Diagnostic and Testing of the Ceph Filesystem

FERMILAB-POSTER-23-178-STUDENT

Nehemyah Green | 2023 GEM Fellow | Dr. Kenneth Herner, Dr. Andrew Norman, and Dr. Michael Kirby
Fermilab Accelerator Laboratory: Computing Sector

Abstract

Fermilab's experiments rely on substantial data, often reaching hundreds of petabytes, resulting in expensive storage needs. The current storage infrastructure in place for interactive analysis use incurs significant costs, with each terabyte of storage costing hundreds of dollars. Consequently, a transition to **Ceph** is underway. This poster will explain the advantages of **Ceph** and why it's the preferred choice for the laboratory.

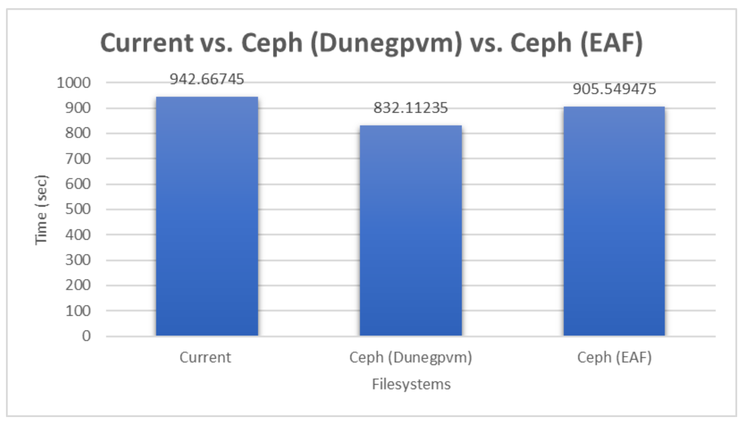
Background

- The goal is to move to a design that uses advanced analysis methods and machine learning techniques.
- Researching and developing new technologies that are cost-effective solutions for the end users' analysis.
- The **Deep Underground Neutrino Experiment (DUNE)** and **Fermilab Elastic Analysis Facility (EAF)** will be used to test **Ceph**.
  - **DUNE** is a state-of-the-art neutrino detector. The goal of this detector is to understand the neutrino.
  - The objective is to test the performance of the new file system on **DUNE**, a data-intensive experiment at Fermilab. Demonstrating its seamless operation with **DUNE** provides confidence that it will be effective across all of Fermilab's experiments. The **DUNE** the machine will be used is dunegpvm15. The machine is a four-CPU virtual machine for interactive end-user analysis
  - **EAF** is a facility that is built around production. It is meant to reduce the size of datasets for use by collaboration in the analysis.

Methods

- Diagnostics and measurements were done by using **Darshan.**
  - **Darshan** is a tool that is designed to obtain a realistic picture of application I/O behavior with the least amount of overhead, including characteristics like patterns of access inside files.
- The code was run four times simultaneously and took the average of the results. This simulates a more realistic load inside the machines.

Results



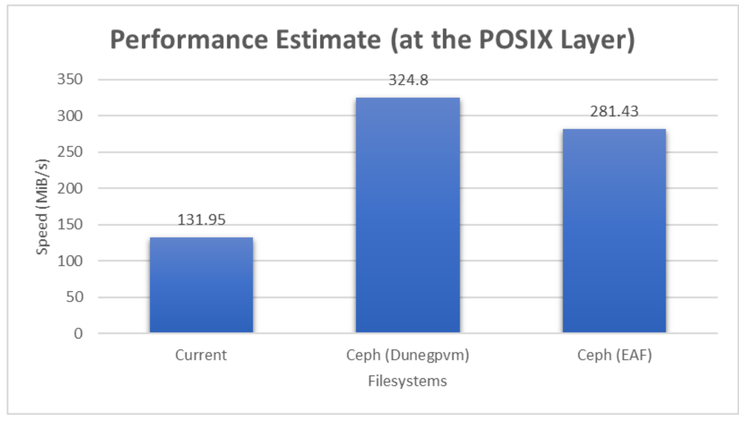


Fig 1. & Fig 2. Both figures compare the performance of the filesystems. The first figure compares the time it the program to run. The second compares the speed of the files moved.

Run Times Compared		
Current (sec)	Ceph (Dunegpvm) (sec)	Ceph (EAF) (sec)
942.6677	832.3432	904.7543
942.6687	832.0349	907.2664
942.6687	832.0371	904.6783
942.6647	832.0342	905.4989
942.66745	**832.11235**	**905.549475**

Table 1. Results of Figure 1 in tabular form.

