Fermilab | U.S. DEPARTMENT OF ENERGY | Office of Science

SciDAC-Data, A Project to Enabling Data Driven Modeling of Exascale Computing

L. Aliaga, M. Mubarak, P. Ding, A. Tsaris, A. Lyon, A. Norman, R. Ross

CHEP 2016

11 October 2016

In partnership with:

Argonne NATIONAL LABORATORY

SciDAC-Data Project

Scientific Discovery through Advanced Computing

SciDAC-Data is a specially designed program within the Office of Science of the U.S. Department of Energy:

- *to develop the Scientific Computing Software and Hardware Infrastructure needed to use HPC computers.*

- *to advance DOE research programs in basic energy sciences, biological and environmental research, fusion energy sciences, and high-energy and nuclear physics.*

http://www.scidac.gov

Overview

- As part of the SciDAC-Data project, we are working to analyze the historical information collected by the Fermilab data center over the past two decades
  (See ***https://indico.cern.ch/event/505613/contributions/2227943***).

- The current Intensity Frontier (IF) experiments are producing data 10-12 PB/year.

- HEP experiments rely on efficient performance of data storage, management and analysis.

- Understanding this performance may help HEP experiments to design storage and caching policies.

- Realistic analysis is possible using data-driven analyses and simulations.

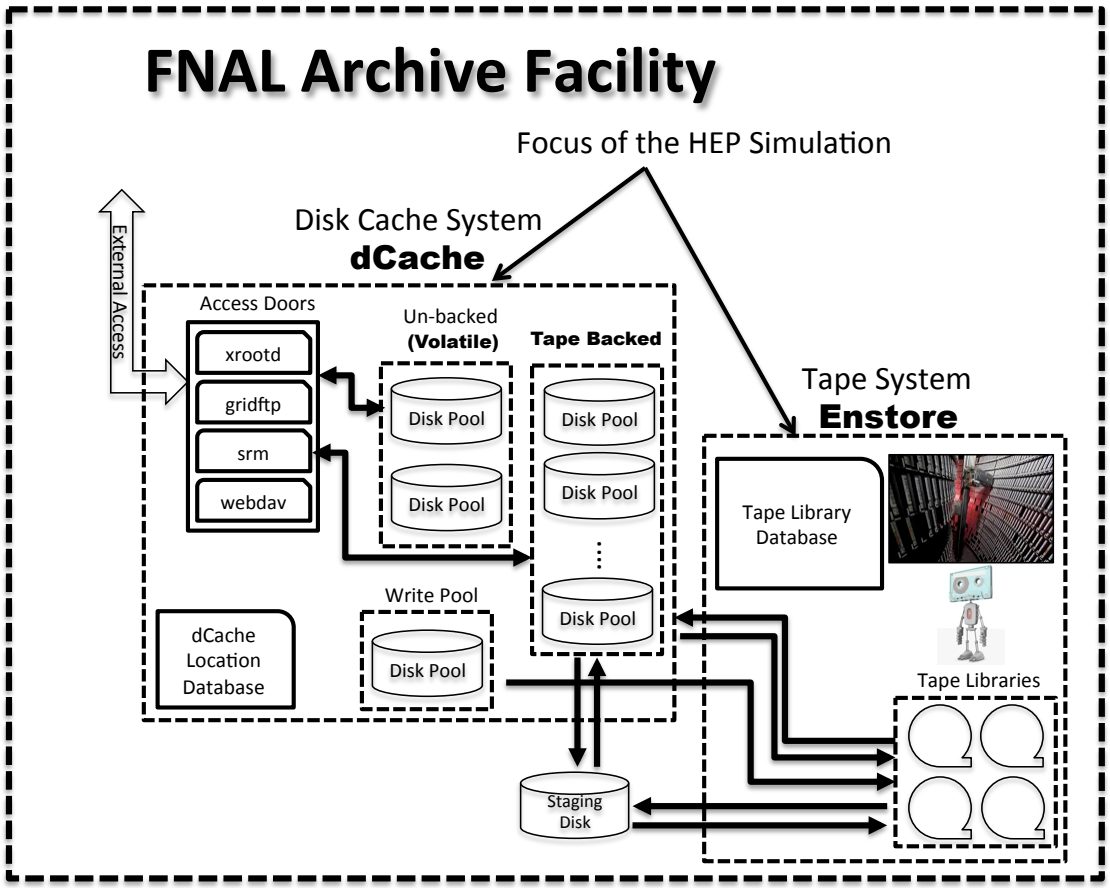
Fermilab Archive Facility

Three layer system:

Tape *system enstore for a long term storage (~374 PB).*

*Accessed via a distributed caching system (**dCache**).*

*Job requests are scheduled by the **SAM** scheduler (Serial Access to Metadata).*



We want to simulate this system…

This Work

- We have created a discrete-event simulation of the Fermilab Archive Facility.
- This simulation uses job activity logs of Fermilab experiments in order to have real data access patters of the Fermilab Archive Facility layers.
- This talk shows results of the NOvA experiment comprised of more than 9K jobs spanning over the past 5 years.

Using real job patterns, we explore the design of the Fermilab storage system and make performance predictions regarding cache size, file lifetime in cache and tape activity.

