Performance Analysis of Commodity and Enterprise Class Flash Devices

Neal M. Master, Matthew Andrews, Jason Hick, Shane Canon, and Nicholas J. Wright
NERSC, Lawrence Berkeley National Lab, Berkeley, CA 94720
neal.m.master@ieee.org, {mnandrews, jhick, scanon, njwright}@lbl.gov

Abstract—**Five different flash-based storage devices were evaluated, two commodity SATA attached MLC ones and three enterprise PCIe attached SLC ones. Specifically, their peak bandwidth and IOPS capabilities were measured. The results show that the PCI attached devices have a significant performance advantage over the SATA ones, by a factor of between four and six in read and write bandwidth respectively, and by a factor of eight for random-read and a factor of 80 for random-write IOPS. The performance degradation that occurred when the drives were already partially filled with data was recorded. These measurements show that significant bandwidth degradation occurred for all the devices, whereas only one of the PCIe and one of the SATA drives showed any IOPS performance degradation. Across these tests no single device consistently out performs the others, therefore these results indicate that there is no one size fits all flash solution currently on the market and that devices should be evaluated carefully with I/O usage patterns as close as possible to the ones they are expected to encounter in a production environment.**

I. Introduction

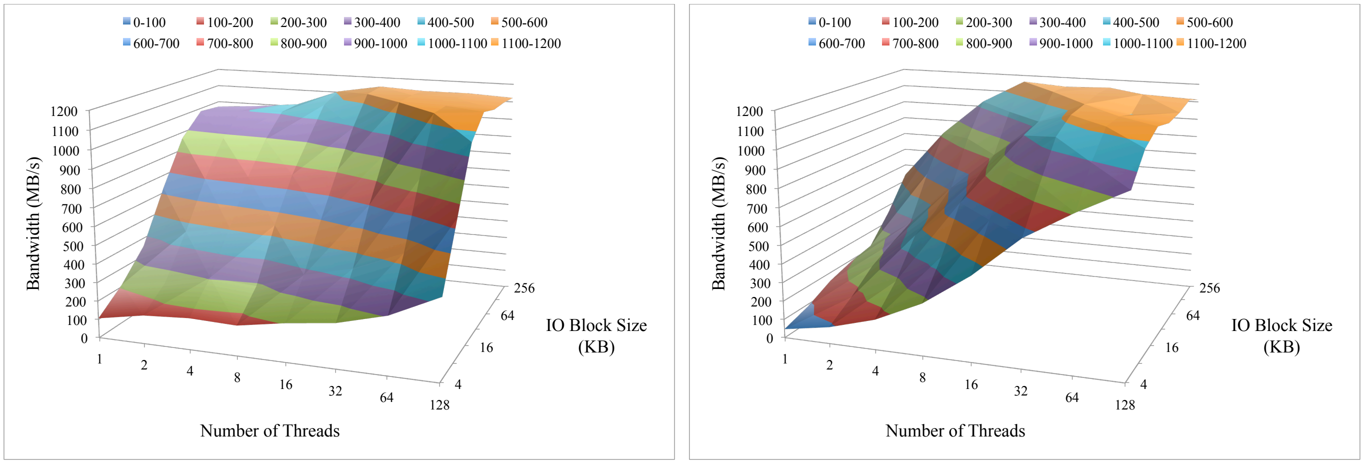
Flash based solid-state storage devices are expected to have a large impact upon the storage hierarchy in high-performance computing (HPC) systems. Indeed flash devices are already beginning to be deployed in large HPC installations, notably at the San Diego Supercomputer Center [1] and Lawrence Livermore National Lab. [2] Interestingly, these two deployments use different kinds of flash technology, the San Diego one is based on Intel SATA drives, whereas the Livermore one is based upon FusionIO PCIe attached cards. This difference is representative of the current knowledge of flash technology in HPC (and enterprise computing in general), there are many issues still to be explored and no consensus solution has emerged yet. This uncertainty arrises because flash can be used in storage in a number of ways. Accordingly, there are a range of products available at both commodity and enterprise levels. These vary by connection type, PCIe or SATA, as well as which kind of flash technology they use, SLC (Single Layer Cell) or MLC (Multi Layer Cell). In this work we explore the performance characteristics of five flash devices, three PCIe SLC, and two SATA MLC devices.

