Power Patterns: Understanding the Energy Dynamics of I/O for Parallel Storage Configurations

Maya Purohit
Colby College
Waterville, Maine, USA
mmpuro26@colby.edu

Valerie Hayot-Sasson (advisor)
University of Chicago
Chicago, Illinois, USA
vhayot@uchicago.edu

Kyle Chard (advisor)
University of Chicago
Chicago, Illinois, USA
chard@uchicago.edu

Abstract

As HPC applications become more I/O intensive, understanding their power consumption patterns is necessary to develop energy-saving solutions. Here, we evaluate the energy consumption of I/O operations on two popular HPC parallel file systems: Lustre and DAOS. We develop models to predict the energy usage of sequential writes and evaluate their accuracy against our gathered benchmarks. Our models can be used to enhance the accuracy of energy-predicting frameworks by allowing them to consider storage configuration when estimating total energy consumption.

1 Introduction

HPC workloads have significant I/O that is often neglected when considering energy consumption. While others have modeled energy usage of I/O for scientific computing [2], energy consumption estimates are typically based on computation and do not consider the impacts of network transfers and data storage [3]. We expand prior work in energy-prediction to consider the total energy consumption of task-related I/O on parallel file systems by investigating the energy usage of Lustre and DAOS.

2 Testbed

We deploy Lustre and DAOS on bare-metal nodes in the Chameleon Cloud [4]. As seen in Table 1, our Lustre[1] deployment has one Client Node, one Metadata Server/Target (MDS/T) with one 223.6GB SSD and one Object Storage Server/Target (OSS/T) with 16 2TB SSDs. For DAOS[5], we deploy one Client/Admin Node with one 447.1GB SSD and two DAOS Storage Server nodes with 15TB of NVMe storage.

Device	CPU Model	Storage Model	Active Storage Power (W)	Idle Power (W)
Lustre Client	Intel(R) Xeon Gold 6126	DELL 28F3R	2.2	207.9
Lustre MDS/T	Intel(R) Xeon Gold 6126	DELL 28F3R	2.2	207.9
Lustre OSS/T	Intel(R) Xeon CPU E5-2650	Seagate Enterprise Capacity HDD	6.02	112.3
DAOS Client/Admin	Intel(R) Xeon Gold 6240R	DELL VPP5P	2.4	256.5
DAOS MD/S	AMD EPYC 7352	Express Flash NVMe	7.6	245.1
DAOS Storage Server	AMD EPYC 7352	Express Flash NVMe	7.6	245.1

Table 1: Testbed hardware description

3 Measurements

We measured total, CPU and RAM power consumption for each node in both file systems by using the Running Average Power Limit (RAPL) interface on Intel machines and the Intelligence Platform Management Interface (IPMI) from the Baseboard Management Controller (BMC). We conducted experiments writing/reading files ranging in size from 1B to 10GB to/from the disk and memory of the two client nodes and to the Lustre and DAOS file systems. We measured power at one second intervals and performed 3 repetitions. The TCP network protocol was used for all experiments.

4 Analysis of Energy Consumption

Figures 1 and 2 compare the total energy consumption for writing/reading to/from local disk, memory, and Lustre or DAOS. We perform pairwise-t-tests to determine the significance of the differences in energy expenditure. Further, we isolate the impact of I/O by subtracting the idle power of each machine from the power metrics.

4.1 Energy consumption of I/O for Lustre

Figure 1 shows that the total energy consumption of Lustre exceeds the energy expended when writing and reading locally to disk and memory. Specifically, our results reveal that Lustre expends approximately 1.9x as much as energy as disk (p = 0.0068) and memory ($p = 3.6x10^{-5}$) when writing and as much as 30x more energy compared to disk (p = 0.0021) and memory (p = 0.0001) when reading. As our configuration includes one client node, OSS/T, and MDS/T, we could not take full advantage of Lustre's parallel capabilities, which may reduce its performance.

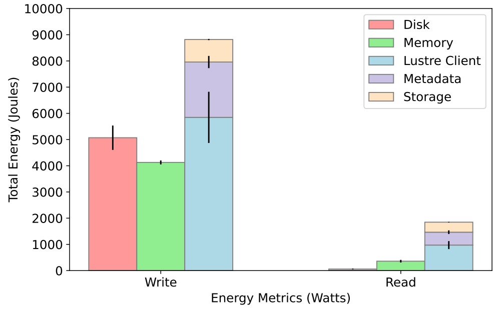

Figure 1: Total Energy Consumption Analysis of Writing/Reading to/from Disk, Memory and all Lustre Machines

4.2 Energy consumption of I/O for DAOS

Figure 2 shows that the energy expended when writing to DAOS is comparable to memory (p = 0.88). This may be due to kernel-bypassing, particularly on the client node, through the DFuse daemon. Additionally, DAOS binds NVMe storage to the SPDK library and SCM storage to the PMDK library, allowing the file system to have access to storage directly through user space and increasing energy efficiency.

