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

Maya Purohit [1] Valerie Hayot-Sasson (Advisor)[2] Kyle Chard (advisor)[2]

[1]Colby College [2]University of Chicago

Introduction

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 describe the energy usage of sequential writes and evaluate their accuracy against our gathered benchmarks. Our models can be utilized to enhance the accuracy of energy-predicting frameworks by allowing them to consider storage configuration when estimating total energy consumption.

Research Methods

While others have modeled energy usage of I/O for scientific computing [1], energy consumption estimates are typically based on computation and do not consider the impacts of network transfers and data storage [2].

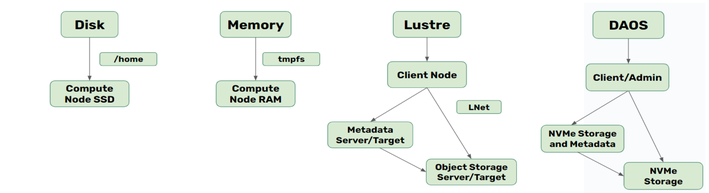

Figure 1. Visual Representation of our Disk, Memory, Lustre and DAOS Configurations

Measurements

We list the three tools used to gather power metrics for each node:

IPMI Power (FreeIPMI Tool)
CPU Power (Powerstat using RAPL Interface)
RAM Power (Turbostat using RAPL Interface)

Experimental Design

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 repeat each write and read operation (flushed cache) three times.

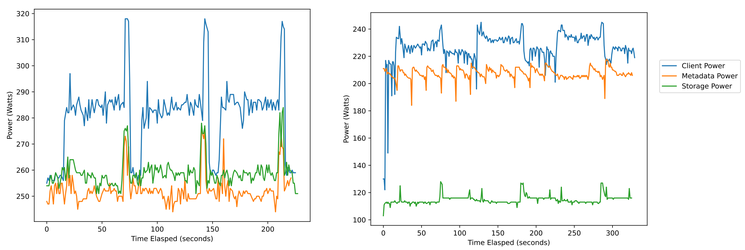

Figure 2. Total Power Consumption of Each Node in DAOS (left) and Lustre (right) for Write Operation

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

Table 1. Testbed hardware description

Total Energy Consumption of Lustre

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. Lustre's poor energy performance is likely due to TCP network transfers between nodes and limited parallelism due to small file sizes.

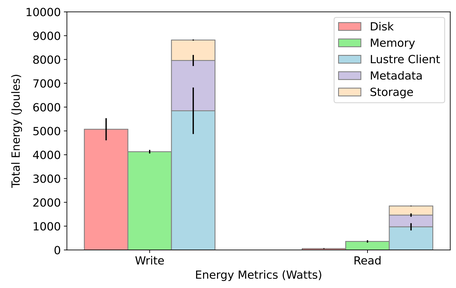

Figure 3. Comparison of Total Energy Consumed for I/O Operations on Disk, Memory, and Lustre

Total Energy Consumption of DAOS

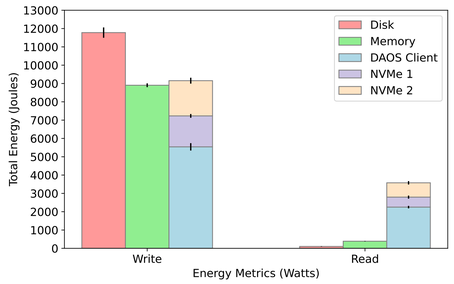

Figure 4. Comparison of Total Energy Consumed for I/O Operations on Disk, Memory, and DAOS

DAOS is comparable to memory ($p = 0.88$) in energy performance when writing, likely due to its use of energy-efficient hardware (NVMe and SCM Storage) and kernel bypassing. DAOS does not perform as well as memory ($p = 1.3x10^{-7}$) or disk ($p = 7.9x10^{-8}$) for read operations. The DAOS client experiences a larger increase above idle power than disk and memory.

Total Energy Modelling

Models

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

Where each component is defined as follows:

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

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

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

Model Design

Components from datasheet:

1. Active Storage Power (W): P_{Storage}
2. Active NIC Power (W): P_{NIC}
3. Bandwidth (GB/s)
4. # files: # write operations

Using our understanding of Lustre and DAOS, we propose a model in which the energy consumption of the client and storage nodes is impacted by file size and the energy consumption of the metadata node is impacted by the number of I/O operations. We isolate the impact of I/O from our observed values when testing accuracy.

Model Accuracy for Parallel File Systems

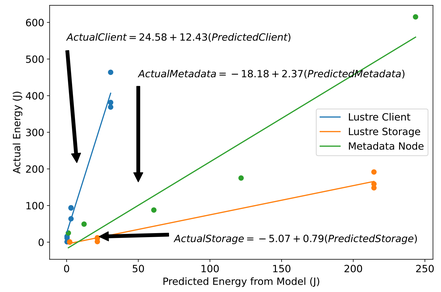

Figure 5. Energy relationship between model prediction and observed energy (J) of I/O for Lustre Nodes

The values predicted by our models E_{client} ($R^2 = 0.9689$) and $E_{metadata}$ ($R^2 = 0.9295$) typically underestimate the total energy expenditure for each node by a factor of 12 and 2.37, respectively. Our model for $E_{storage}$ ($R^2 = 0.9732$) overestimates energy expenditure by a factor of 0.79 (large P_{Storage}). There is a strong linear relationship between predicted and observed values.

Both models for E_{client} ($R^2 = 0.9689$) and $E_{storage}$ ($R^2 = 0.9734$) underestimate total energy expenditure by a factor of ~3 for the DAOS Client and NVMe storage nodes. This indicates that there may be other components in each node that also increase energy expenditure as file size and number of I/O operations increase. We must expand our model to incorporate additional factors.

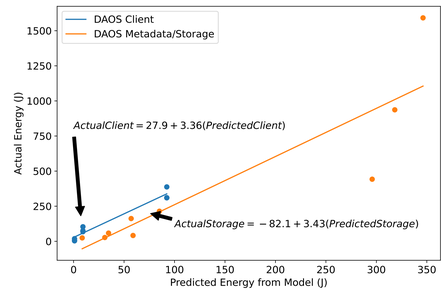

Figure 6. Energy relationship between model prediction and observed energy (J) of I/O for DAOS Nodes

Conclusions and Future Work

Lustre always expends more energy than local storage but DAOS is comparable in performance to memory when writing due to its utilization of energy-efficient hardware.

Currently, our models over/underestimate the total power consumption of each node in the parallel file systems. We would like to expand them to account for other factors that contribute energy.

We would like use the same hardware for Lustre and DAOS to make more direct comparisons between them regarding energy expenditure.

We would like to explore differences in energy when using Infiniband as opposed to Ethernet.

References

[1] 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

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