II. Methods

Five devices were evaluated: three enterprise flash storage solutions and two commodity solid state drives. The first enterprise storage device was the RamSan-20 from Texas Memory Systems. This device has 450GB of SLC NAND flash, uses a PCI-Express x4 connection. [3] The second was a Fusion IO ioDrive Duo with 320GB of SLC NAND flash and a PCI-Express x4 connection. [4] The ioDrive Duo is seen by the user as two 160GB “slots” that can be used in parallel to improve performance. In this work however, we decided to use a single slot so as to better control the parameter space. The final enterprise flash storage device was the Virident tachIOn. This device has 400GB of SLC NAND flash and uses a PCI-Express x8 connection. [5] All of the PCI devices have an FPGA controller on the cards themselves. Our two commodity devices were from Intel and OCZ. The Intel X25-M has 160GB of MLC NAND flash and uses a Serial ATA connection. [6] The OCZ Colossus has 250GB of MLC NAND flash and also uses a Serial ATA connection. [7]

To evaluate the bandwidth characteristics of each device, we used IOzone. [8] We varied the IO block size exponentially from 4KB to 256KB (4KB, 8KB, 16KB, ... , 256KB) while also varying the level of concurrency. We split the IO equally among 1, 2, 4, ... , 128 threads. While varying these two parameters, we measured the bandwidth to determine the optimal block size(s) and level(s) of concurrency. We did one of these scans for each of strided-read and strided-write and the mixed workload (50% read 50% write). We also used IOzone to measure IOPS, but controlled the variables differently. IOPS bound applications, like databases, tend to use small IO blocks. Therefore, when measuring IOPS we used a 4KB block size and varied the number of threads exponentially from 1 to 128 (1, 2, 4, 8, ..., 128). Each IOzone test was repeated five times to ensure consistency. All the IOzone tests were performed using the ext3 filesystem, starting from a freshly created partition.

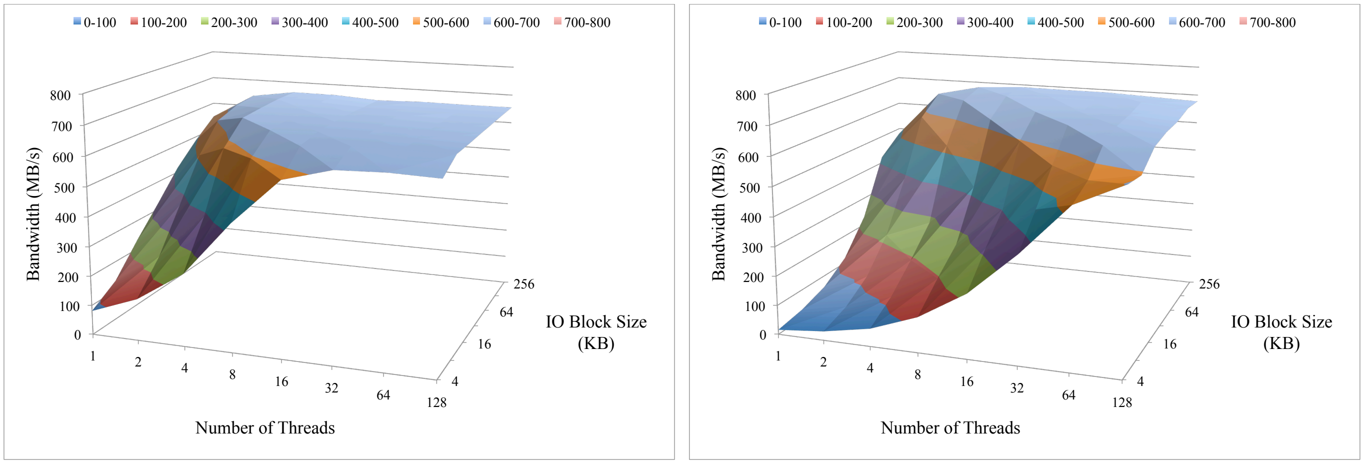
We also performed an experiment to try to understand how the performance characteristics of each device was effected by how full it was, under a sustained workload. We considered the bandwidth and IOPS of the devices after each had been filled with random data up to varying percentage of total capacity. We used the standard Linux utility `dd` to create files on each device and filled each device to 30%, 50%, 70%, and 90% of the device capacity in separate experiments. We then used FIO[9] to randomly write within the file for 1 hour. This was typically long enough to observe the steady state bandwidth and IOPS characteristics. For the bandwidth experiments we used 128KB blocks and 128 threads (64 threads for the SATA devices). For the IOPS experiments we used 4 KB blocks and 16 threads.



