A Study of NVRAM Performance Variability under Concurrent I/O Accesses

Anthony Kougkas, Hariharan Devarajan, and Xian-He Sun
Illinois Institute of Technology, Department of Computer Science, Chicago, IL
{akougkas, hdevarajan}@hawk.iit.edu, sun@iit.edu

***Index Terms*—Performance Variability, Performance Modelling, Benchmarking, NVRAM, NVME, SSD, Non-Volatile**

I. Introduction

Modern HPC applications generate massive amounts of data. However, the improvement in the speed of disk-based storage systems has been much slower than that of memory, creating a significant *I/O performance gap* [1], [2]. To reduce the performance gap, the storage subsystem is going through extensive changes, by adding multiple levels of memory and storage in a hierarchy [3]. Newly emerging hardware technologies such as High Bandwidth Memory (HBM), Non-Volatile RAM (NVRAM), Solid-State Drives (SSD), and dedicated shared buffering nodes (e.g., burst buffers) have been also introduced to alleviate this issue [4], [5]. Several new supercomputers employ such low latency devices to deal with the burstiness of I/O [6], [7], reducing the peak I/O requirements for external storage [8]. For example, Cori system at the National Energy Research Scientific Computing Center (NERSC) [9], uses CRAY's Datawarp technology [10]. Los Alamos National Laboratory Trinity supercomputer [11] uses burst buffers with a 3.7 PB capacity and 3.3 TB/s bandwidth. Summit in Oak Ridge National Lab will also employ fast NVMe storage for buffering, based on the first developer machine already deployed [12]. NERSC demonstrated [13] an improvement of 60% performance on balanced usage over applications not using burst buffer acceleration. However, they also stated that when two compute nodes share a burst buffer node, then their accesses compete for bandwidth which resulted in significant degradation in performance for both job. This phenomenon is even stronger for data-intensive applications which spend significantly more time in I/O. As multiple layers of storage are introduced into HPC systems, the complexity of data movement among the layers increases significantly, making it harder to take advantage of the highspeed or low-latency storage systems [14].

II. Our Approach

In this study, we aim to explore and uncover any performance variability of NVRAM devices. The difference of the medium (i.e., flash-based vs spinning drives) dictates different access concurrency, device bandwidth and latency, sensitivity to random access, and other performance variabilities such as garbage collection and data fragmentation.

A. Experimental Environment

As our testbed we use Chameleon systems [15]. Specifically, we used the bare metal configuration on the storage hierarchy nodes that have several storage devices, NVRAM included. Table 1 demonstrates the specifications of each device used. Even though this is an exploration of how NVRAM handles concurrent accesses, we included all devices as a comparison.

Table 1: Device specifications

Device	RAM	NVRAM	SSD	HDD fast	HDD
Model	M386A4G40DM0	Intel DC P3700	Intel DC S3610	Seagate ST600MP0005	Seagate ST9250610NS
Connection	DDR4 2133Mhz	PCIe Gen3 x8	SATA 6Gb/s	12Gb/s SAS	SATA 6Gb/s
Capacity	512 GB(32GBx16)	1 TB	1.6 TB	600 GB	2.4 TB
Latency	13.5 ns	20 us	55-66 us	2 ms	4.16 ms
RPM	-	-	-	15000	7200
Buffer	-	-	-	128 MB	64 MB

As our driver program, we used our own synthetic benchmark. Each process writes 64MB requests in a file-per-process pattern. We increase the number of concurrent processes while the total I/O remains 2GB (i.e., weak-scaling). We define a new metric, *medium-sensitivity*, as the rate at which each storage medium experiences bandwidth reduction due to concurrent access:

```
Medium-Sensitivity=(#Processes/#Lanes)*((MaxBW-RealBW)/MaxBW)
```

B. Initial Results

As it can in Figure 1, the NVRAM demonstrated sensitivity very close to the main memory. Specifically, for write operations, RAM has sensitivity value of 0.43, NVRAM has a value of 3.1 whereas the traditional drives 30 and 31 respectively. Same trends can be seen for read operations. These results are only the first step towards a more detailed study on performance variability of NVRAM we plan to do.

