2023 IEEE International Conference on Cluster Computing Workshops (CLUSTER Workshops) | 979-8-3503-7062-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/CLUSTERWorkshops61457.2023.00017

I/O characterization and performance evaluation of large-scale storage architectures for heterogeneous workloads

Olga Kogiou
Florida State University
Tallahassee, FL, USA
ok22b@fsu.edu

Hariharan Devarajan
Lawrence Livermore National Laboratory
Livermore, CA, USA
hariharandev1@llnl.gov

Chen Wang
Lawrence Livermore National Laboratory
Livermore, CA, USA
wang116@llnl.gov

Weikuan Yu
Florida State University
Tallahassee, FL, USA
wy3@fsu.edu

Kathryn Mohror
Lawrence Livermore National Laboratory
Livermore, CA, USA
mohror1@llnl.gov

***Abstract*— HPC systems traditionally supported compute-centric workloads. However, the increasing reliance on data has led to a shift towards data-dependent workloads. This transition has necessitated storage technologies that enable fast data sharing among workflow parts, but diverse I/O requirements demand tailored solutions. New HPC architectures incorporate specialized software layers like Datawarp, IME, and VAST. However, user-driven storage system selection may lead to improper choices. Our investigation compares VAST with GPFS and Lustre filesystems across multiple machines, measuring performance, scalability, and identifying suitable I/O behaviors. This work provides a guide for selecting the appropriate storage system to optimize data access based on user requirements.**

***Keywords*— VAST, IOR benchmark, HPC applications**

I. Introduction

In the last decade, the increasing reliance on data has led to a shift towards data-dependent workloads. Different parts of the workflow have diverse I/O requirements which has necessitated tailored solutions to these specific needs. However, the selection of the proper technology i.e. storage subsystem is left to the user which can result in improper decisions for a particular workload. In this work we aim to profile VAST storage system [1] by exploring its performance characteristics and scalability as compared to GPFS [5] and Lustre [4] on multiple machines such as Lassen, Quartz and Ruby LLNL machines [6].

II. Test Methodology

A. VAST on Lassen

VAST architecture on Lassen consists of 10 DNodes and 16 CNodes, as shown on Figure 1. DNodes host the file system's SCM and SSDs using NVMe over Fabrics (NVMeoF), while CNodes bridge the client-facing network and VAST's internal NVMe fabric and are Network File System (NFS) Servers. The data path involves synchronous replication to SCM and organizing writes into multi-gigabyte stripes based on similarity and lifetime and is performed by the CNodes. Filled stripes are compressed and migrated to QLC [1] drives asynchronously. The performance analysis is focused on the NFS as well as the protocol used.

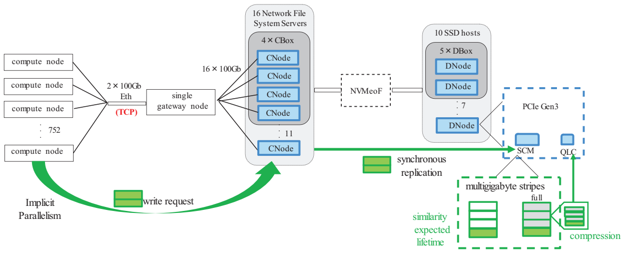


Fig. 1. High-level VAST architecture on Lassen and Implicit Parallelism scheme.

B. Performance Datasets

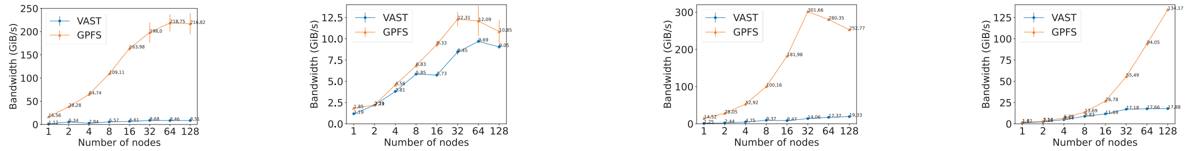
To test the performance of the storage subsystems, we measured the bandwidth using different workloads generated with the IOR benchmark [2]. All tests were performed using file-per-processor pattern [3] and POSIX API since this is a simple yet common pattern and can be used when initially testing storage systems. For our tests we kept the I/O size large enough to ensure that the cache size is exceeded where our segments were 3000 and our block and transfer sizes are 1 MiB.

