Comparing Cache Utilization Trends for Regional Scientific Caches with Transfer Learning Models

Erica Wang[1], Alex Sim (advisor)[2], K. John Wu (advisor)[2]
[1]California Institute of Technology, [2]Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY
Office of Science

ABSTRACT

To enhance data sharing and reduce access latency in scientific collaborations, High Energy Physics (LHC CMS experiment) employs regional in-network storage caches. Accurate predictions of cache utilization trends help design new caching policies and improve capacity planning. This study leverages the SoCal cache access trends to improve the prediction on the newer caches in Chicago and Boston through transfer learning. We also investigate the impact of doubling the Chicago cache's storage capacity on its cache hit rate.

BACKGROUND INFO

- High-Luminosity Large Hadron Collider expects data volume increase 10x by 2029 and regional data cache need to improve data sharing.
- Study uses data from Southern California Petabyte Scale Cache, the Chicago Regional Cache, and the Boston Regional Cache
- SoCal cache consists of 23 XCache nodes with 2 PB storage capacity
- Chicago cache consists of 6 XCache nodes with 340 TB storage capacity
- Boston cache consists of one XCache node with 150 TB storage capacity.

Table 1: Summary data access of regional caches

		# of accesses	Cache misses (TB)	Cache hits (TB)
6/2020-3/2024 SoCal	Total	30,938,524	11,837.38	26,686.73
	Daily average	22,114	8.46	19.08
10/2022-3/2024 Chic	Total	9,054,942	13,315.37	12,494.41
	Daily average	16,553	24.34	22.84
8/2023-3/2024 Bost	Total	27,754,028	87,311.58	6,173.02
	Daily average	114,214	359.31	25.40

CACHE UTILIZATION

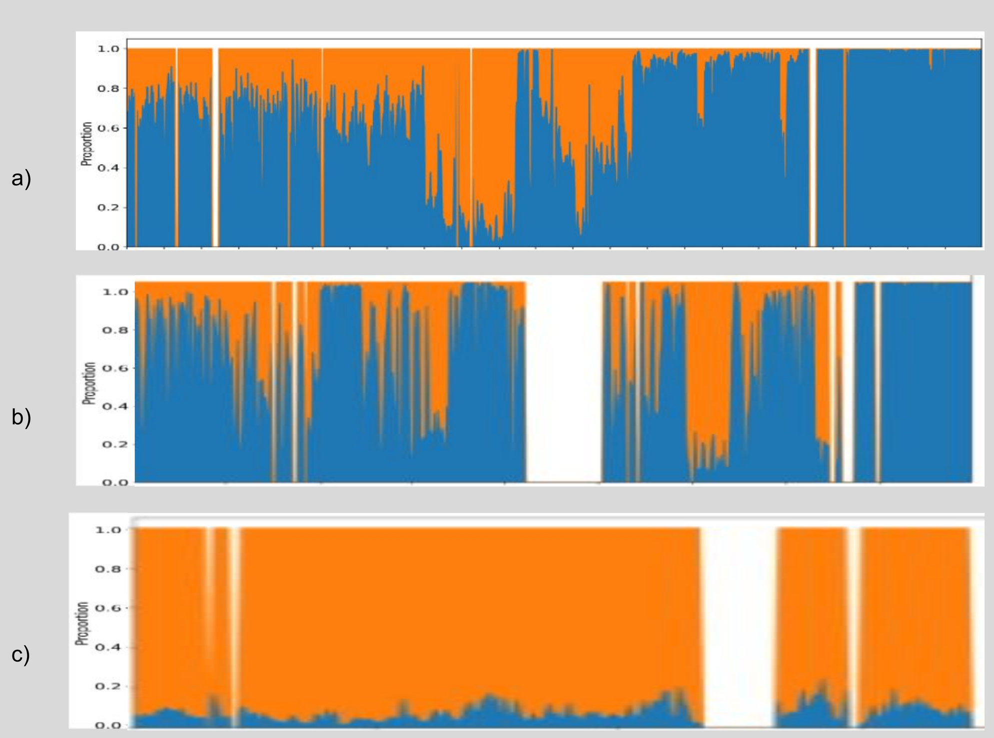

Figure 1: Overview of daily volume of cache hits and misses for: a) SoCal cache, with 69.3% of the traffic reduction overall and 94% reduction during the last year, b) Chicago cache, with 48.4% of the traffic reduction, c) Boston cache, with 6.6% of the traffic reduction.