(a) Write

(b) Read

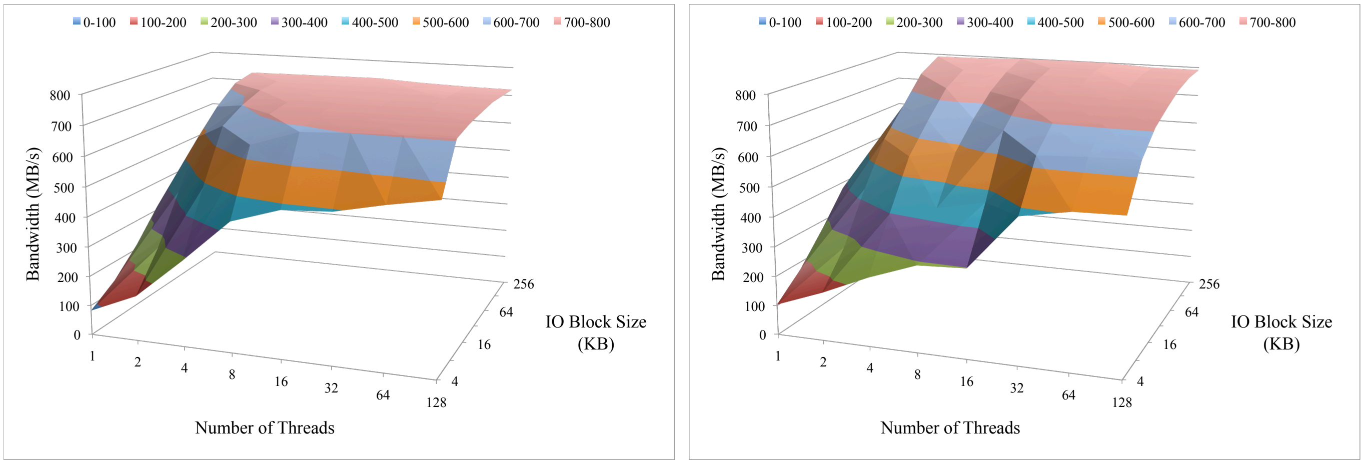
Fig. 1. Virident tachIOn (400GB) Bandwidth Plots



(a) Write

(b) Read

Fig. 2. TMS RamSan 20 (450GB) Bandwidth Plots



(a) Write

(b) Read

Fig. 3. Fusion IO ioDrive Duo (Single Slot, 160GB)

III. Results

A. Peak Bandwidth Plots

The bandwidth surface measurements show a few key distinctions between the enterprise and commodity devices as well as differences between vendors within each category. Figures 1 through 4 show some typical results.

Figures 1, 2,and 3 show the measured write and read bandwidths for the Virident tachIOn, TMS RamSan and FusionIO ioDrive cards respectively. All the devices reach the peak of the PCIe connection that they use for both read and write; because the tachIOn card is a PCIe-8x card and the others are PCIe-4x its peak bandwidths are approximately twice that of the TMS and FusionIO cards. The principle differences between the devices is in the shapes of the surfaces, i.e. how many threads and which block sizes achieve saturation.

