Draft 10/21/10
Information in this document is preliminary and subject to change.

Flash SuperCharger™

Delivering Higher Performance to Large-Scale Flash Media Storage Systems with Lower Cost

Abstract

Flash SuperCharger delivers its greatest relative benefit to high speed Flash Memory mass storage systems such as SAN appliances and on-board Raid storage systems.

Most large-scale storage designers prefer to use a Raid-5 configuration for their mass storage. An eight drive Raid-5 set yields 75% more addressable storage than an equivalent Raid-10 set. Unfortunately, Raid-5 by itself is not feasible in an ordinary Flash environment. The practical 4KB random write is likely to be less than 2% of the random read speed and not more than 10% in the best conditions. Thus a 4KB random read/write mix of 70/30 will spend more than 80% of its saturated time performing random writes.

One would expect Raid-10 to be significantly faster. But again due to the limits of saturation performance, controller limitations and other factors, asymmetry is only 4:1 or less for most models of SSDs, while even the fastest performing SSDs cannot exceed 2:1. This is still problematic.

Flash SuperCharger, while allowing the use of lower-cost Raid-5, avoids these problems because of its inherent linear writing methodology. Random writing of small data elements is slow in Raid-5 sets because any random write requires some operation from each of the drives in the set.

When a small block of data is written to an eight drive Raid-5 set, the Raid-5 logic must random read blocks from six drives in order to compute the parity, and then perform one data write and one parity write. Conversely, if the write is the width of the stripe (for instance 64KB x 7 in an eight drive array), seven data writes will be made as will one parity.

Because SuperCharger linearizes random writes as clusters in FIFO order, and then writes this data on erase block boundaries, it is always able to write at the linear speed, and to avoid the IOPS limitations of Raid controllers. Similarly, on an identical surface basis, SuperCharger inherently writes almost half the data of Raid-10 to almost double the number of data surfaces. As a result, SuperCharger delivers 4KB random write performance three to ten times faster than a Raid-10 set. SuperCharger enhanced Raid-5 arrays also have four times the Flash media life of Raid-10 systems with identical drive counts.

SuperCharger's capacities do not end at simple speed improvement. SuperCharger also includes optional TPC compliance to assure data integrity, and does this in a manner which increases rather than reduces overall apparent write performance.

This set of capabilities gives the storage designer a broad range of capabilities. He can build truly massive storage systems of up to 30 addressable terabytes in a "4u" form factor with

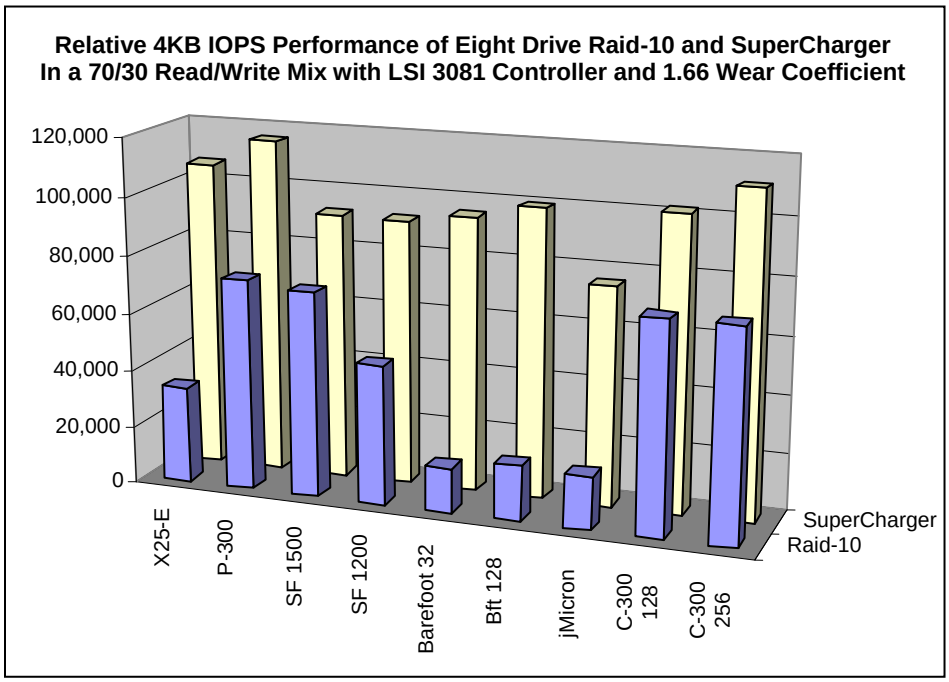
composite practical 4KB random IOPS rates in excess of a half million. He can also tune his system for performance and life or minimum cost. These are capabilities not previously seen.

