Understanding Flash SSD Performance

Douglas Dumitru
CTO EasyCo LLC
August 16, 2007
DRAFT

Flash based Solid State Drives are quickly becoming popular in a wide variety of applications. Most people think of these solid state devices as just another hard disk, but their performance characteristics can be very different.

To start this discussion off, lets see a table of drive performance parameters for a couple of hard disk drives and a couple of Flash SSDs.

	Drive Model	Description	Seek Time			Latency	Read XFR Rate		Write XFR Rate	
			Track to Track	Average	Full Stroke		Outer Tracks	Inner Tracks	Outer Tracks	Inner Tracks
Hard Drives	Western Digital WD7500AYYS	7200 RPM 3.5” SATA	0.6 ms	8.9 ms	12.0 ms	4.2 ms	85 MB/sec	60 MB/sec*	85 MB/sec	60 MB/sec*
	Seagate ST936751SS	15K RPM 2.5” SAS	0.2 ms	2.9 ms	5.0 ms*	2.0 ms	112 MB/sec	79 MB/sec	112 MB/sec	79 MB/sec
Flash SSDs	Transcend TS8GCF266	8GB 266x CF Card	0.09ms				40 MB/sec		32 MB/sec	
	Samsung MCAQE32G5APP	32G 2.5” PATA	0.14ms				51 MB/sec		28 MB/sec	
	Sandisk SATA5000	32G 2.5” SATA	0.125ms				68 MB/sec		40 MB/sec	

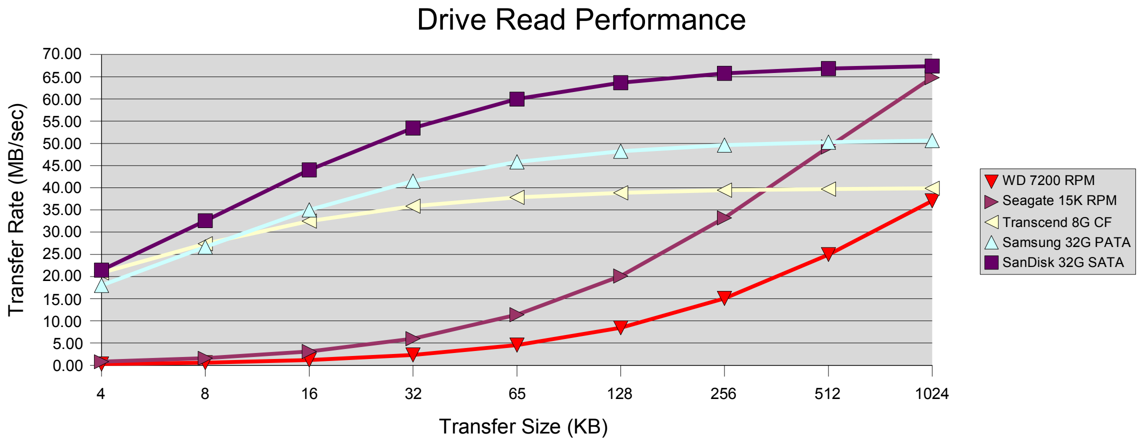
* Figure is an estimate

Of course there are other hard drives and Flash SSDs on the market, but this mix is a good starting point to work with.

Read Performance

The write performance of Flash SSDs is very different from the read performance. So we will talk about read and write performance separately.

When comparing read performance between hard disks and Flash SSDs the specs are often misleading. Flash SSDs have much better access times, but typically are slower in terms of transfer rates. This means that you have to consider the block size when evaluating read performance. Here is a chart that shows the average read performance for these five drives at varying block sizes:



This is actually a very telling chart. It shows that Flash SSDs far outperform hard disks unless the transfer size is very large. Even a CF card outperforms a 15K SAS drive for block sizes of 256K or less.

But I Work With Really Large Files

A lot of people underestimate how prevalent small IO operations are. Lets say you have a file server with word processing files on it. The average file size might be 250K. So what will your average IO size be? In all probability, the average read size will be somewhere around 100K and the average write size will be even lower around 50K. The reason for this is that there are disk blocks in use that are not used for the file's data. The file system needs directory entries, free space tables, and other structure to maintain consistency. Some file systems even have backup copies of critical elements. So reading the 250K will likely involve 3-5 small reads before the file is even located. Writing the file is similarly involved except that many file systems will do small writes even if you don't actually save the file. For example, Linux will write the ATIME (Access Time) just because you read the file even if you never update it. And this all assumes that your disk is not fragmented and that the file is written to 100% linear space.

Flash SSD Write Performance

It is with write performance that Flash SSDs become problematic. The issue here is the internal structure used within the Flash storage array. This structure includes a collection of bytes called an "erase block". When you write to a Flash SSD, the drive itself cannot just update the sectors you are changing, but must merge your changes with existing data to update a complete erase block. As Flash SSDs have gotten faster and larger, erase blocks have grown as well. Flash erase blocks used to be 16K in length. Now they are 1 Megabyte for small SSDs extending up to as large as 4 Megabytes for some models.

