Proceedings of the 3rd Petascale Data Storage Workshop held in conjunction with Supercomputing '08,
November 17, 2008, Austin, TX.

Comparing Performance of Solid State Devices and Mechanical Disks

Milo Polte, Jiri Simsa, Garth Gibson
Carnegie Mellon University
School of Computer Science
Pittsburgh, PA, 15213

Abstract—**In terms of performance, solid state devices promise to be superior technology to mechanical disks. This study investigates performance of several up-to-date high-end consumer and enterprise Flash solid state devices (SSDs) and relates their performance to that of mechanical disks. For the purpose of this evaluation, the IOZone benchmark is run in single-threaded mode with varying request size and access pattern on an ext3 filesystem mounted on these devices. The price of the measured devices is then used to allow for comparison of price per performance. Measurements presented in this study offer an evaluation of cost-effectiveness of a Flash based SSD storage solution over a range of workloads.**

In particular, for sequential access pattern the SSDs are up to 10 times faster for reads and up to 5 times faster than the disks. For random reads, the SSDs provide up to 200x performance advantage. For random writes the SSDs provide up to 135x performance advantage. After weighting these numbers against the prices of the tested devices, we can conclude that SSDs are approaching price per performance of magnetic disks for sequential access patterns workloads and are superior technology to magnetic disks for random access patterns.

I. Introduction

For more than thirty years now, there has been a technological gap [1] in the memory hierarchy between the access times of random-access memories and mechanical disks. Moreover, this gap has been widening due to unbalanced improvement of these technologies. To this day, the difference between the access times of random-access memories and mechanical disks spans six orders of magnitude. This has severe implications on performance of applications that are not able to contain their working data set within random-access memory and incur seeks to disk.

Ever since the identification of the access gap, there has been ongoing research trying to bridge this gap. Its goal has been a technology that provides persistent storage with high performance, a suitable interface and competitive price and capacity. Trends of recent years [2] indicated that Flash could be a good candidate for the access gap technology.

Although Flash based SSDs have engendered much excitement, they are still relatively expensive and their performance may vary dramatically based on access pattern. Flash based SSD is a new technology with wildly varying capabilities so far, yet it is mature enough to have been adopted by a leading disk array manufacturer [3]. To better understand the behavior of Flash based SSDs, this study examines the performance of several high-end consumer and enterprise Flash SSDs comparing their performance to a few generally available mechanical disks. The comparison is carried out using the IOZone benchmark [4] that executes a series of microbenchmarks of varying access sizes and patterns.

The following section describes our experimental setup. Section III presents performance measurements and analysis of cost-effectiveness for the measured devices. Section IV discusses our plans for future work. Finally, this study is concluded in Section V.

II. Experimental Setup

For each of the measured devices we run the following IOZone command:

```
iozone -Raz -e -i 0 -i 1 -i 2 -s 4g -U
```

The meaning of the above parameters is:

- `-Raz` generate a report for all possible request sizes
- `-e` sync before ending each test
- `-i 0` run sequential read and write test
- `-i 1` run random read test
- `-i 2` run random write test
- `-s 4g` use 4GB for the size of the test file
- `-U` remount the filesystem before each test

In particular, every test was performed repeatedly for request sizes ranging from 4KB to 64KB, on an ext3 file system, remounted with no special options before every test. Because the system was running the benchmark in isolation, caching might have had significant influence on the performance. To avoid cache interference during the measurements, I/O requests covered an entire 4GB file, larger than the available memory.

As we shall see, Flash SSDs are challenged by the access pattern of random writes. It is no surprise that delayed write buffering, in the ext3 file system and in the device itself, can be important to achieved performance. For this study we use default settings—ext3 can buffer until IOZone issues the sync and the device uses default write buffering. Delayed write buffering can effect the amount of data that might be lost if power were to fail and all devices have some mechanism to defer responding to an application until each write is non-volatile, but few high performance systems use these mechanisms, often preferring to battery back up their servers, so we have not explored the performance implications of non-default buffering configuration.