In the following pages, this paper will explain what makes SuperCharger a consistently superior solution and back this up with test data and clear reasoning. To do so, we need to talk about the mechanics of flash including performance, wear and the impact of topology. We also need to talk about TPC compliance, Raid controllers, and the impacts of scale upon performance. At the end, we hope you will have a general understanding of why SuperCharger offers profoundly superior value.

Table of Contents

1. A Quick Comparison of Raid-10 and SuperCharger

Given that the tests in the latter half of this document demonstrate that simple Raid-5 and Raid-6 solutions just don't cut it in terms of performance, we want to start by giving you the bottom line comparison of the real world performance and cost of Flash storage subsystems built with either Raid-10 technology or SuperCharger enhanced Raid-5 technology. Here's what relative overall real-world performance looks like in graphic terms with different models of Flash SSDs:



SuperCharger managed Raid-5 and Raid-6 sets always outperform the same number of drives in a Raid-10 set. We could demonstrate it with graphics that look more impressive. But what this graph says that if you take the fastest flash drive available, the Marvel/Micron P-300 enterprise SSD, and run it with SuperCharger, you will get 58% more total throughput per second than using the same drives in a Raid-10 configuration.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets Adjusted for SATA-2 Raid Controller Limitations and Average Case Wear Equivalent to 60% Free Space

	Single Drive Statistics					Eight Drive Raid-10		Raid-5 with SuperCharger	
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	33,227	17.92	106,491
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	73,077	16.46	116,031
SandForce 1500	100	529	22,836	20,000	90	10.58	71,169	9.10	92,109
SandForce 1200	128	289	22,836	10,500	90	6.01	48,230	4.36	92,109
Barefoot 32GB	32	92	15,858	2,354	110	7.65	15,374	5.32	95,307
Barefoot 128GB	128	240	18,741	2,782	130	4.99	19,120	3.76	99,874
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	18,336	4.79	75,788
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	73,077	4.34	101,794
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	73,077	4.18	111,517

Not only that. You will also get 20% more usable space for the same expenditure, and if you could persuade Marvel to offer a dial-a-yield feature you could get 50% additional usable space per dollar of expenditure.

Moreover, SuperCharger's 4x wear enhancement of Raid-5 relative to Raid-10, discussed in detail in Section 10, means that almost all systems can be built with inexpensive MLC Flash Memory rather than eMLC or SLC memory. Thus, you could build the same system using the C-300 variant of this drive at a fifth the Raid-10 cost and almost the same performance.

Finally, it points out that the improvement of other brands of media is relatively even higher. With SuperCharger, even the poorest performing drive set outperforms the best drive set in a Raid-10 environment.

In the following pages, we will walk through the steps by which this table and graph were arrived at, returning at the end to this and sister tables with other sets of performance assumptions.

2. How Flash SuperCharger Works

Flash SuperCharger is block device management software running at the system level. Super-Charger linearizes all random writes, writing clusters of data in FIFO order. It maintains a memory lookup table so that the correct location of every storage element is known.

Flash operates on the principal of erasing a block and then filling its contents in a linear manner, first to last byte. MLC (Multi Level Cell) Flash memory typically has an erase block size of 512KB. But a number of these blocks are written in parallel to achieve speed, and so erase blocks typically have an effective length of two to eight megabytes.