It is because of these large erase blocks that Flash SSDs are so slow at random writes. The "random write" performance of our three example drives is:

Drive	Random Writes/sec
Transcend 8G CF	47
Samsung 32G PATA	24
SanDisk 32G SATA	13

Your first reaction to this chart should be "that's terrible". You are right. These numbers are generally accurate for small random block writes. As the block size starts to approach the erase block size, you will start to get the stated linear write speed of the drive.

How Writes Impact Overall Performance

Because Flash SSDs have write performance that is so much worse than their read performance, the overall performance mix with reads and writes can be confusing. If you are doing pure reads, a Flash SSD will typically be 20x faster than a hard disk for small random reads. If you are doing pure random writes, the same drive might be 15x slower than a hard disk. But what about a 50-50 mix. You might think that the performance would balance out, but you would be wrong. Here are rough ratios for 4K operations with various read/write percentages (this table is for a SanDisk SATA5000 drive).

% Writes	Total IOPS	Performance vs 15K SAS Hard Drive
0%	5400	20x better
5%	252	1.25x better
10%	130	1.5x worse
20%	65	3x worse
50%	26	8x worse
100%	13	16x worse

This shows how even a very small percentage of writes can destroy the overall performance of a Flash SSD. It is for this reason alone that Flash SSDs, by themselves, are not very effective with random update applications like on-line databases, mail queues, and other environments that involve a lot of small updates.

Improving Write Performance

There are a couple of techniques that can improve write performance. Not all of these are available to everyone though.

OS Write Caching

You can turn write caching on with most operating systems. This will let the system buffer writes. This can make a drive “appear” to write faster, but the drive will have to do the actual writes eventually. If you are dealing with databases or are concerned about file system corruptions, OS write caching is not considered a good choice.

Flash Specific File Systems

Many devices that are designed to use Flash storage use Flash optimized file systems. File Systems like JFFS and YAFFS are optimized to minimize random writes. They also are designed to minimize the possibility of drive wear-out. Fortunately, with current Flash SSDs, the drives high endurance numbers of 1,000,000 write/erase cycles combined with large capacity means that drives will last over a decade before wear-out is even an issue.

Drive Write Caching

Some drives have internal memory that will buffer writes. This has the same effect as OS write buffering but is usually protected against power failures with batteries or capacitors that will push writes to flash. In general the effectiveness of write caching is limited because write intensive applications tend to overrun the cache pool anyway. Examples of drive write caching are the STEC Mach-8 and the Texas Memory System Flash RamSan-500.

Multiple Concurrent Erase Blocks

The Mach-8 and TMS RamSan-500 also feature internal arrays of drives so there are multiple erase blocks that can be manipulated in parallel. This pushes the sustained write rate of the STEC Mach-8 to about 800 writes/sec, which is a lot better than a “traditional” drive.

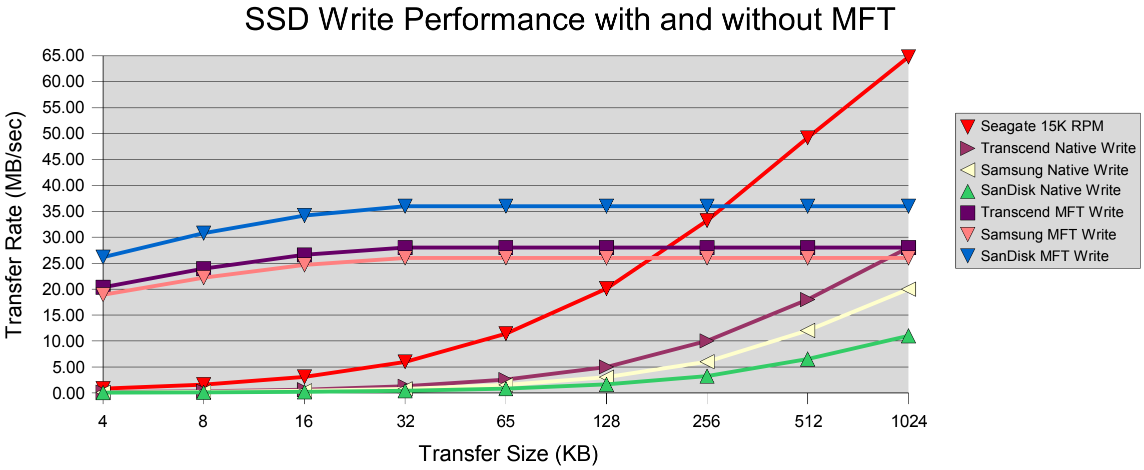
Unfortunately, the cost of these solutions is really high. Where a "commodity" Flash SSD might be $25/GB, the RamSan-500 is about $400/GB and the STEC Mach-8 is expected to be over $200/GB.

Massive Flash Over Commit and Caching

The STEC IOPS series of drives go a step further. They have almost double the internal Flash storage as their stated capacity, lots of memory (512 MB), and a RISC processor. This lets them run 10,000 write IOPS, although it is unclear if this is only a 512 byte number. If it is, then the 4K rate might be quite a bit lower. Unfortunately, these drives were recently quoted at $18,000 for a 32GB drive or over $500/GB.