RESEARCH QUESTION

How can we use measurement data from SoCal cache to predict Chicago and Boston cache utilization trends? If we double Chicago cache size, how will access and utilization trends change?

LSTM AND TRANSFER MODELING

- Goals:
  - Model network and cache utilization trends
  - Plan additional cache deployments
- Traditional cache replacement policies (LRU or FIFO) rely on straightforward algorithms or statistical analysis
  - Assume fixed patterns, cannot adapt to temporal access patterns
- LSTM models excel at capturing complex dependencies in long-term sequential data
  - Leverages temporal locality of data
  - Adapts to dynamic access patterns
  - Incorporates robust input (time, hit count, etc)

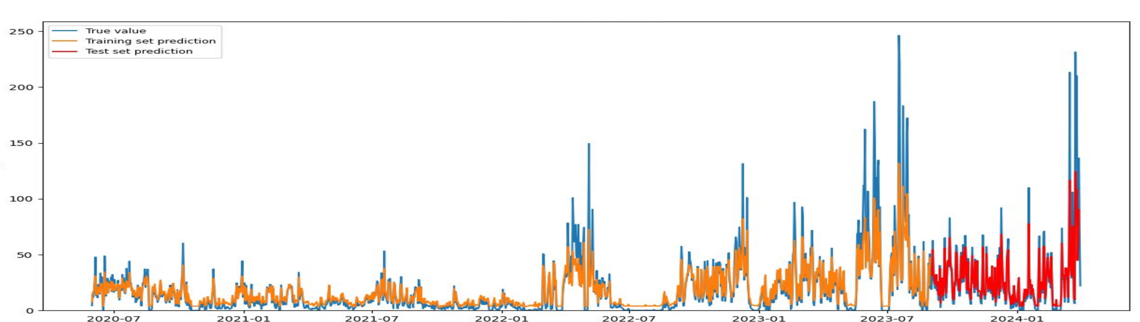

Figure 2: Daily cache hit volume prediction of cache hits of the SoCal cache. The normalized test RMSE is 0.5, indicating an accurate prediction.

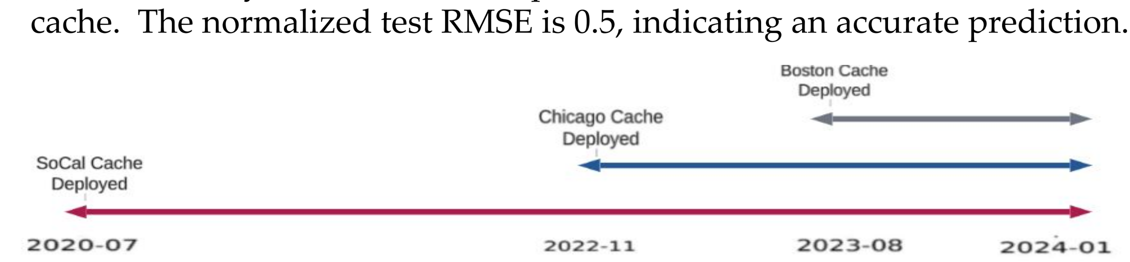

Figure 3: Timeline of regional cache deployment. The SoCal cache has a much longer history providing more training and testing data.