For the PCIe-4x cards, the TMS RamSan and the FusionIO ioDrive, the write surfaces have very similar shapes, the only difference being the slightly greater ability of the Ramsan with blocks of <8KB. The read surfaces show more differences; a greater number of threads and/or larger block size is required to saturate the TMS Ramsam card. For the Virident card the surfaces look qualitatively different. Both bigger blocks and a greater number of threads are required in order to reach saturation, also, with block sizes of <32 KB for writes saturation is never reached. We also measured the surfaces for random read and write for all three cards (not shown). They show very similar properties.

The trends for the SATA drives are very different. A representative bandwidth plot for a SATA drive is shown in Figure 4. This particular Figure is for the Intel device and sequential read. It should be noted that the surfaces for all the other measurements for both the Intel and the OCZ drive are qualitatively similar. In addition to having significantly lower peak values (see Figure 5), we see that these devices are much less sensitive to variations in concurrency. For most workloads, including sequential read and write as well as random read and write, there was a slight benefit to using a single thread. There also tended to be a sharp decline when using 128 threads. This seems reasonable since these devices are intended for commodity desktop use; most desktop applications use relatively few threads. The SATA devices also show much greater variation with block size and only reach saturation with block sizes >64 KB typically.

It is also interesting to note that while block erasures typically make write operations much slower than read operations when dealing with flash, not all of these devices showed this difference. (See Figure 5.) The TMS Ramsan card showed less than a 4% difference between peak read and peak write bandwidths. Similarly, the Virident TachIOn card showed no difference at all; the peak values for both read and write were equal. This is not unique to the PCI cards, the OCZ device also had equal peak read and write bandwidths. The Fusion IO card and the Intel drive, on the other hand, show the asymmetry that we expected from flash storage. The Fusion IO card demonstrated about a 15% difference between read and write operations. The Intel drive was even more asymmetric with peak read bandwidth double the peak write bandwidth.