Drive Type	Model	Erase cycles	Capacity	Price	Dollars/Gigabyte	Access Time	Ship Date
Consumer SATA SSD	MTron Mobi	100,000	16 GB	$370	$23.13	0.1 msec	2008
Consumer SATA SSD	Memoright GT	100,000	16 GB	$510	$31.88	0.1 msec	2008
Enterprise SATA SSD	Intel X25-M	10,000	80 GB	$730	$9.13	0.085 msec	2008
Enterprise SATA SSD	Intel X25-E	100,000	32 GB	$810	$25.31	0.085 msec	2008
Enterprise PCIe SSD	FusionIO ioDrive	100,000	80 GB	$2400	$30.00	0.05 msec	2008
7200 RPM SATA Drive	Seagate Barracuda 7200.11	∞	750 GB	$110	$0.15	4.2 msec	2006
10K RPM SCSI 320 Drive	Seagate ST3300007LW	∞	300 GB	$350	$1.17	4.7 msec	2005
15K RPM SCSI 160 Drive	Seagate ST3300655LC	∞	300 GB	$425	$1.42	3.9 msec	2005

TABLE I
SUMMARY OF STORAGE DEVICE ATTRIBUTES

Also, IOZone is single by default single-threaded in its data accesses—that is, only one access is issued to ext3 at a time. This may lead to the device experiencing less concurrent accesses than it can support, because Flash SSDs have multiple independent banks of Flash chips and disks can seek sort for shortest access first. As a consequence, our results do not achieve the highest performance possible with the device, as might be expected from its data sheet.

Devices compared in the experiment are described in the Table I. The SSDs selected for the comparison are high-end consumer and enterprise devices representing the state of the art devices as of 2008. In particular, the Memoright GT represents a state of the art consumer device, the MTron Mobi is typical of consumer drives found in laptops [5], [6], and the Intel devices as well as the FusionIO device are marketed as enterprise devices. The main difference between the Intel X25-M and Intel X25-E is, besides the capacity, the technology they are based on. Intel X25-M is uses multi-level Flash cells, while Intel X25-E uses single-level Flash cells. The former provides for higher capacity, while the latter for higher durability. The magnetic disks selected for the comparison were those readily available to us. Although they are two to six years old, the performance of their current equivalents does not differ dramatically.

The reported access times for the SSDs come from [5] and vendors data sheets. The performance numbers for the rotating drives also come from data sheets [7]. The price and capacity is reported for modern comparable disks.

Due to availability of the interconnect hardware, different machines were used for the measurements. The experiments for the MTron Mobi, Memoright GT and 7200 RPM drive were run on an Intel Dual Core 3.2 GHz machine with 2 GB of DRAM running Linux 2.6.22 and connected via a SATA/300 connection. The experiments for the Intel X25-M, Intel X25-E and the FusionIO ioDrive drives were run on an AMD x64 3 GHz machine with 4 GB of DRAM running Linux 2.6.24 and connected via a SATA/300 connection. The experiments for the 10K RPM drive were run on a 2.66GHz Pentium 4 with 1 GB of DRAM running Linux 2.6.18 and an Adaptec 3960D SCSI controller. The experiments for the 15K RPM SCSI drive were run on a 2.66GHz Xeon machine with 0.9 GB of DRAM running Linux 2.6.12 and an LSI Logic LSI53C1030 SCSI controller. All tests used the up-to-date version of the ext3 filesystem for their kernel, mounted without special options.

Since IOZone is designed to test the speed of the storage subsystem, it assumes that its execution is bottlenecked by the storage device. Under such assumption the above differences in the operating system, CPU speed, and DRAM size should not significantly affect the outcome of the measurements. In our measurements, this was true for the magnetic disks. However, for some of the SSDs we have not seen the advertised performance. We conjecture this is not due hardware but due to the combination of the single-threaded nature of IOZone and the ext3 layer, resulting in a shallow queue depth visible to the device. Since these two components were common to all the setups we believe that we can still derive value from our measurements.

III. MEASUREMENTS

In flash a non-empty page must be erased before it can be written to. Since a single erase is slower than the actual write, for efficiency pages are grouped into erase blocks that may span tens or hundreds of pages. On some devices, however, a write to any of these pages will result in the entire block being erased and rewritten, an inefficiency referred to as 'write amplification'. Moreover, erasing is destructive and the typical number of block erase cycles ranges from tens to hundreds of thousands depending on the type of Flash media, after which page writes may begin to fail. This poses many challenges out of which we comment on those related to Flash SSD's performance. For more detailed discussion see [8].