CODES Simulation Model

- This simulation uses **CODES** framework (***http://www.mcs.anl.gov/projects/codes***) developed by **Argonne** National Laboratory and Rensselaer Polytechnic Institute (**RPI**).

- CODES simulates large-scale HPC workloads and scientific workflows, high-performance network, storage systems and distributed data-intensive systems.

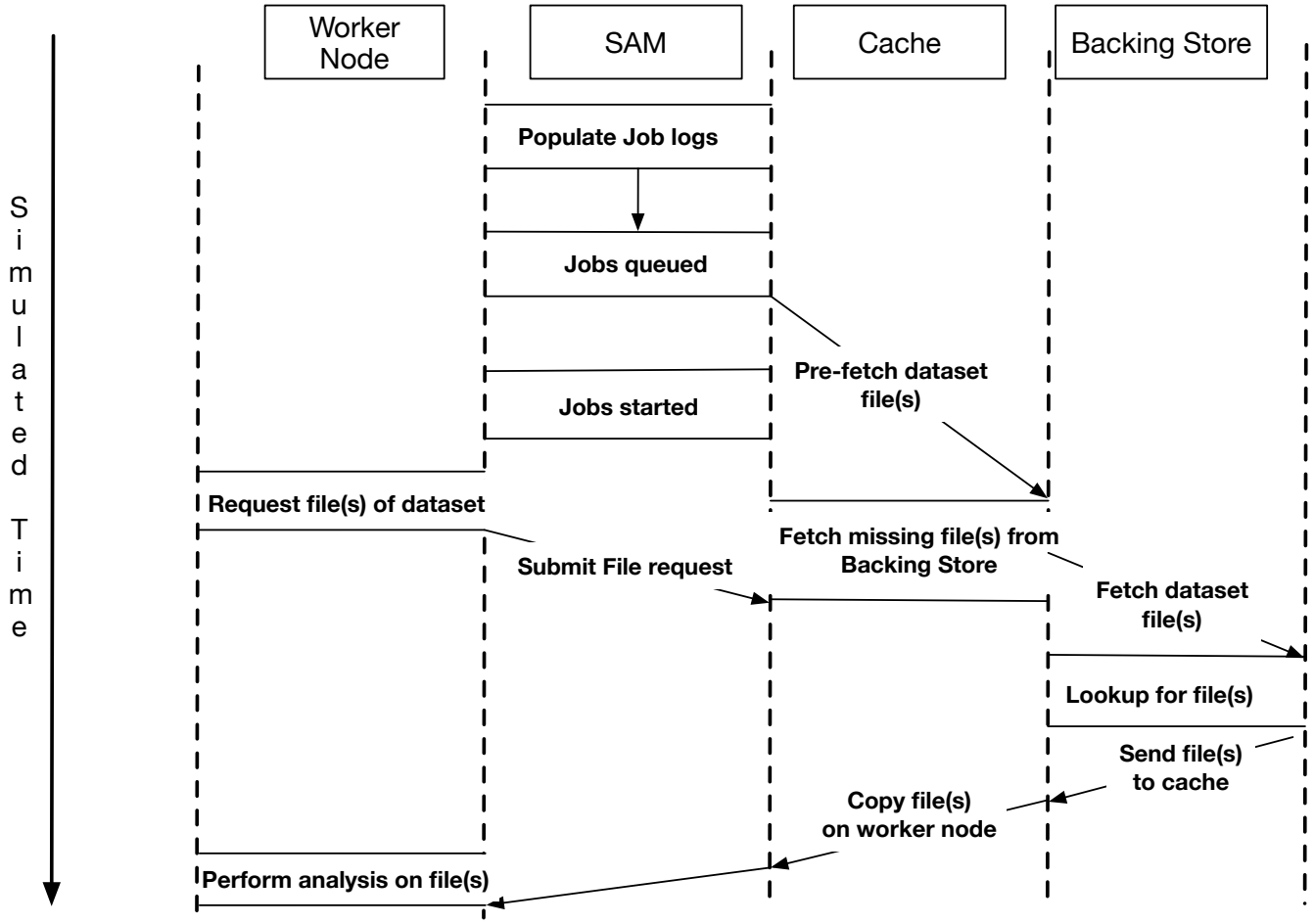
- CODES uses the Rensselaer Optimistic Simulation System, **ROSS (**http://carothersc.github.io/ROSS): a discrete-event simulation framework, which has been shown to process billions of events per second on modern HPC systems.

- Simulations in CODES/ROSS comprise of Logical Processes (LPs) where each LP represents a distinct entity in the system.

- The Logical Processes interact with each other via time-stamped messages or events and they are mapped to physical cores on compute nodes where each core maintain its own local simulation time.