Drive Block ReMapping

EasyCo's patent pending Managed Flash Technology (MFT) involves real-time remapping the data layout of the drive dynamically. This is currently implemented as a software layer inside of the host operating system. This allows the use of commodity drives while achieving random write performance that utilizes about 80% of the drives available linear write bandwidth. This lets a Transcend CF card write at 7,000 4K writes/sec or 28 MB/sec. A single SanDisk SATA drive writes at over 8,000 4K writes/sec.



The Impact of RAID Arrays

Flash SSDs have standard disk drive interfaces and are easy to use with hardware or software RAID. For reads, the RAID arrays tend to scale just like disk drives. If you have an efficient controller or software setup, you should see linear improvements based on the number of drives. One example we have tested here is:

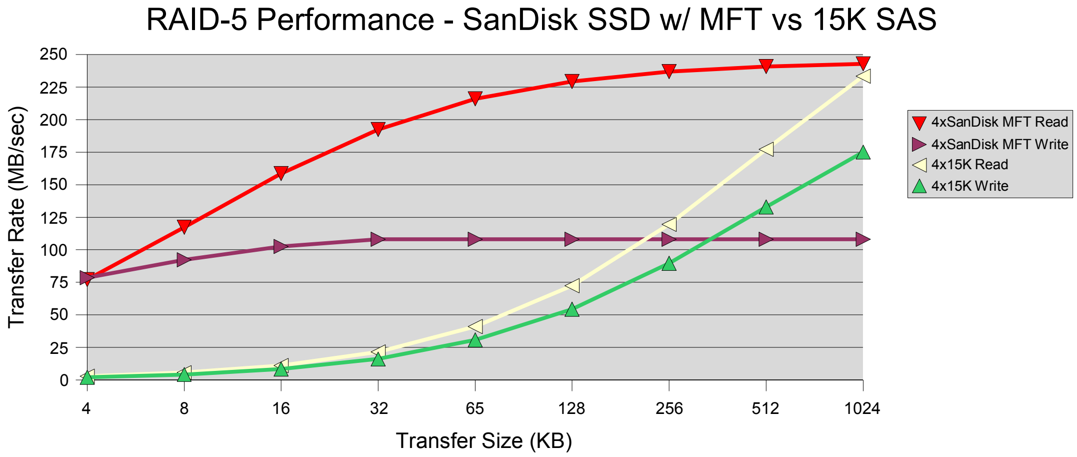
4 Samsung PATA drives
4 PATA to SATA bridge adapters
Highpoint PCIe 4 port raid adapter setup RAID-5
Electrical tape to hold this mess together
Red Hat Enterprise Linux 5.0 / x86 / 32-bit

Random read performance single thread is about 3600 4K reads/sec. With 20 threads, this jumps to about 14,000 4K reads/sec or 56 MB/sec. Linear read speed maxes out at about 140MB/sec. This implies that the PCIe 1x controller is hitting an internal bandwidth limit. The drives should theoretically deliver about 200 MB/sec. You should note that 14,000 random read IOPS would take an array of 70 15K SAS drives.

On writes, the linear write performance of the array is about 80 MB/sec. Theoretical would have been 84 MB/sec, so this is actually quote good. The random write rate ends up at about 35 writes/sec with multiple threads. This shows how RAID-5 and random writes don't really go together. If the array were RAID-0, you would expect a 4 drive random write rate of about 90 writes/sec. With RAID-10, you would expect a rate of about 45 writes/sec. Larger arrays would scale somewhat with RAID-5, but would scale linearly with RAID-0 or RAID-10.

If you take this same array and run it through the EasyCo's Managed Flash Technology mapping layer, the random write performance goes up to about 17,000 4K writes/sec or almost 500x the performance of the bare drives running RAID-5.

This chart shows 4 SanDisk SATA drives using the MFT management layer comparing performance to 4 15K SAS hard drives, both running RAID-5.



In order to equal the small block read and write performance of a 4 drive RAID-5 SanDisk SSD array with MFT, you would need a raid-10 array of about 80 15K SAS drives. The SSDs would draw about 3 watts of power while the SAS array would need nearly a kilowatt.

Conclusion:

Flash SSDs are rapidly developing with performance that already eclipses hard disk drives in many aspects. Each generation of Flash SSDs tend to get denser and faster. Solutions to Flash limitations, including poor random write performance, are finally coming to market. Together, these insure an increase in adoption of Flash storage solutions in performance critical enterprise servers.

About the Author:

Doug Dumitru is CTO of EasyCo LLC and the lead inventor of the “Managed Flash Technology” block device random write optimizing software. He has been working with large, centralized database servers for 20 years. He is co-owner of EasyCo LLC, a privately held company with offices in Pennsylvania and California. Mr. Dumitru can be reached at doug@easyco.com or 610 237-2000 x43.