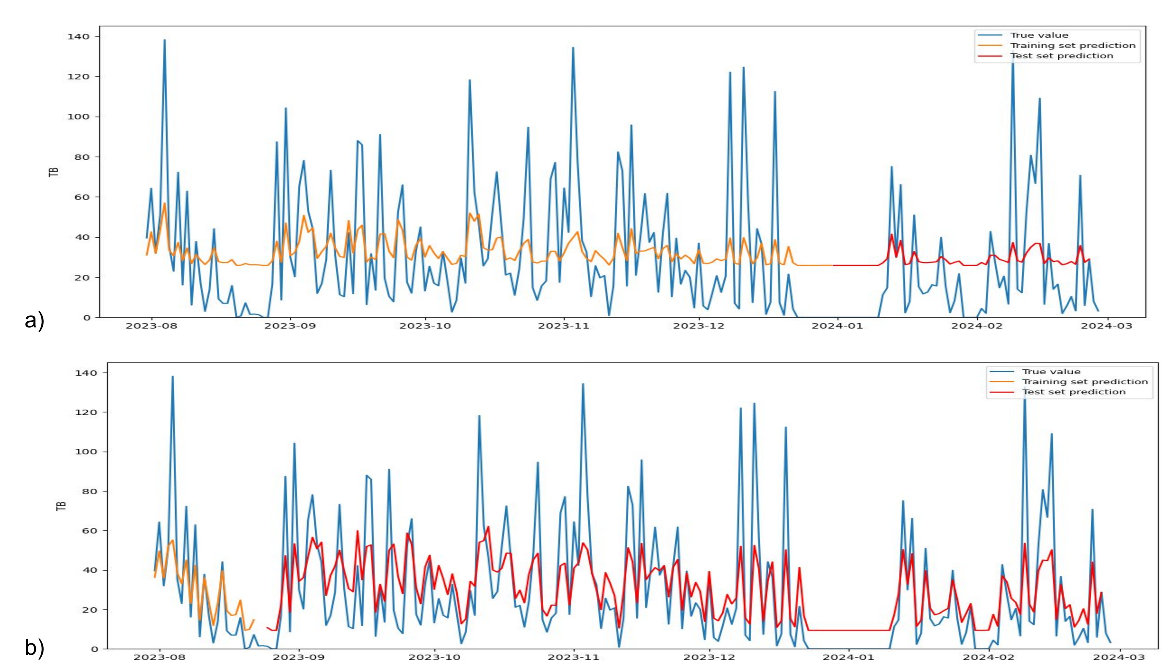

Figure 4: Daily cache hit volume predictions for the Boston regional cache for a) no transfer learning, 80% training size and b) transfer learning, 15% training size. Normalized RMSE error drops from 0.84 (without transfer learning) to 0.64 (with transfer learning).