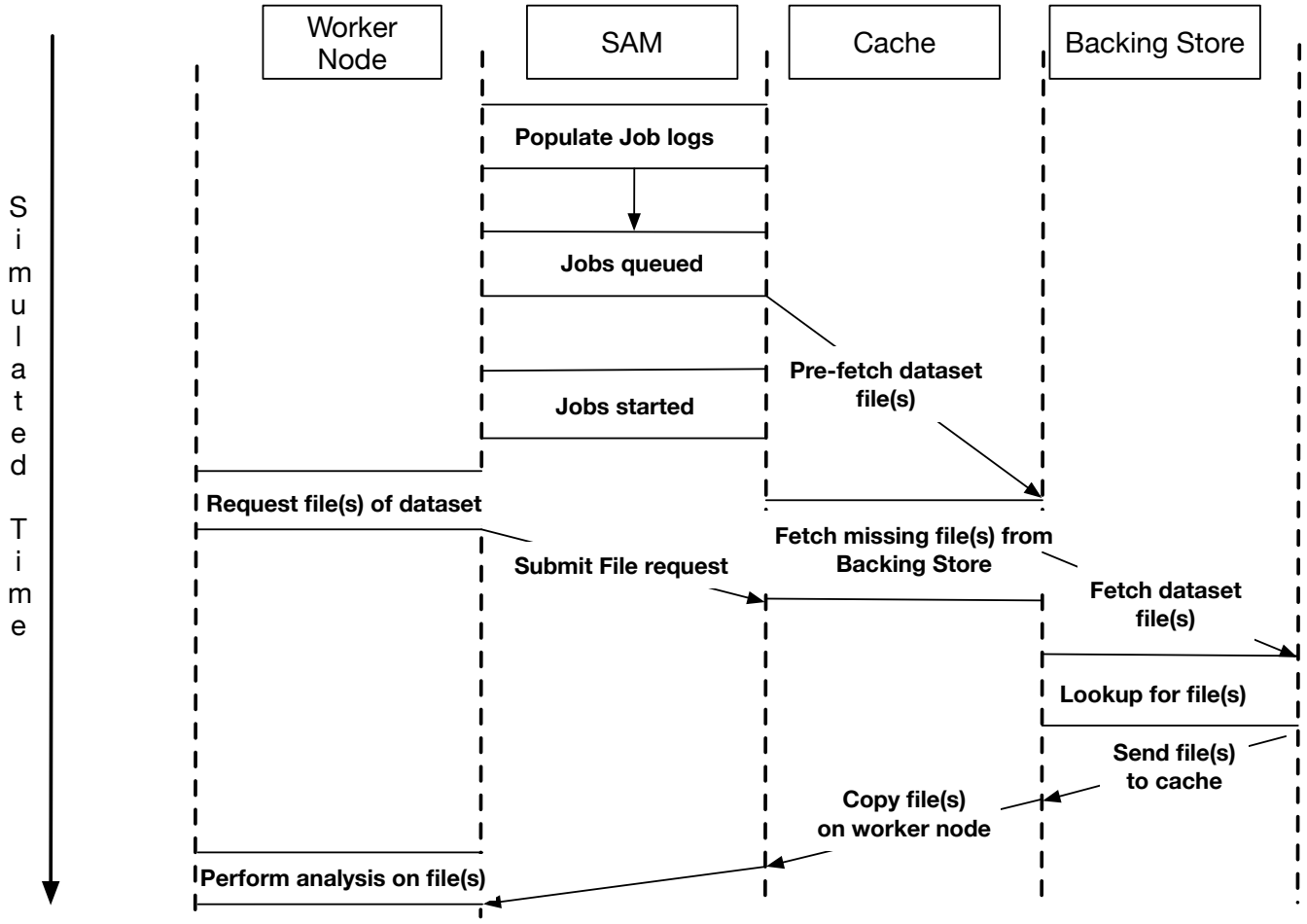
HEP Model in CODES

- **4 Logical Process types:**
  - SAM scheduler.
  - Worker nodes.
  - Cache.
  - Tape.