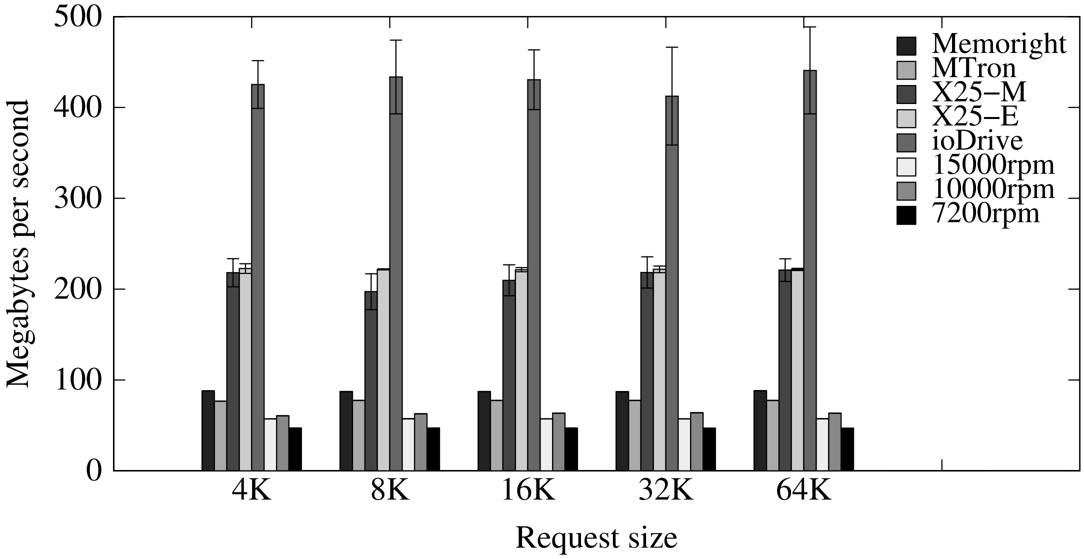
First, issuing a write to a non-empty page is expensive, because it requires the page to be erased. Second, small writes can wear out pages that were not being written to, decreasing the lifetime of the device. In order to overcome these problems, the SSD controller may keep a pool of pre-erased blocks of pages and implement a log-structured strategy for relocating data being written to more effective locations. Additionally an SSD typically is organized as four to ten banks of Flash chips and capable of independent, concurrent access. The SSD controller provides good performance only as long as it is able to maintain a big enough pool of pre-erased blocks of pages to service incoming write requests, and gather enough data to be written to make every erase-write cycle useful and keep all Flash banks busy.

In the following subsections we present figures summarizing the outcome of IOZone tests for sequential reads, sequential writes, random reads and random writes. For the sequential

Request size	Memoright GT	MTron Mobi	Intel X25-M	Intel X25-E	FusionIO ioDrive	15K RPM	10K RPM	7200 RPM
4 KB	87.91	76.85	218.06	222.63	425.28	57.16	60.54	47.05
8 KB	87.34	77.39	197.23	221.75	433.48	57.21	62.70	47.08
16 KB	87.25	77.40	209.73	221.42	430.48	57.20	63.62	47.07
32 KB	87.08	77.39	218.36	221.78	412.43	57.19	63.98	46.96
64 KB	88.09	77.38	220.96	221.64	440.71	57.27	63.61	46.94

TABLE II
SEQUENTIAL READS IN MBS PER SECOND

Fig. 1. Sequential reads



Request size	Memoright GT	MTron Mobi	Intel X25-M	Intel X25-E	FusionIO ioDrive	15K RPM	10K RPM	7200 RPM
4 KB	85.60	75.37	60.06	181.09	162.30	55.86	64.01	42.43
8 KB	93.88	76.17	64.95	175.21	157.42	55.48	60.37	42.10
16 KB	89.57	76.36	66.38	177.04	166.41	55.69	60.82	41.64
32 KB	84.87	70.64	68.77	177.99	165.65	56.02	63.69	41.82
64 KB	100.33	76.43	68.08	178.61	161.83	55.51	63.36	41.84

TABLE III
SEQUENTIAL WRITES IN MBS PER SECOND

reads and writes the bars represent measured performance in megabytes per second. For the random reads and writes the bars represent measured performance in terms of achieved I/O operations per second (IOPS).