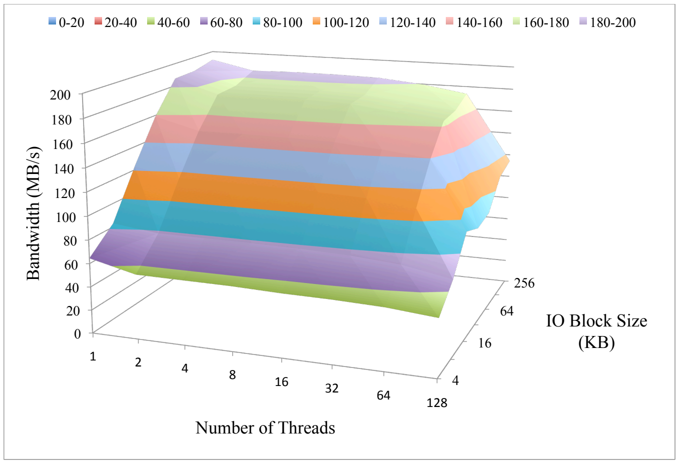


Fig. 4. Intel X-25M (160GB) Sequential Read

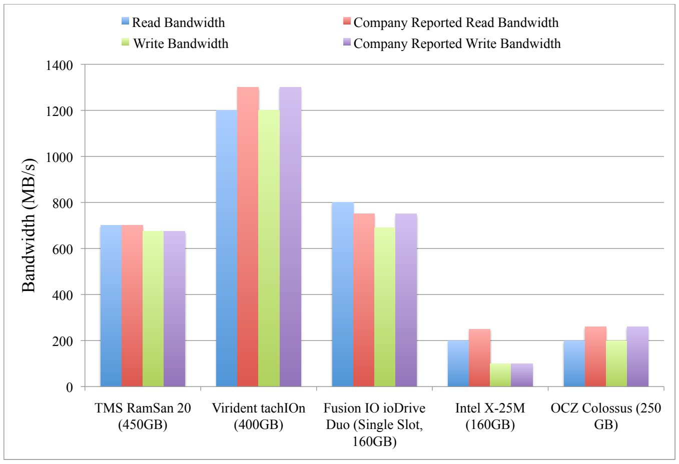


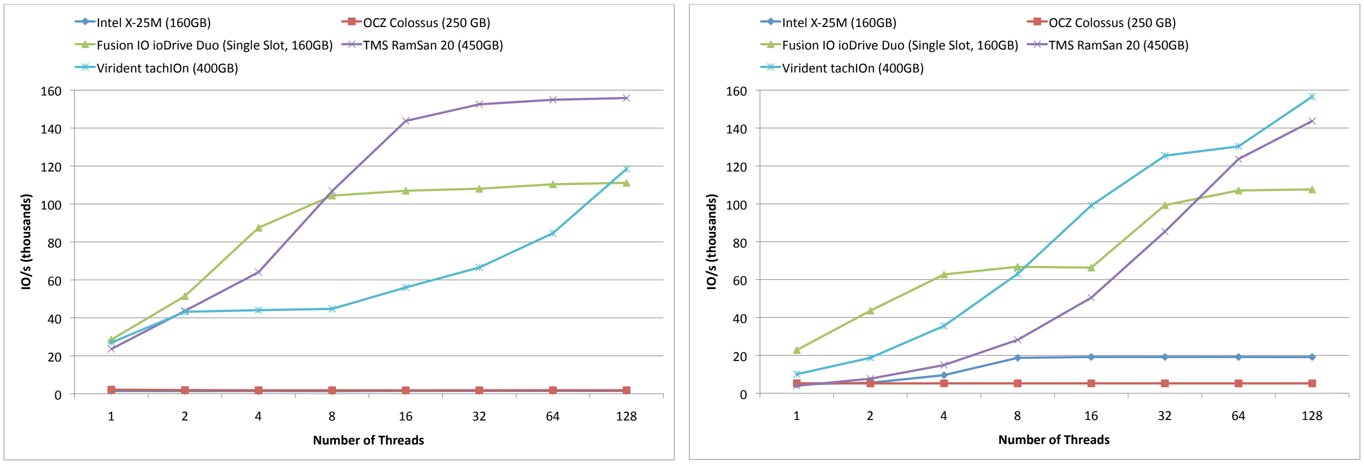
Fig. 5. Peak Bandwidth Values

B. Peak IOPS Plots

The results of our IO/s measurements are shown in Figure 6. In this case we measured the IO/s rate for random write and read with 4KB block sizes for each of the devices as a function of concurrency.

For both the random-read and write cases the SATA devices vastly underperform the enterprise ones, with less than 10% of the performance. Although interestingly, the Intel drive performs relatively well on random-reads, and achieves almost 20K IO/s which is more than 10× the random-write value and 4× the OCZ value. Also, in contrast to almost all the other measurements we made on the SATA drives, there is a dependence upon concurrency, with the peak value not being reached until 8 threads are used.

The Fusion-IO and TMS PCI cards both show similar behavior for random-write, they both reach saturation, at 8 and 16 threads respectively. However, the peak value for the TMS card is almost 160K IO/s whereas the Fusion card



(a) Random Write

(b) Random Read

Fig. 6. IO/s with 4KB Blocks

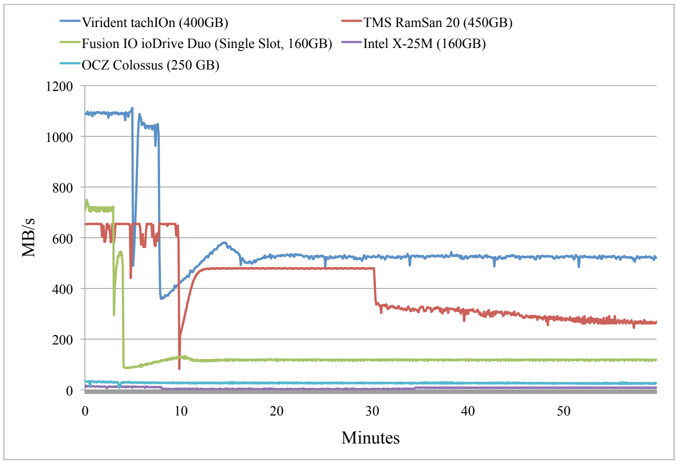


Fig. 7. Transient Random-Write Bandwidth Degradation (90% Capacity)

reaches approximately 110K IOPS. The Virident tachIOn card is the poorest performing for write IO/s, only equaling the performance of the Fusion card with 128 threads, which seems a little unlikely to happen under a realistic workload. Also it doesn't reach saturation. This is in contrast to the bandwidth measurements, where the PCIe-16x interface provided it with a larger performance advantage.