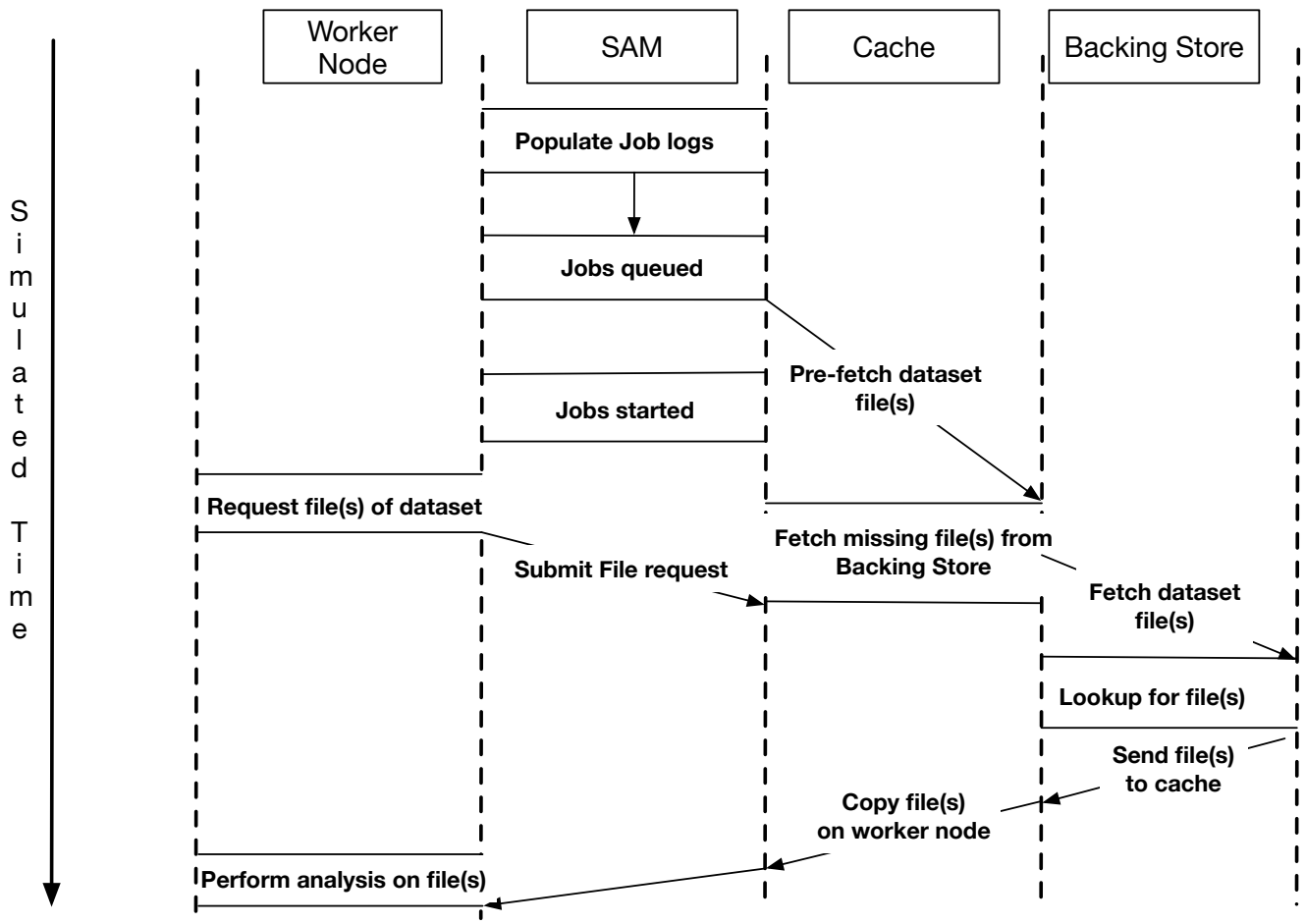
SAM Logical Process

- Replicates the behavior of the SAM storage management and scheduler.
- The simulation keeps a catalog for the files: ID, location (cache or tape), size, etc.