Before starting the tests, we issued several writes to every logical address using the dd command. This is done to ensure that every logical block of the device has been accessed, as though the device has been in use for some time. In particular, we want the SSD controller to erase blocks while servicing write requests.

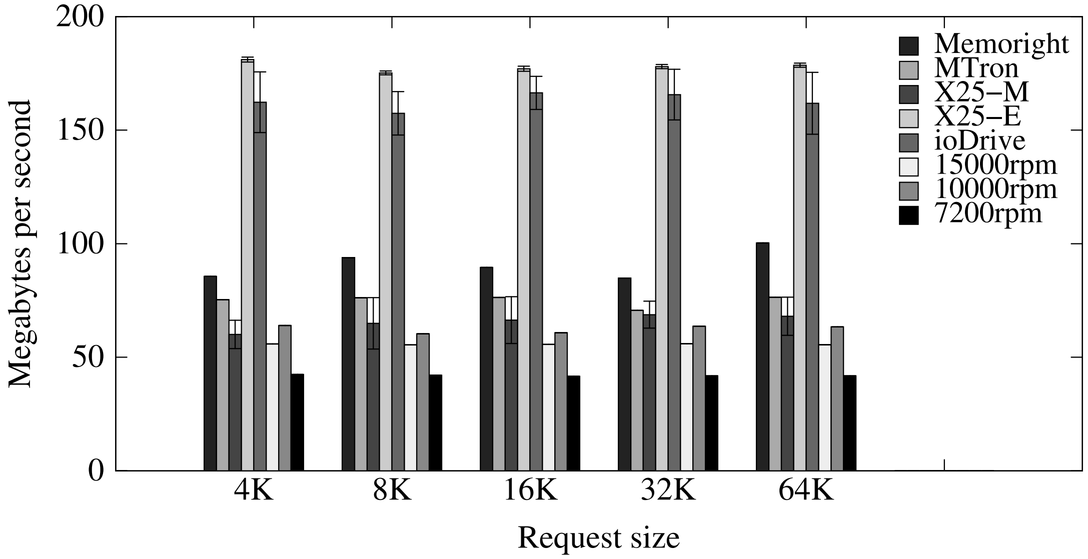
Moreover, due to the complex nature of some controllers, their performance can vary run from run. Therefore, each test was run five times for Intel X25-M, Intel X25-E and FusionIO ioDrive—the only devices that exhibited varying performance during our measurements. For these devices the graphs below give the average performance over five runs as well as error bars denoting the standard deviation. We did not rerun the test five times for all devices due to time constraints.

A. Sequential reads

In Fig.1 are the results of the sequential read test. The consumer SSDs achieve performance just below 90 MBs per second, the Intel SSDs performance just above 200 MBs per second and the FusionIO SSD performance above 400 MBs per second. The magnetic disks provide between 45 to 65 MBs per second.

Note that reading sequentially from the SSDs is a win over the 7200 RPM disk with a factor of 2x to 10x, while their relative cost is still slightly higher. Since a read bandwidth similar to that of SSDs can be achieved with a RAID 0 striping on a number of disks, for a given price, magnetic disks can provide comparable sequential read performance and much higher capacity.