For random-read the performance is qualitatively different. At low concurrencies, the FusionIO ioDrive card is the best performing. However, at eight threads the performance advantage passes to the Virident tachIOn card, which with 128 threads achieves almost 50% greater performance than the FusionIO card. In contrast to the write case the TMS Ramsan card is never the best performing, although with 128 threads it is very close to the Virident card.

Interestingly, the peak random-read and random-write values achieved were not significantly different for the Fusion IO ioDrive and the TMS Ramsan PCI cards, in contrast to "conventional wisdom" that says that writes on flash are significantly worse than reads.

C. Degradation Experiments

Due to the intricacies of flash, specifically the need to erase whole blocks at a time, the performance of a device can be affected by how full it is, and by which IO patterns were used to fill it.

We begin by describing our degradation experiments for large block, 128KB I/O. The measured bandwidth as a function of time is shown in Figure 7. Typically within the first 15 minutes of the experiment, we see quite a bit of noise, which is most likely due to the flash controller switching algorithms as the device transitions from having spare blocks into the process of actually utilizing these blocks. Within about 30 mins all of the devices have reached a steady state often with a drastic decline in random-write bandwidth.

Our results, which show the steady state bandwidth as a fraction of peak for each device achieved as a function of fullness are shown in Figure 8. For the SATA drives the performance degradation is significant, although it shows no variation with time or fullness, typically 5-10% of the peak is observed right from the beginning of the experiment.

For the PCI cards, the performance degradation is significant. In this case the Virident tacIOn card is the best performing. It shows the lowest deviation from peak with 30-70% fullness, and is equal to the TMS Ramsan at 90% fullness. The FusionIO ioDrive card performs almost identically to the TMS Ramsam one for 30% and 50% but for 70% and 90% fullness is significantly worse, it only achieves 15% of its peak bandwidth with 90% fullness.

We also performed the same degradation experiments using 4KB blocks, to explore the performance degradation of IO/s. The only SATA drive to show significant degradation is the Intel SATA drive, which consistently shows about 20% of peak performance regardless of the fullness. The FusionIO

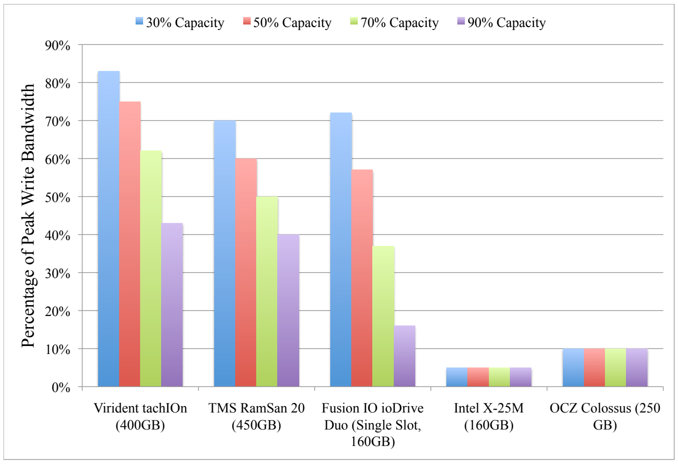


Fig. 8. Steady-State Random-Write Bandwidth Degradation

ioDrive is the only PCI card to show significant performance degradation, with an almost 50% performance hit at 90% fullness.