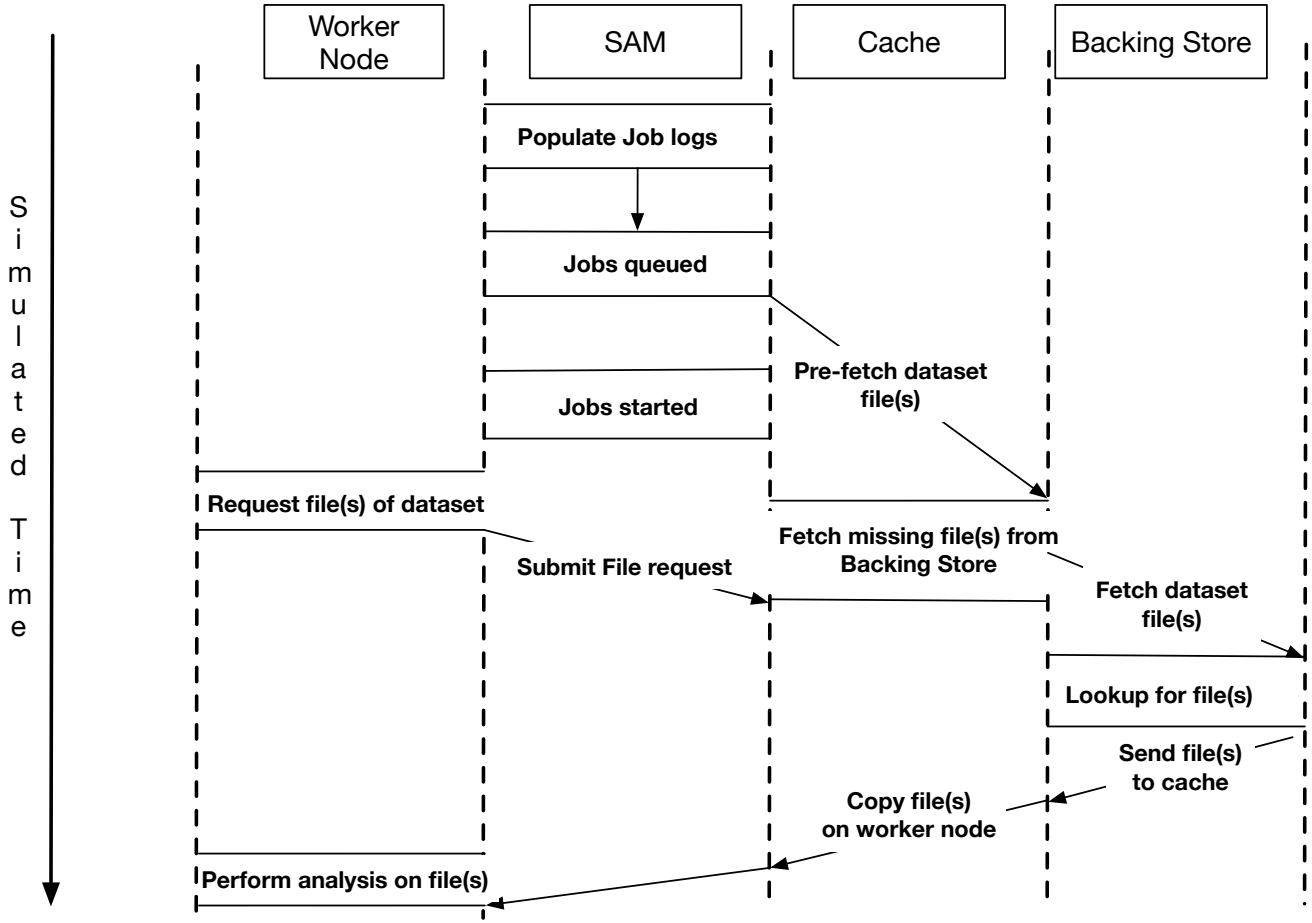
Cache and Tape Logical Processes

- Startup latency and bandwidth cost associated with values based on actual configuration.
- Last Recent Used policy for cache access is used in this simulation.
- Capacity of the cache can be configured in the simulation: allows us to explore the impact of different cache sizes on the workflow.



SAM Working…

- Set of request for accessing files (in cache or tape) belonging to a specific dataset.
- Worker cores are assigned to the job where each worker core processes a set of files from the dataset.
- Files are accessed by SAM prefetching from tape.
- When a worker core finishes processing a file, it would request another file to SAM.
- If the file is not in cache, it requests the file from the tape.



The process continues until all the files have been consumed by the worker cores.

CODES Output

CODES reports important statistics for analysis:

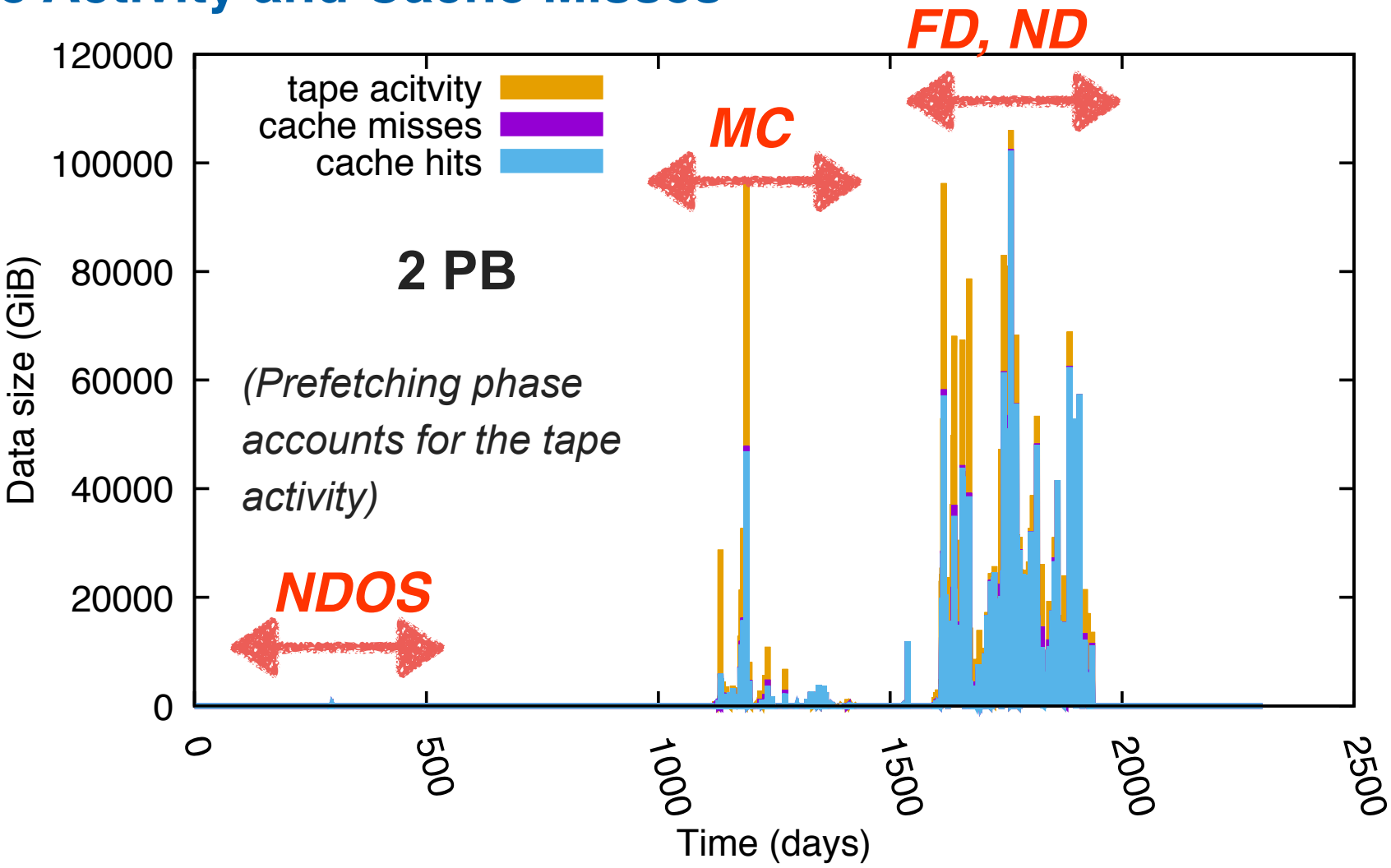
- File eviction times from cache.
- Cache hits and misses.
- Cache utilization over time.
- Start and end times of jobs.
- Tape activity over time.
- Other information…