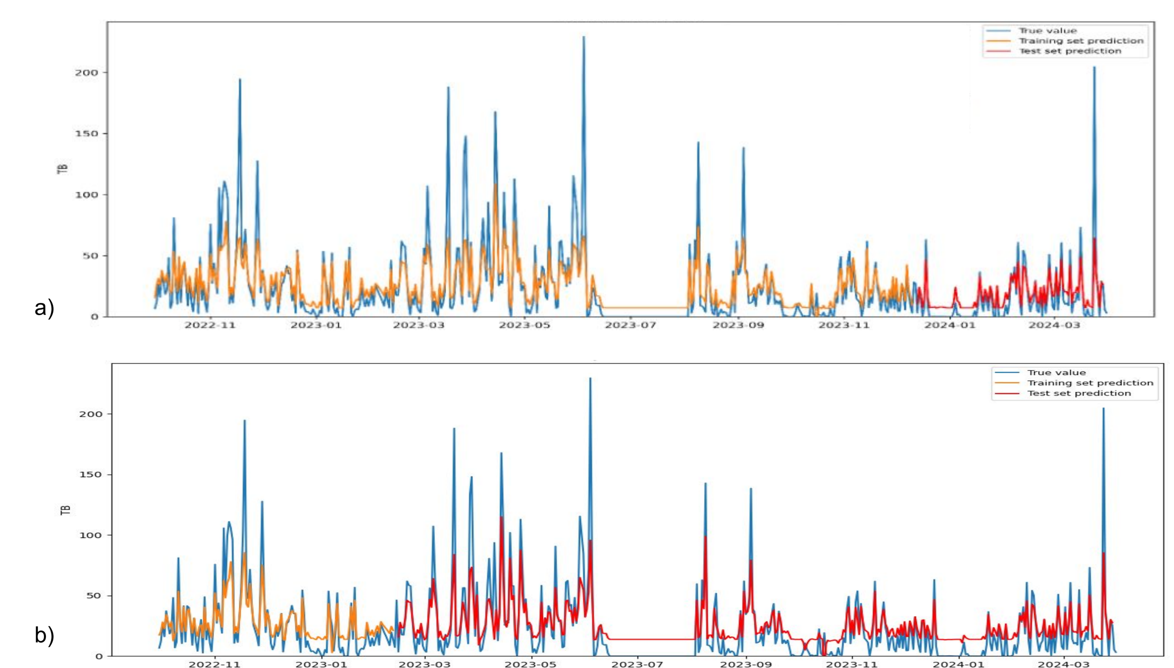

Figure 5: Daily cache hit volume predictions for the Chicago regional cache for a) no transfer learning, 80% training size and b) transfer learning, 20% training size. Normalized RMSE is 0.50 before and after transfer learning.

PREDICTIONS WITH INCREASED STORAGE SIZE

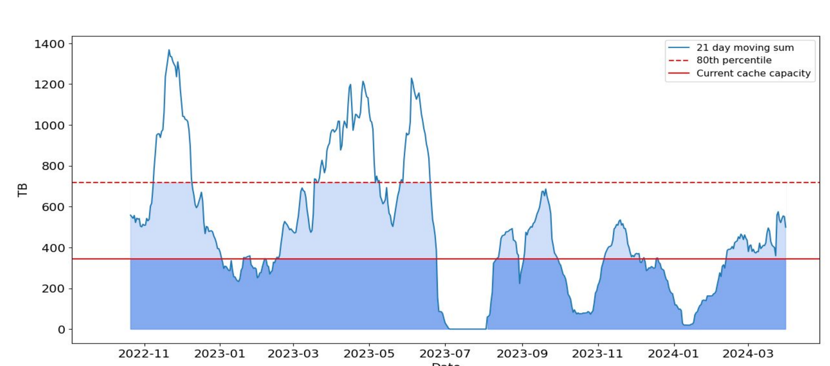

Figure 6: This figure shows the sizes of files if each of them stay in the cache for 21 days. The Chicago cache (340 TB) is able to accommodate the request bytes on 50% of days. Doubled storage capacity predicts 80%.

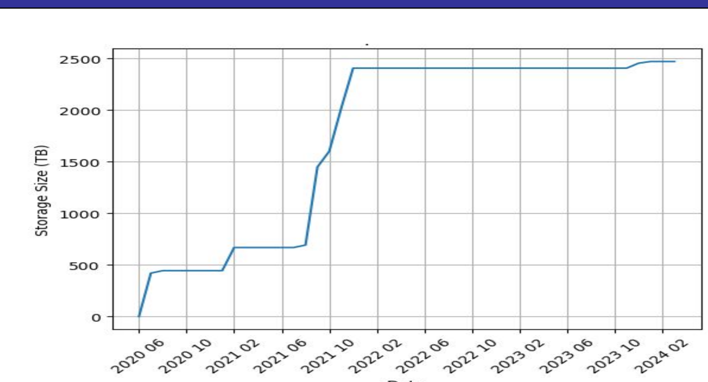

Figure 7: Increase in SoCal cache storage size from June 2020 to Mar 2024. Chicago and Boston storage sizes remain constant since deployment.

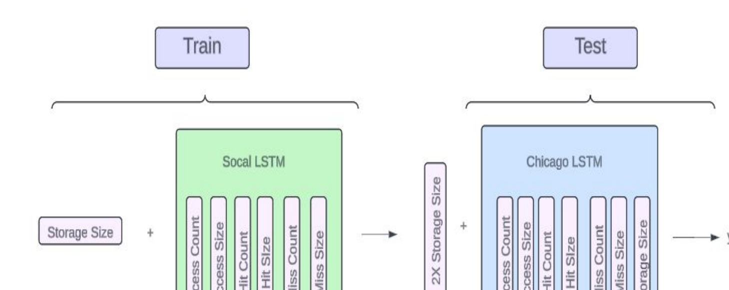

Figure 8: Diagram of doubled storage size prediction methodology. Storage size is added as a feature and then doubled.