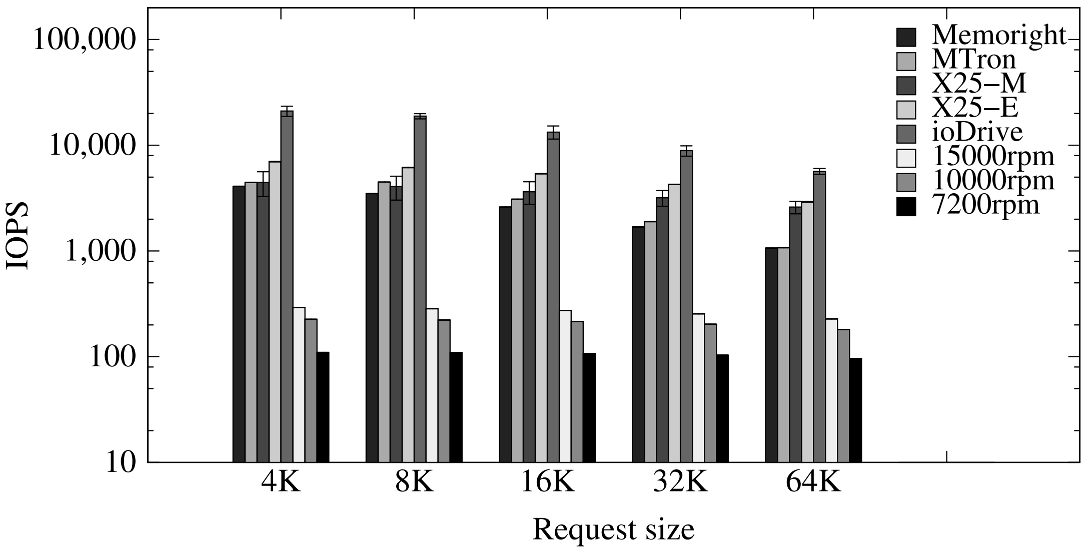
Fig. 2. Sequential writes



Request size	Memoright GT	MTron Mobi	Intel X25-M	Intel X25-E	FusionIO ioDrive	15K RPM	10K RPM	7200 RPM
4 KB	4101.25	4458.00	4450.75	6998.00	21093.35	291.50	226.00	109.75
8 KB	3484.00	4493.25	4068.05	6155.00	18853.60	283.88	222.75	109.00
16 KB	2602.00	3092.88	3634.84	5382.00	13334.04	273.25	215.88	107.38
32 KB	1689.19	1892.97	3192.00	4274.00	8895.71	254.44	203.00	103.59
64 KB	1067.78	1075.27	2601.59	2917.00	5659.12	227.92	180.44	96.06

TABLE IV
RANDOM READS IN IOPS

Fig. 3. Random reads



Another observation—that might be at first confusing—is that the 15K RPM disk is performing worse than the 10K RPM disk. This can be contributed to the faster interconnects of our specific 10K RPM disk (see Table I), and their relative ship dates.

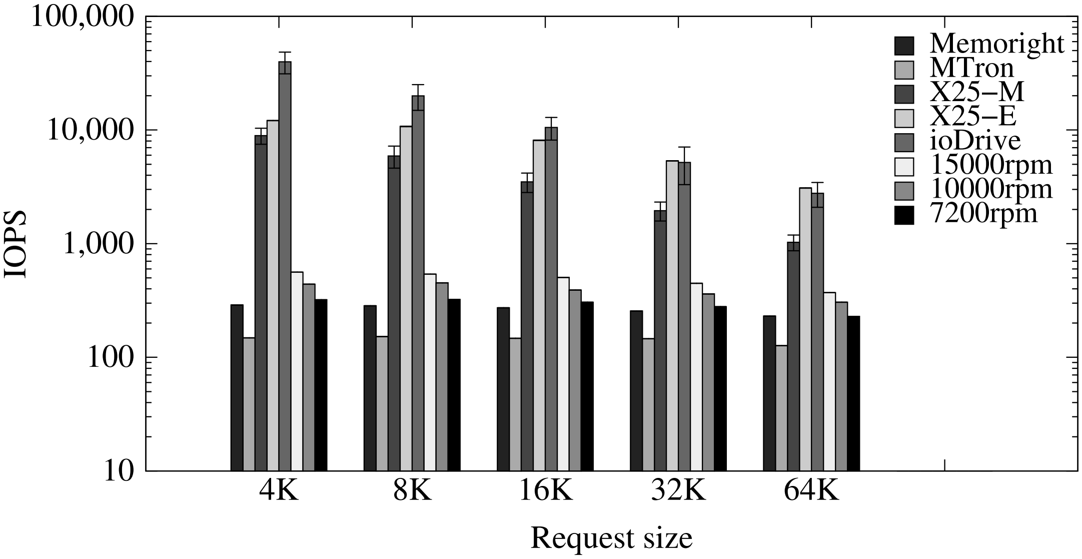
B. *Sequential writes*

In Fig.2 are the results of the sequential write test. First, note that the achieved performance is fairly constant across all request sizes for both SSDs and disks. While one would expect this kind of behavior for SSDs, it is a bit surprising to