Performance estimate (at the POSIX Layer)					
Current		Ceph (Dunegpvm)		Ceph (EAF)	
Transferred (MiB)	Speed (MiB/s)	Transferred (MiB)	Speed (MiB/s)	Transferred (MiB)	Speed (MiB/s)
21365.1	134.25	21365.1	325.68	21365.1	283.72
21365.1	133.18	21365.1	335.12	21365.1	281.31
21365.1	130.96	21365.1	322.56	21365.1	278.29
21365.1	129.41	21365.1	315.84	21365.1	282.40
21365.1	**131.95**	**21365.1**	**324.80**	**21365.1**	**281.43**

Table 2. Results of Figure 2 in tabular form, along with data transfer totals.

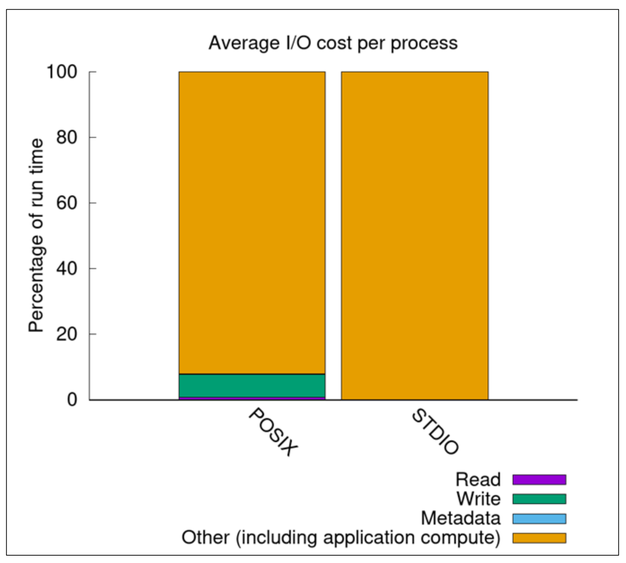


Fig 3. The percentage of time the machine spent on reading, writing, Metadata, and computing for the Ceph Filesystem.

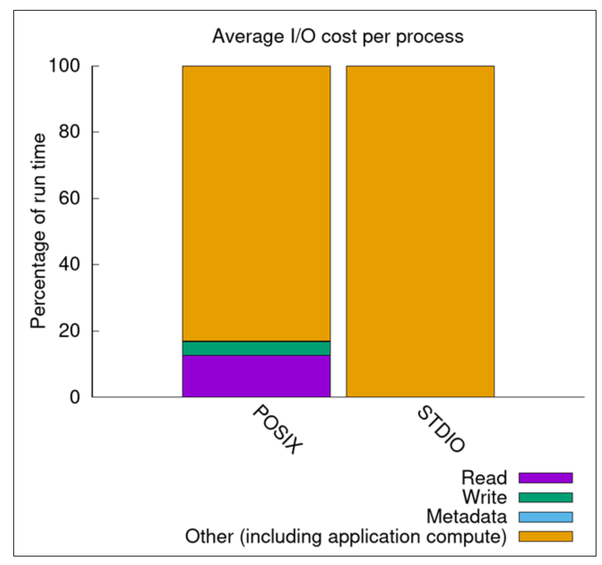
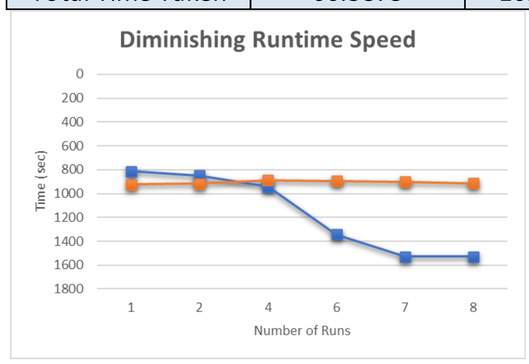


Fig 4. The percentage of time the machine spent on reading, writing, Metadata, and computing for the current Filesystem.

Reading and Writing		
	Ceph (Dunegpvm)	Current
Run Time	832.3432	942.6677
% of Run Time	8%	18%
Total Time Taken	66.5875	169.6802

Table 3. Compares the percentage of read and write time together of the two filesystems.



# of Runs	Current (sec)	Ceph (EAF) (sec)
1	813	925
2	847	918
4	943	889
6	1344	895
7	1529	902
8	1529	914

Fig 5. & Table 4. The figure compares Ceph (EAF) and the Current system. It shows how much the runtime diminishes vs the number of runs. The table shows the results in tabular form.

Conclusion & Discussion

- The results clearly show that the **Ceph** filesystem is more than capable of replacing the current one.
- **Ceph** performance speed was more than twice as fast as the current infrastructure that we are using.
- **Ceph** percentage of read and write time was less than the current system by about 40%.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.