained LSTM model, performed multi-step prediction, and summed up. The comparison between Hawkes and LSTM was made by computing the mean squared error (MSE) against the true values. The errors of Hawkes and LSTM were compared across 200 datasets. We choose the top 200 because rest of the datasets are not accessed more than 1000 times, therefore not likely to be popular anyways.

III. Results

As seen in Table I, only 7 out of 30 top datasets pass the KS test. It is *unlikely* that, in the scale of the entire study period, access pattern of top datasets is Hawkes. This is likely because the popularity of a dataset over a long period of time is highly dependent on human behavior, e.g. what topic is currently popular among HEP community. Thus, the global trend of popularity cannot be captured using a simple model.

Monthly predictions seem to give a more promising result. For 161 out of 200 datasets, Hawkes had a lower error than LSTM. In Fig. 3, few prediction results are shown.

TABLE I
Results of Fitting Hawkes on Top 30 Datasets

Rank	p				
1	**0.050**	11	0.017	21	0.002
2	**0.390**	12	0.000	22	0.000
3	0.015	13	**0.164**	23	0.002
4	0.040	14	0.040	24	0.002
5	0.000	15	0.000	25	0.000
6	0.000	16	0.000	26	**0.094**
7	0.019	17	0.000	27	0.000
8	0.000	18	0.000	28	**0.091**
9	0.000	19	0.000	29	0.000
10	**0.056**	20	0.003	30	**0.114**

In the case of top right, LSTM outperforms Hawkes. For bottom right, LSTM "hallucinates" and produces a constant value. The constant value prediction is caused by mostly-zero training examples. Hence, LSTM is heavily affected by the learning examples, while Hawkes is much less affected by the anomalous learning examples. This is further confirmed by the fact that for the majority of the predictions, Hawkes outperformed LSTM. It is also important to note that, despite the fact that many of the top datasets are not Hawkes in the scale of the entire study period, but when used to make predictions about short term accesses, Hawkes performs well.

IV. Conclusion

Efficient access to large-scale datasets is fundamental to advancing scientific discovery. Improving cache management through accurate prediction of dataset popularity therefore represents a critical opportunity to enhance scientific workflow performance. This work demonstrates the potential of Hawkes processes for modeling dataset popularity dynamics. Although Hawkes processes do not fully capture the global behavior of highly accessed datasets, our results indicate their effectiveness as short-term local predictors, yielding statistically significant improvements in prediction accuracy.

Future research directions include estimating the underlying probability density function to enable probabilistic characterization of anomalous access patterns. Given the absence of a closed-form PDF for this process, Monte Carlo simulation combined with histogram-based density estimation will be employed to establish probability thresholds and quantify the likelihood of extreme access events.

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] 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 *International Conference on Computing, Networking and Communication (ICNC 2023)*, IEEE. IEEE, 2023.

[2] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.