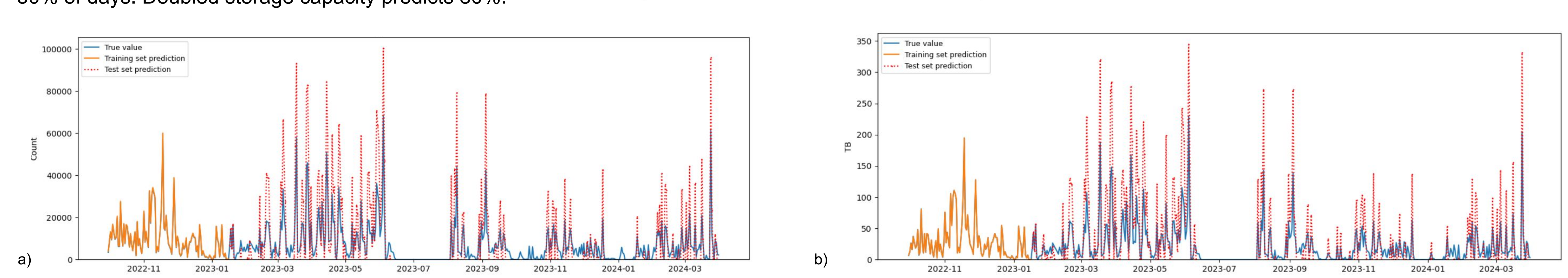

Figure 9: After doubling the storage size, we again predict (red dotted line) the a) daily cache hit counts b) daily cache hit sizes for the Chicago cache with transfer learning.This prediction confirms that the cache hit rate is 80% similar to the analysis shown in Figure 7.

DISCUSSION

- LSTM models based on PyTorch with 128 units, dropout rate 0.04, and tanh activation functions have produced low-error predictions (Fig. 2).
- SoCal LSTM models have better predictions than other caches due to its longer history.
- Experiments were performed with transfer learning (training on SoCal data and fine-tuning in Chicago or Boston data).
- Transfer modelling allows improved predictions with decreased training sizes.
- Chicago cache's total access size 21-day moving sum suggests that increasing its capacity could greatly improve its cache performance (Fig. 6).
- SoCal cache storage size has gradually increased from 420 TB to 2.5 PB from July 2020 to Mar 2024 (Fig. 7), which leads to the cache hit rate to be about 94% during the last year.
- Additional experiments included storage size as a feature in LSTM transfer models to predict cache utilization as Chicago storage size increases.

Table 2: Chicago transfer model with doubled storage size feature predictions

		Increased Avg Cache Hit (Count)	Increased Avg Cache Hit (TB)
Daily	Normal	18257	53.65
	Doubled	33177	96.7
Hourly	Normal	272	0.83
	Doubled	411	0.97

CONCLUSION

- A model trained with one cache could improve predictions about other regional caches.
  - Boston predictions and RMSE error are improved with transfer learning for as small as 15% of fine-tuning sizes.
- Chicago regional cache would see improved performance if storage capacity were increased. When doubling storage size, the Chicago cache predicts (Fig. 9, Tbl. 2)
  - 81% increase for daily hit counts and 80% increase for daily hit sizes.
  - 50% increase for hourly hit counts and 17% increase for hourly hit sizes.
  - Overall byte accommodation of 80%.
- Future work:
  - Model impact of varying storage capacity
  - ex. 3x capacity, 4x capacity

ACKNOWLEDGEMENTS

Thanks to my mentors and collaborators, Alex Sim and John Wu. This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Dept. of Energy under Contract No. DE-AC02-05CH11231, and also used resources of the National Energy Research Scientific Computing Center (NERSC). This work was also supported by the U.S. Dept. of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program.