III. Experimental Results and Key Observations

A. Different Access Patterns

Our first test was conducted on Lassen machine where we scaled 40-process nodes to 128 and tested the subsystems using sequential and random workloads. As shown in Figure 2, VAST can handle random accesses comparably to GPFS. That is because for random patterns the cache of GPFS gets thrashed more and therefore it can be avoided. Unlikely, cache usage is not avoided for sequential accesses and therefore GPFS seems to be able to perform and scale better. This is an expected GPFS behavior, due to its multiple levels of disks and caches with a total storage capacity of 24 PB. GPFS also utilizes the same Mellanox InfiniBand fabric as the compute nodes which enables

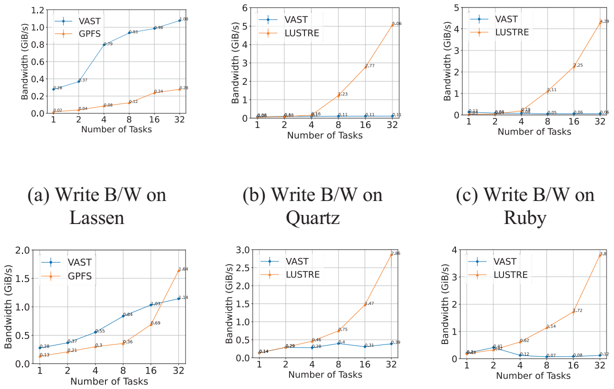
979-8-3503-7062-1/23/$31.00 ©2023 IEEE
DOI 10.1109/CLUSTERWorkshops61457.2023.00017



(a) Sequential Write B/W (b) Random Write B/W (c) Sequential Read B/W (d) Random Read B/W

Fig. 2. Write and Read B/W for node scaling test on Lassen for Sequential and Random Accesses.

high- bandwidth, low-latency storage access without the need for Ethernet routing. VAST on the other hand, reaches its global maximum on 32 nodes, possibly due to a network bottleneck from the use of NFS and TCP link with the client side.



(a) Write B/W on Lassen (b) Write B/W on Quartz (c) Write B/W on Ruby

(d) Read B/W on Lassen (e) Read B/W on Quartz (f) Read B/W on Ruby

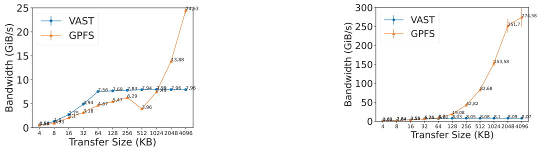
Fig. 3. Write and Read B/W for single node test on Lassen, Quartz and Ruby.

B. *Understanding the Raw Performance of Storage Systems on Different Machines*

Scaling the processes to 32, we conducted a single node test on Lassen, Quartz and Ruby with the use of write fsync. This system call synchronizes the write requests with the rates of files present in the backbone storage of the subsystem and can provide a better insight into its raw performance. As shown in Figure 3, VAST's performance is best on Lassen where there is a 2×100Gb Ethernet on a single gateway node TCP protocol connection which allows VAST to leverage from the system's topology and outperform GPFS. On the other hand, there seems to be an early saturation of VAST on Quartz and Ruby probably due to the use of NFS along with the TCP link with the compute nodes which is 1×40Gb Ethernet on 8 gateway nodes on Ruby and only 2×1Gb on 32 gateway nodes on Quartz. Lustre demonstrates the ability to scale exponentially which is expected since its storage capacity is of 20 TB and it is configured similarly to GPFS using the Infiniband of the compute nodes on Ruby and Quartz.

C. *Various Transfer sizes*

In our final test we test the handling of various transfer sizes on Lassen having 32 nodes and 40 processes per node. From the results shown in Figure 4 it can be concluded that VAST shows the ability to deal with smaller transfer sizes better than with larger ones. For sizes larger than 64 KiB, GPFS seems to be able to scale and give enough bandwidth while VAST appears to be saturated due to the network bottleneck. For smaller transfer sizes however, VAST and GPFS perform comparable which highlights VAST's ability to deal with small I/O requests.



(a) Write B/W (b) Read B/W

Fig. 4. Write and Read B/W for various transfer sizes on Lassen

IV. Conclusion and Future Plans

In conclusion, VAST could be selected as an alternative to other well-used HPC storage subsystems such as GPFS and Lustre when dealing with smaller and random workloads. However, due to the use of NFS over TCP on Lassen, Quartz, and Ruby, VAST saturates early when dealing with larger I/O requests and therefore it is not possible to test its full abilities. A possible remedy could be the exploration of other modes of VAST such as the use of multipath and RDMA. Our future plans include the testing of VAST with metadata and deep learning workloads.

Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under the DOE Early Career Research Program (LLNL-ABS-852643).

References

[1] "Universal Storage Explained," 2021. [Online]. Available: https://vastdata.com/whitepaper/.

[2] "Introduction — ior 4.1.0+dev documentation," *Readthedocs.io*. [Online]. Available: https://ior.readthedocs.io/en/latest/.

[3] H. Shan and J. Shalf, Using IOR to analyze the I/O performance for HPC platforms. No. LBNL-62647. Berkeley, CA (United States, 2007.

[4] P. Braam, "The Lustre Storage Architecture," *arXiv [cs.OS]*, 2019.

[5] F. Schmuck and R. Haskin, "GPFS: A Shared-Disk file system for large computing clusters," in *Conference on file and storage technologies*, 2002.

[6] "Compute platforms," *Llnl.gov*. [Online]. Available: https://hpc.llnl.gov/hardware/compute-platforms.