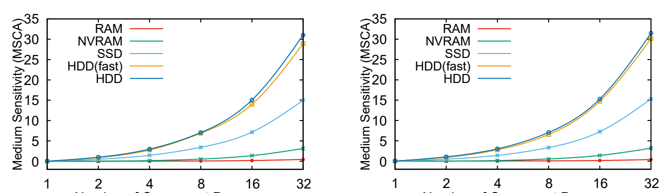

Figure 1: Performance Variability (left figure - Write, right figure – Read)

References

[1] Dong, Bin, Xiuqiao Li, Limin Xiao, and Li Ruan. "A new file-specific stripe size selection method for highly concurrent data access." In *Grid Computing (GRID), 2012 ACM/IEEE 13th International Conference on*, pp. 22-30. IEEE, 2012.

[2] Shoshani, Arie, and Doron Rotem, eds. *Scientific Data Management: Challenges, Technology, and Deployment*. CRC Press, 2009.

[3] Bent, John, Gary Grider, Brett Kettering, Adam Manzanares, Meghan McClelland, Aaron Torres, and Alfred Torrez. "Storage challenges at los alamos national lab." In *Mass Storage Systems and Technologies (MSST), 2012 IEEE 28th Symposium on*, pp. 1-5. IEEE, 2012.

[4] A. M. Caulfield, L. M. Grupp, and S. Swanson, "Gordon: using flash memory to build fast, power-efficient clusters for data-intensive applications," ACM Sigplan Notices, vol. 44, no. 3, pp. 217–228, 2009.

[5] S. Kannan, A. Gavrilovska, K. Schwan, D. Milojicic, and V. Talwar, "Using active nvram for i/o staging," in Proceedings of the 2nd international workshop on Petascale data analytics: challenges and opportunities. ACM, 2011, pp. 15–22.

[6] N. Mi, A. Riska, Q. Zhang, E. Smirni, and E. Riedel, "Efficient management of idleness in storage systems," *ACM Transactions on Storage (TOS)*, vol. 5, no. 2, p. 4, 2009.

[7] Y. Kim, R. Gunasekaran, G. M. Shipman, D. Dillow, Z. Zhang, B. W. Settlemyer *et al.*, "Workload characterization of a leadership class storage cluster," in *Petascale Data Storage Workshop (PDSW), 2010 5th*. IEEE, 2010, pp. 1–5.

[8] Lawrence Livermore National Lab, "Large Memory Appliance/Burst Buffers Use Case." [Online]. Available: https://asc.llnl.gov/CORAL-benchmarks/Large memory use cases llnl.pdf

[9] NERSC, "Cori system burst buffer design." [Online]. Available: https://www.nersc.gov/users/computational-systems/cori/burst-buffer/

[10] CRAY Inc, "Datawarp technology," 2017. [Online]. Available: http://www.cray.com/sites/default/files/resources/CrayXC40-DataWarp.pdf

[11] Los Alamos National Lab, "Trinity specs." [Online]. Available: http://www.lanl.gov/projects/trinity/specifications.php

[12] Whitt, Justin L, "Oak Ridge Leadership Computing Facility: Summit and Beyond," 2017. [Online]. Available: https://indico.cern.ch/event/618513/contributions/2527318/attachments/1437236/2210560/SummitProjectOverviewf gjlw.pdf

[13] W. Bhimji, D. Bard, M. Romanus, D. Paul, A. Ovsyannikov, B. Friesen, M. Bryson, J. Correa, G. K. Lockwood, V. Tsulaia, S. Byna, S. Farrell, D. Gursoy, C. Daley, V. Beckner, B. V. Straalen, D. Trebotich, C. Tull, G. Weber, N. J. Wright, and K. Antypas, "Accelerating science with the nersc burst buffer early user program," 2016

[14] A. M. Caulfield, J. Coburn, T. Mollov, A. De, A. Akel, J. He, A. Jagatheesan, R. K. Gupta, A. Snavely, and S. Swanson, "Understanding the impact of emerging non-volatile memories on high-performance, io-intensive computing," in *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society, 2010, pp. 1–11.

[15] Chameleon.org. Chameleon system, 2016. [Online]. Available: https://www.chameleoncloud.org/about/chameleon/