Simulation Using Real Job Logs from NOvA

Specifications:

- Workload comprises the analysis 9536 different datasets over 5 years that represents 53 K analysis campaigns for NOvA experiment (***https://www-nova.fnal.gov***).
- The bandwidth configuration for cache is 1.25 GiB/s and tape is 200 MB/s.
- 60 K worker nodes.
- The next slides show results from our simulation using **2 PB**, **1 PB** and **20 TB** cache sizes.

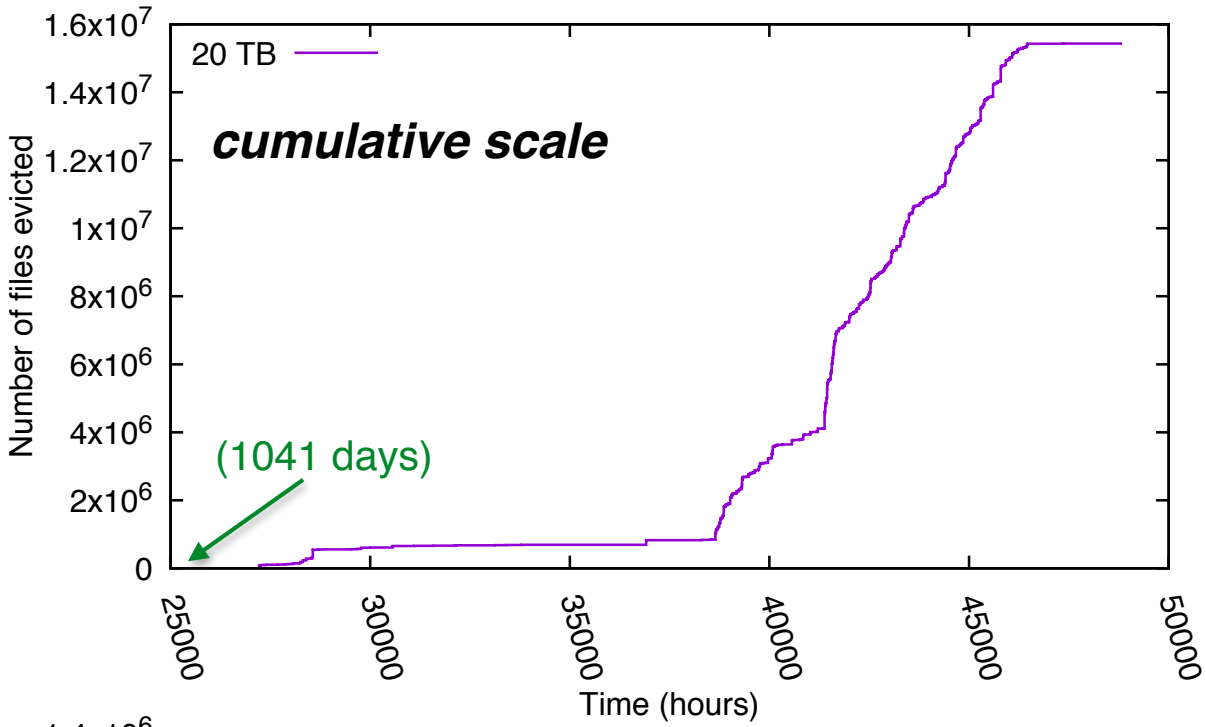
Tape Activity and Cache Misses