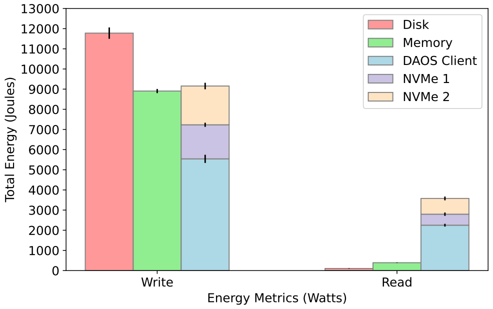

Figure 2: Total Energy Consumption (J) Analysis of Writing/Reading to/from Disk, Memory and all DAOS Machines

Interestingly, DAOS does not perform as well as memory (p = $1.3x10^{-7}$) or disk (p = $7.9x10^{-8}$) for read operations. Our results show that disk is the most optimal in this case by a factor of 10 and 30 compared to memory and DAOS, respectively. In contrast to disk, I/O task duration is short but the increase above idle power is large for memory and DAOS. This may be a result of network transfers between nodes in DAOS or inaccurate IPMI reportings due to small file sizes.

We note that our recorded idle energy for Lustre (528.1 W) is less than the total idle energy for DAOS (746.7 W) by approximately 218.6 W.

5 Energy Model

We define the total energy consumption for I/O in a parallel file system to be:

$$\text{Total Energy of I/O} = E_{client} + E_{metadata} + E_{storage} \tag{1}$$

Where each component is defined as follows:

$$E_{client} = \frac{\text{file size}}{\text{storage bandwidth}} \cdot (P_{\text{Storage}} + P_{\text{NIC}}) \tag{2}$$

$$E_{metadata} = \frac{\text{metadata size}}{\text{network bandwidth}} \cdot (P_{\text{Storage}} + P_{\text{NIC}}) \cdot \text{\# files} \tag{3}$$

$$E_{storage} = \frac{\text{file size}}{\text{network bandwidth}} \cdot (P_{\text{Storage}} + P_{\text{NIC}}) \tag{4}$$

*Since DAOS stores metadata in SCM storage and larger data in NVMe storage, we model the energy consumption of each DAOS Storage Server as the sum of $E_{storage}$ and $E_{metadata}$.

5.1 Model Validation

Figures 3 and 4 compare the total energy consumption predicted by our models for writing with our experimentally observed values. In order to isolate the impact of I/O, we subtracted the average power consumption measured when performing computation from power measures taken when performing both I/O and computation.

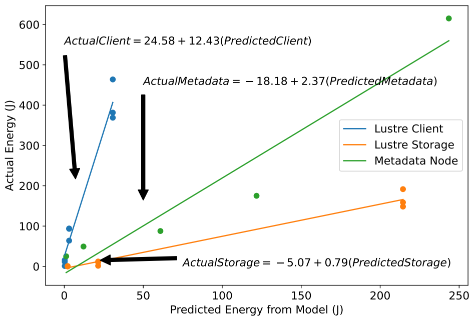

Figure 3: Energy relationship between model prediction and observed energy (J) of I/O for Lustre Nodes

Figure 3 shows the relationship between the observed and predicted energy values for the client, metadata, and storage nodes of Lustre. The values predicted by our models E_{client} and $E_{metadata}$ typically underestimate the total energy expenditure for each node by a factor of 12 and 2.37, respectively. However, our model for $E_{storage}$ overestimates energy expenditure by a factor of 0.79.

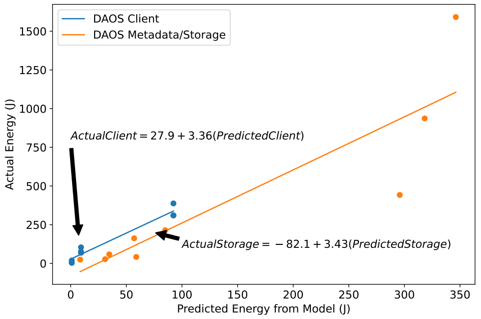

Figure 4: Energy relationship between model prediction and observed energy (J) of I/O for DAOS Nodes

Figure 4 shows the relationship between the observed and predicted energy values for the client and storage servers of DAOS. Both models for E_{client} and $E_{storage}$ underestimate total energy expenditure by approximately a factor of 3.

Since our predictions currently over/underestimate our observed energy consumption values, we may need to expand our model to include additional energy contributors.

6 Conclusions

Our results show that while Lustre always expends more energy than local storage, DAOS is comparable in performance to memory when writing. We have developed a model to predict the I/O-related energy consumption of a given task for parallel file systems. However, it currently over/underestimates the observed energy usage of each node. We will continue to explore ways to expand the model to consider additional energy contributors and repeat experiments using a power meter to obtain more accurate measurements.

References

[1] Peter Braam and Philip Schwan. 2002. Lustre: The intergalactic file system. (01 2002).

[2] Rafael Ferreira da Silva, Henri Casanova, Anne-Cécile Orgerie, Ryan Tanaka, Ewa Deelman, and Frédéric Suter. 2020. Characterizing, Modeling, and Accurately Simulating Power and Energy Consumption of I/O-intensive Scientific Workflows. *Journal of Computational Science* 44 (2020), 101157. https://doi.org/10.1016/j.jocs.2020.101157

[3] Alok Kamatar, Valerie Hayot-Sasson, Yadu Babuji, Andre Bauer, Gourav Rattihalli, Ninad Hogade, Dejan Milojicic, Kyle Chard, and Ian Foster. 2024. GreenFaaS: Maximizing Energy Efficiency of HPC Workloads with FaaS. arXiv:2406.17710 [cs.DC] https://arxiv.org/abs/2406.17710

[4] Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons Learned from the Chameleon Testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

[5] Jay Lofstead, Ivo Jimenez, Carlos Maltzahn, Quincey Koziol, John Bent, and Eric Barton. 2016. DAOS and Friends: A Proposal for an Exascale Storage System. In *SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 585–596. https://doi.org/10.1109/SC.2016.49