IV. Conclusions

Five different flash-based storage devices were evaluated, two commodity SATA attached MLC ones and three PCIe attached SLC ones. Specifically, their peak bandwidth and IOPS capabilities were measured. The results show that the PCI attached drives have a significant performance advantage over the SATA ones: by a factor of between four and six in read and write bandwidth respectively, and by a factor of eight for random-read and a factor of 80 for random-write IOPS.

The SATA drives appear to be primarily limited by the SATA interface itself, especially when it comes to bandwidth. The PCI devices are much more capable, however they are also more resource intensive. Each of them comes with a driver which runs on the host client and uses cpu cycles. (Our measurements of CPU usage showed at most a 25% load on one CPU core.)

The write bandwidth and IOPs performance for the PCI cards are very interesting. In contrast to the oft stated wisdom about flash, that there is a large asymmetry between the read and the write performance, we did not observe that in this case. However, it was true that a large number of threads (or large blocks) are needed in order to saturate the PCI devices and achieve peak values, especially in the IOPS cases.

We also measured the performance degradation that occurred when the drives were already partially filled with data, which is essentially a measure of the sustained performance achievable. The measurements showed that significant bandwidth degradation occurred for all the devices, presumably because of some combination of the grooming algorithm used and the amount of 'spare' flash storage that is on each device available for grooming. For some use cases in fact it maybe optimal to use them at less that the manufactures stated capacity. In contrast to the bandwidth case only one of the PCIe and one of the SATA drives showed any IOPS performance degradation. The FusionIO drive was the PCI one that showed performance degradation, which probably implies that more driver optimizations remain to be performed in this particular case, an advantage the PCI cards have over the SATA ones.

For the PCI cards the performance is also significantly influenced by the drivers. In fact during the course of this work we had to update the drivers from the original ones supplied to us after discovering anomalies for two of the devices. Also we had to ensure that the driver tasks were correctly pinned to the CPU cores closest to the PCI devices. The reflects the increased complexity of the Flash Translation Layer (FTL) on the PCI devices. For the SATA devices performance is a function of the SATA interface capabilities and the firmware/controller combination. For the PCI devices there is the extra layer of the drivers that run on the host and the choices that the manufacturer makes to partition the work to be done between the host and the controller on the card. It is almost certain therefore that with future driver releases some of the performance measurements described here will change.

Across all of these tests no single device consistently out performed the others, either from an absolute performance or a price/performance perspective. Therefore these results indicate that there is no one size fits all flash solution currently on the when market and that devices should be evaluated carefully with I/O usage patterns as close as possible to the ones they are expected to encounter in a production environment.

In future work we plan to look at the performance of these devices with specific applications, including databases and other data intensive HPC applications. We also plan to use the devices together, as part of a parallel filesystem.

V. Acknowledgements

This work was funded in part by the DOE award DE-FC02-06ER25767 the Petascale Data Storage Institute and by the Advanced Scientific Computing Research (ASCR) in the DOE Office of Science under contract number DE-C02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, under Contract No. DE-AC02-05CH11231.

References

[1] J. He, J. Bennett, and A. Snavely, "DASH-IO: an empirical study of flash-based IO for HPC," in *IEEE and ACM, Supercomputing 2010*, November 13-19, 2010.
[2] "Lawrence Livermore Teams with Fusion-io to Redefine Performance Density," http://www.fusionio.com/press/Lawrence-Livermore-Teams-with-Fusion-io-to-Re-define-Performance-Density/.
[3] "RamSan-20," http://www.ramsan.com/products/ramsan-20.asp.
[4] "ioDrive Duo Data Sheet," http://community.fusionio.com/media/p/461.aspx.
[5] "Virident Products," http://www.virident.com/products.php.
[6] "Intel X25-M and X18-M Mainstream SATA Solid-State Drives," http://www.intel.com/design/flash/nand/mainstream/technicaldocuments.htm.
[7] "OCZ Colossus Series SATA II 3.5" SSD," http://tinyurl.com/y9gfemv.
[8] "IOzone Filesystem Benchmark," http://www.iozone.org/.
[9] "fio," http://freshmeat.net/projects/fio/.