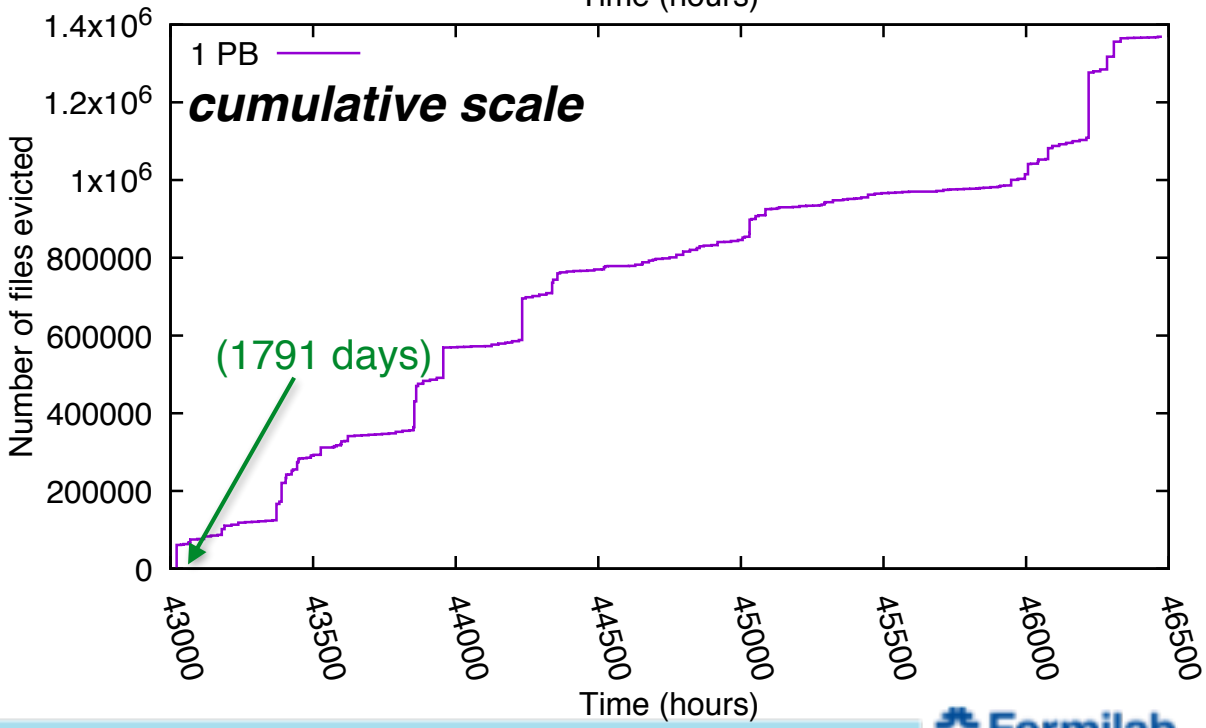


With 2 PB, the cache size is large enough to fit the fetched data, files are not evicted and minimal cache misses are observed.

Impact of Cache Size

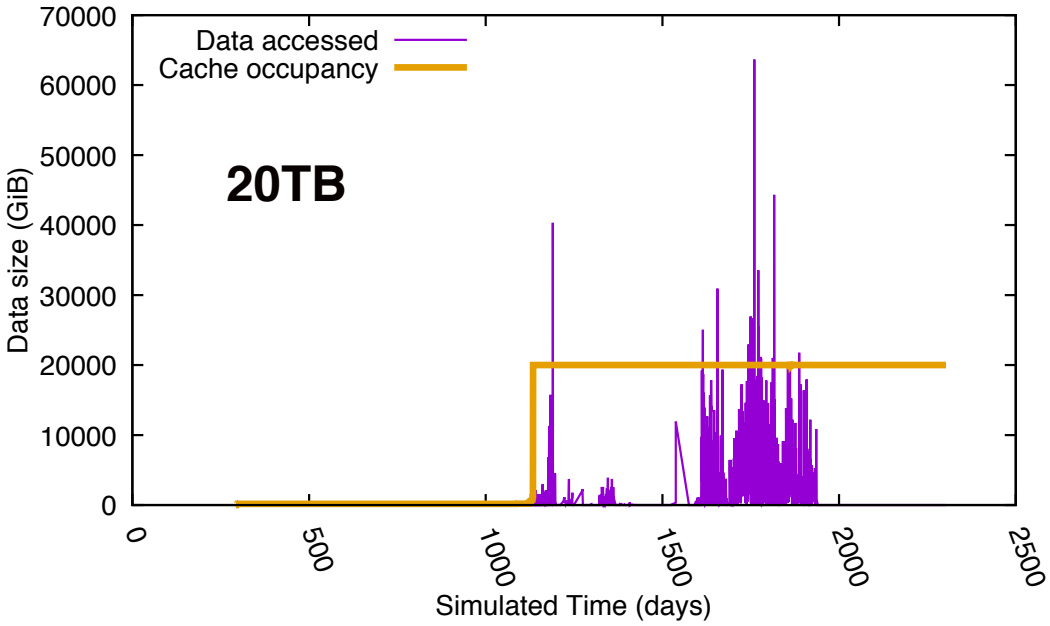
- With 20 TB, we see around 16 M files getting evicted since the cache size is small to handle all the data that is being requested
- With 1 PB, a reasonable number of files getting evicted from the cache.
- With 2 PB, none of the files get evicted.