Request size	Memoright GT	MTron Mobi	Intel X25-M	Intel X25-E	FusionIO ioDrive	15K RPM	10K RPM	7200 RPM
4 KB	288.75	148.50	8917.75	12104.50	39865.35	561.00	440.25	320.75
8 KB	284.38	152.00	5910.80	10755.38	19973.53	539.25	452.13	322.88
16 KB	273.00	147.19	3500.70	8112.66	10534.23	504.25	390.88	305.31
32 KB	255.91	145.59	1952.68	5354.01	5190.39	447.63	361.00	279.13
64 KB	230.70	126.67	1028.73	3089.24	2772.53	370.58	305.73	229.14

TABLE V
RANDOM WRITES IN IOPS

Fig. 4. Random writes



see it for disks as well. This can be attributed to the filesystem, which gathers written data and may issue device writes in the same actual size.

The consumer SSDs and the Intel X25-M achieve performance between 75 and 100 MBs per second, while the Intel X25-E and FusionIO SSD performance lies between 150 and 175 MBs per second. The magnetic disks provide between 45 to 65 MBs per second. Thus writing sequentially to an SSD is 1.5x to 5x faster than writing to a disk.

Using the same line of reasoning as in the previous subsection sequential write performance of SSDs is not a winning argument over a disk array. Consequently, SSD based storage solutions based on these products will not be cost-effective—in terms of price per performance and capacity—for workloads with predominantly sequential access patterns.

C. Random reads

In Fig.3 are the results of the random read test. Note that the vertical axis is displayed using a log scale in terms of I/O operations per second (IOPS). All SSDs follow a similar trend of decreasing IOPS with increasing request size, while disks exhibit a more constant IOPS across the request sizes, all relatively smaller compared to a magnetic disk track.

The consumer SSDs and the Intel X25-M achieve performance around 4,000 IOPS for 4KB reads. The Intel X25-E SSD achieves performance of 7,000 IOPS for 4KB reads. By far the best is the performance of the FusionIO SSD that achieves up to 20,000 IOPS for 4KB reads. As expected magnetic disk performance is in the range of 100 to 200 IOPS.

For small requests, the performance of the SSDs is 40x to 200x higher than that of a magnetic disk, with FusionIO being the fastest one, while their cost factors are relatively smaller. The take away is that for small random read workloads, SSDs are undoubtedly superior technology to disks.

D. Random writes

Fig.4 contains the results of the random write tests. Again, note that the vertical axis is displayed using a log scale in terms of I/O operations per second (IOPS). All enterprise SSDs follow a similar trend of decreasing IOPS with increasing request size, while both the consumer SSDs and magnetic disk exhibit a more constant IOPS across these request sizes.

The consumer SSDs we tested perform even worse than the disks for small random writes, achieving between 100 and 200 IOPS. It is likely that for these devices every write operation requires a block of pages to be erased and thus greatly reduce performance for small writes.

On the other hand, the more sophisticated controllers of the enterprise SSDs, Intel X25-M, Intel X25-E and FusionIO achieve excellent performance of 9,000, 12,000, and 40,000 IOPS respectively. We conjecture this is due to inherent log-structured pattern of writing, maintaining a pool of pre-erased

	Sequential reads ($ per MBs/sec)	**Sequential writes ($ per MBs/sec)**	**Random reads ($ per IOPS)**	**Random writes ($ per IOPS)**
Memoright GT	$5.80	$5.96	**$0.12**	$1.77
Mtron Mobi	$4.81	$4.91	**$0.08**	$2.49
Intel X25-M	**$3.35**	$12.15	**$0.16**	**$0.08**
Intel X25-E	**$3.64**	$4.47	**$0.12**	**$0.07**
FusionIO ioDrive	$5.64	$14.79	**$0.11**	**$0.06**
15K RPM	$7.43	$7.61	$1.46	$0.76
10K RPM	$5.37	$5.08	$1.55	$0.80
7200 RPM	**$2.34**	**$2.59**	$1.00	$0.34