By writing linearly on full block boundaries, SuperCharger is able to always write at 95% of the linear speed of the device, while minimizing the number of erase cycles.

When writing full speed, SuperCharger is typically writing a new block of data every 50 to 100 milliseconds. Conversely, if a half second passes and insufficient data has accumulated to fill a full block, SuperCharger will write the available data, flushing this out to the next available stripe of the drive set. A series of these is the functional equivalent of a full write.

Each of these writes is atomic and contains not just the data to be saved, but also metadata. Metadata is used to rebuild the server's memory table when a mount occurs after shutdown. Typically, a three terabyte array will mount in 1 to 3 minutes depending upon the number of drives in the array and their speed.

You will ask, can't SuperCharger "lose" data? There are two answers to this.

The first, and less expensive answer, is that if all data is written in FIFO order, it does not matter whether the system crashes a quarter second earlier or later. The data remains in order. It is not the undocumented spaghetti one gets when ladder sorts change the written order of data. Failure to write all data only becomes catastrophic when it is impossible to confirm what data is related to what.

That said, if a purist approach is needed, EasyCo offers the further refinement of software to utilize "Non-volatile DDR Flash backed-up RAM," assuring that in either a power fail or system hang, data not yet committed to disk can be recovered. This is an inexpensive solution as SuperCharger requires less than one gigabyte of RAM even for multi-terabyte arrays.

3. Compliance with TPC – Avoiding Data Loss in a Server Environment

In the description above, we have seen that, unlike Flash SSDs by themselves, SuperCharger does not lose data catastrophically. Inherently, all it can lose is the last fraction of a second of data, but because of its FIFO methodology, all data remains in order.

To reduce the chance of catastrophic loss, some "Enterprise class" manufacturers have added features. For instance, SandForce includes the capacity to build a Super Capacitor into the SSD, assuring that any data in SSD memory is flushed to flash media, and then reconstructed when power returns. Others, such as Intel and Marvel, include the ability to turn off RAM cache, thus assuring that data is not reported as written until it is physically written. Not all of these are "perfect" solutions. For instance, when cache is turned off on the Marvel design, the performance plunges more than 80% from 45,000 to 8,000 and change when used with SLC memory. MLC and eMLC memory would theoretically be worse.

But the problem of all these approaches from a SAN perspective is that they are piecemeal and don't solve the problem of data receipted in RAM memory but not yet written to the SSD set.

Conversely, in the first quarter of 2011, SuperCharger will offer a software option giving you the opportunity to build fully TPC ("Transaction Processing Performance Council") compliant systems, while reducing the turn-around time on write acknowledgement to about four micro-seconds.

The methodology is extremely simple: store data to be written in either battery backed RAM or Flash backed RAM. Thus, even if there is a power loss or hang, it is possible to recover data into RAM and then flush it to the SSD set, assuring data is not lost. The Flash backed RAM option is extremely appealing because it looks as if the market price of this product will be $100 to $150 per gigabyte of RAM. SuperCharger nominally needs less than a gigabyte of RAM to hold this in-transit data, though more may be needed to comply with interleave requirements.

TPC is just one step in building robust systems, but it is an important one.

4. Comparative Cost and Performance for Flash SSDs

The table below was compiled from independent test and manufacturer's data available on the Internet, as well as extensive testing of Indilinx BareFoot and SandForce based SSDs. The table shows the size and price of various SSDs, and their 4KB random read and write rates, as well as linear write rates. The items highlighted in green are "Enterprise" drives. Those highlighted in blue are tested devices, and those in blue and white are all "Workstation" drives[1].

The principle difference between server and workstation drives is the percentage of storage that is visible to the computer. Workstation drives typically have seven to thirteen percent free space, and are typically sold with an addressable size of 128GB or 120GB. or a multiple of these, with the disclaimer that the size is the number of "billions of bytes." Thus manufacturers play on the differential between 1,000 and 1,024-cubed to attain 7% free space. Conversely, server drives typically have free space of 27% and a reported storage capacity of 100GB, which in the fine print is 100 billion bytes. That said, the X25-E, available in 32GB and 64GB sizes, is also a server drive made with SLC memory, and has substantial free space.