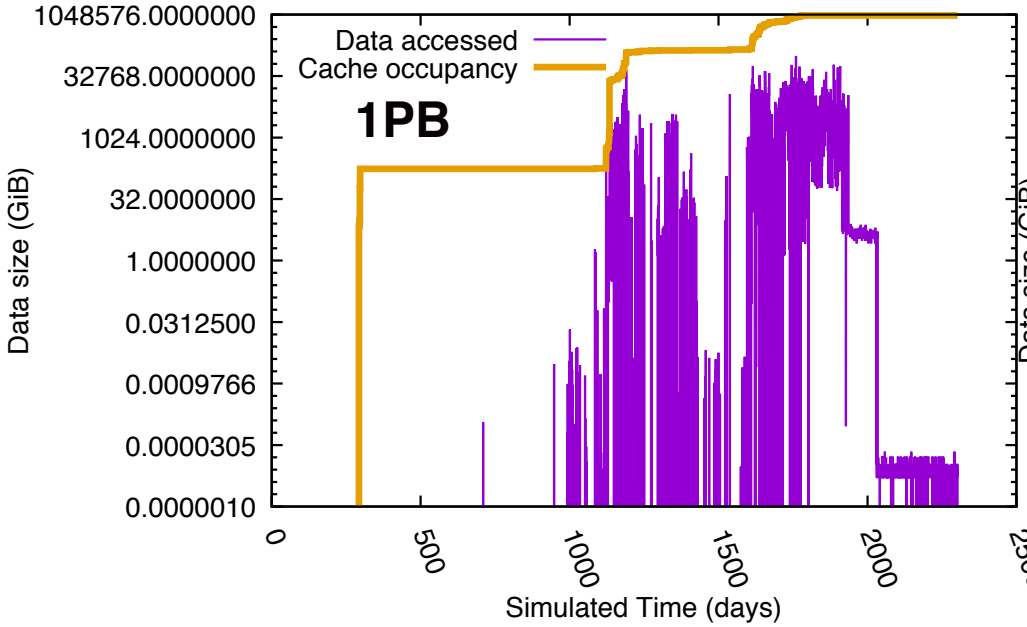


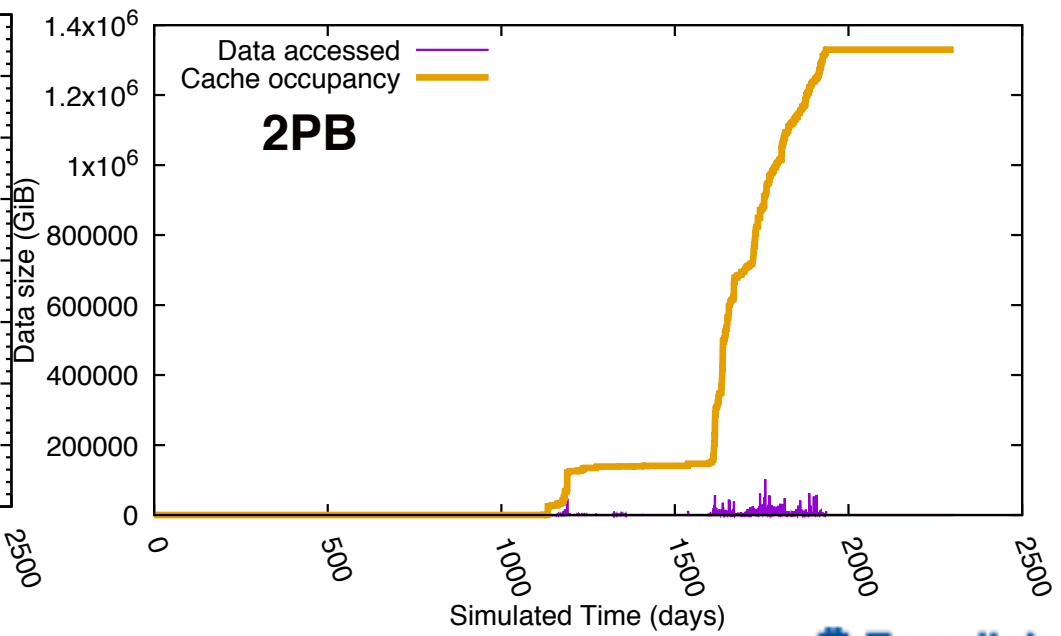


Cache Utilization

- With 1 PB and 20 TB, requests for data access get more than the occupied cache size.
- With 2 PB, requests for data access stay beyond the actual size of the cache.







Next Steps

- Expand our model by experimenting with multiple caching policies and running simulations with **mixed dataset** of NOvA, CDF, D0, MINOS, MicroBooNE, MINERvA and Mu2e experiments.
- Validate our simulation performance predictions with the performance statistics reported by Fermilab.

Extrapolation to HPC Scale

- CODES also has developed detailed network and storage models of current and future extreme-scale architectures.
- CODES provides the HPC interconnect and storage models as pluggable components so that the model developers can plug and play different HPC components with minimal changes to the code.
- CODES will enable us to do performance prediction of the HEP jobs on the next generation extreme-scale systems: parameter tuning of the HEP workloads and adapt them to the HPC environment.

Summary

- Using the CODES simulation framework, we develop an end-to-end simulation of the Fermilab data center that can assist data scientists in effective data handling and cache optimizations.

- An accurate end-to-end simulation can help gain insight into the behavior of these complex systems and answer questions on data storage and cache policies, identify bottlenecks and quantify the value of adding new hardware.

- We use real data from Fermilab that correspond to NOvA. We have designed the simulation to report detailed metrics on the performance of the storage model such as cache lifetime, cache hits and misses, tape activity over time etc.