TABLE VI
PRICE PER PERFORMANCE

	Sequential reads ($ per MBs/sec)	**Sequential writes ($ per MBs/sec)**	**Random reads ($ per IOPS)**	**Random writes ($ per IOPS)**
Memoright GT	$3.92	$4.25	$0.06	$1.02
Mtron Mobi	$3.70	$4.63	$0.02	N/A
Intel X25-M	$2.92	$10.43	N/A	N/A
Intel X25-E	$3.24	$4.76	$0.02	$0.24
FusionIO ioDrive	$3.43	$4.36	$0.02	$0.03

TABLE VII
PRICE PER ADVERTISED PERFORMANCE

blocks, coalescing writes to minimize rewriting data without changing it and servicing write requests in parallel.

Finally, the disks achieved performance between 300 and 500 IOPS, which is a bit higher than one would expect. We attribute this to ext3 that delays write requests and issues them in groups ordered by their logical block numbers. This allows the disk controller to service several request in a single rotation.

Thus for small requests, the performance of the enterprise SSDs is from 30x to 135x higher than that of a magnetic disk, with FusionIO being again the fastest one. The take away is similar to that for random reads, only this time it is true only for the enterprise SSDs, which for small random write have undoubtedly superior performance to disks.

E. Cost-effectiveness analysis

In order to put the measured performance into economic perspective, this section performs a simple cost-effectiveness analysis—using price per performance as our metric of choice. For the price comparison, we use the current unit-one prices, as many of these products are new and their markets are immature we expect these price to change perhaps dramatically in the coming years. For SSDs prices were obtained from suppliers [5], [9] and [10]. For the magnetic disks the numbers were obtained by surveying online prices using tools such as [11]. The results are summarized in Table VI.

For both sequential reads and writes the 7200 RPM disk is the most efficient device in terms of price per MBs per second. Out of the SSDs, it is the Intel X25-E that provides the best price per throughput for sequential reads and writes.

For random reads all SSDs do very well compared to magnetic disks, with MTron Mobi offering the best ratio of price per IOPS. On the other hand, for random writes only the enterprise SSDs do much better than magnetic disks, with FusionIO ioDrive offering the best ratio of price per IOPS. And it is also the FusionIO SSD ioDrive that offers the best balance of price per IOPS for both random reads and writes.

For comparison we present in Table VII the price per performance of what might be possible with the ideal workload of each device according to vendor data sheets. For example, FusionIO's numbers use IOZone on an XFS rather than ext3 file system.

IV. FUTURE WORK

The results above come from a simple configuration in which a user mounts a Flash SSD as a separate volume and manually manages its content. The performance numbers show that in order to utilize these Flash SSDs effectively, their storage should be employed primarily for small random requests. This access pattern, however, might be a difficult property for a user to correctly associate with files, suggesting a number of research questions dealing with more automated mechanisms for using Flash SSD storage in a larger system, including:

- How to best use Flash SSDs in combination with disks? As a general cache? Metadata stores [12]? Cache for specific data ranges?
- How file system knowledge and algorithms can be used to better partition data between a Flash SSD and a mechanical disk

V. CONCLUSION

In this study, we validate the common sense intuition that Flash based SSDs provide superior performance for small random I/O. Although widely expected, it is worth noting that sophisticated controllers only recently available are needed to make this true for writing.

In conclusion, for sequential access pattern the SSDs are up to 10 times faster for reads and up to 5 times faster than the disks. For random reads, the SSDs provide up to 200x performance advantage. For random writes the SSDs provide up to 135x performance advantage. After weighting these numbers against the prices of the tested devices, we can conclude that SSDs are approaching price per performance of magnetic disks for sequential access patterns workloads and are superior technology to magnetic disks for random access patterns.

The field is rapidly changing, the tested SSDs are just a sample of products coming into availability [6] and new products need to be carefully inspected for their performance characteristics.

Given the difference in the economic value of these devices on sequential versus random workloads, there remain many open questions regarding how to integrate them into a storage stack containing magnetic disks and how to manage the data stored on them. We believe that solid state devices are indeed a promising technology and that Flash storage will pose many interesting research questions for designers of high performance storage systems.

ACKNOWLEDGMENT

The authors would like to thank Shafeeq Sinnamohideen for helping them with running IOZone and Steve Bishop for his technical support on the FusionIO device.

REFERENCES

[1] E. Pugh, "Storage Hierarchies: Gaps, Cliffs and Trends," *IEEE Trans. Magnetics, v. MAG-7*, pp. 810–814, 1971.

[2] "Flash Memory vs. Hard Disk Drives - Which Will Win?" [Online]. Available: http://www.storagesearch.com/semico-art1.html

[3] "EMC in Major Storage Performance Breakthrough; First with Enterprise-Ready Solid State Flash Drive Technology," 2008. [Online]. Available: http://www.emc.com/about/news/press/us/2008/011408-1.htm

[4] "IOZone Filesystem Benchmark." [Online]. Available: http://www.iozone.org/

[5] "DV Nation." [Online]. Available: http://www.dvnation.com

[6] "Storage Search: The Fastest Solid State Disks." [Online]. Available: http://www.storagesearch.com/ssd-fastest.html

[7] "Seagate Barracuda 7200.9 datasheet." [Online]. Available: http://www.seagate.com/docs/pdf/datasheet/disc/ds_barracuda_7200_9.pdf

[8] N. Agrawal, V. Prabhakaran, T. Wobber, J. D. Davis, M. Manasse, and R. Panigrahy, "Design Tradeoffs for SSD performance," in *ATC'08: USENIX 2008 Annual Technical Conference on Annual Technical Conference*. Berkeley, CA, USA: USENIX Association, 2008, pp. 57–70.

[9] "CDW - IT Products and Services for Business." [Online]. Available: http://www.cdw.com/

[10] Fusion IO Coporation, "iodrive datasheet." [Online]. Available: http://www.fusionio.com/iodrivedata.pdf

[11] "Pricewatch." [Online]. Available: http://www.pricewatch.com/

[12] J. Piernas, T. Cortes, and J. M. García, "Dualfs: a new journaling file system without meta-data duplication," in *ICS '02: Proceedings of the 16th international conference on Supercomputing*. New York, NY, USA: ACM, 2002, pp. 137–146.

[13] "IOMeter – I/O Subsystem Measurement and Characterization Tool." [Online]. Available: http://www.iometer.org/

[14] "Intel X25-M 80GB Solid State Hard Drive Review." [Online]. Available: http://www.pcper.com/

[15] "High End Computing Revitalization Task Force (HECRTF), Inter Agency Working Group (HECIWG) File Systems and I/O Research Workshop HECIWG," 2006. [Online]. Available: http://institute.lanl.gov/hec-fsio/docs/HECIWG-FSIO-FY06-Workshop-Document-FINAL-FINAL.pdf

[16] J. Piernas, T. Cortes, and J. M. García, "Dualfs: a new journaling file system without meta-data duplication," in *ICS '02: Proceedings of the 16th international conference on Supercomputing*. New York, NY, USA: ACM, 2002, pp. 137–146.

[17] J. Piernas and S. Faibish, "Dualfs: A new journaling file system for linux," in *Linux Storage and File system Workshop*, 2007, pp. 12–13.

[18] "The Fastest Solid State Disks," 2008. [Online]. Available: http://www.storagesearch.com/ssd-fastest.html

[19] H. Kim and S. Ahn, "BPLRU: A Buffer Management Scheme for Improving Random Writes in Flash Storage," in *FAST'08: Proceedings of the 6th USENIX Conference on File and Storage Technologies*. Berkeley, CA, USA: USENIX Association, 2008, pp. 1–14. [Online]. Available: http://portal.acm.org/citation.cfm?id=1364829

[20] STEC Inc., "$Zeus^{IOPS}$ Solid State Drive." [Online]. Available: http://www.stec-inc.com/downloads/flash_datasheets/iopsdatasheet.pdf

[21] Intel Corporation, "Intel Mainstream SSD Datasheet." [Online]. Available: http://support.intel.com/design/flash/nand/mainstream/index.htm