Comparison of Enterprise and Workstation SSDs							
	Single Drive Statistics						
Model	Media Type	Size GB	Price	Price per Gigabyte	4kb Random Reads	4kb Random Writes	Linear Write mb/sec
Intel x25-E 32gb	SLC	64	699	10.92	35,000	4,800	170
Marvel/Micron P-300	SLC	100	999	9.99	60,000	45,000	275
SandForce 1500	MLC	100	529	5.29	22,836	20,000	90
SandForce 1200	MLC	128	289	2.26	22,836	10,500	90
Barefoot 32gb	MLC	32	92	2.88	15,858	2,354	110
Barefoot 128gb	MLC	128	240	1.88	18,741	2,782	130
jMicron 616 512	MLC	512	1,299	2.54	13,312	2,650	140
Marvel/Micron c-300 128	MLC	128	288	2.25	29,250	29,250	140
Marvel/Micron c-300 256	MLC	256	549	2.14	45,000	45,000	215

In the table above, we have listed each of the products by the brand name of its controller. This said, each controller product has many licensees and is sold under many brands. Accordingly, when buying drives, you should check out the controller being used.

While there is wild disparity in seeming price and performance between all of these, what will become obvious as we look at the various issues is that most have about the same effective performance in multi-drive storage arrays running under SuperCharger, and that SuperCharger will make each perform significantly better than any drive-only Raid configuration, whether it be Raid 5, 6, or 10. Finally, we will see that in some contexts (such as 72 SSDs in a 4u chassis) even the slowest of these devices, the jMicron 616, is practically no slower than the fastest available due to external limitations such as the performance of Raid controllers.

[1] We have been conservative in linear write speed numbers as these would inherently favor Flash SuperCharger. For instance, all SandForce device results have been computed on the basis of 90mb/second linear speed seen in our own testing. However, the linear write speed of Flash is in part a function of its quality and gating methods. Some SandForce based drives with better MLC are known to operate in the 120 to 130 megabyte per second range, which would increase the random write performance of SuperCharger by 33% to 44%.

5. The Limits of Raid Controllers

When building Raid storage systems, we recommend use of the LSI 3081 and its successors, such as the LSI SAS 9211-8i six gigabit controller. The minor reasons for this are because it is relatively inexpensive "dumb" controller, and is actually built into some motherboards, such as SuperMicro, thus freeing a slot. But the most important reason is that it is the fastest controller we have seen. This LSI 3081 can handle about 95,000 IOPS per direction. By comparison, when we tested the "intelligent" Adaptec 508 two years ago, it only delivered about 25,000 random read IOPS, though it has reportedly been improved since then. During that testing period, even a quite inexpensive Hi-Point "dumb" controller delivered 50,000. The general conclusion we have drawn is that while intelligent controllers can accelerate the speed of very large writes, their intelligence tends to get in the way of delivering maximum read or write IOPS.

What does this mean? Using the right Raid controller is very important, but it is also important to realize that even 95,000 IOPS can be a major roadblock. The 256GB Marvel can deliver 60,000 4KB reads per drive. Thus, an eight drive set should be capable of delivering 480,000 random reads. But the Raid controller limits reads to +/- 95,000, or a fifth of theoretical possibility.

As important, this chokes the Raid-10 and Raid-5 random 4KB write rates as well. Because Raid-10 delivers two buss writes per write, the maximum throughput for Raid-10 is 95,000/2 IOPS: 47,500 IOPS. Similarly, as writing to a Raid-5 set requires IOPS to all the SSDs, Raid-5 small random write output will be limited to 95,000/6 IOPS: 15,625 IOPS. This limit can be seen in the Raid-5 WITHOUT SuperCharger test results in Section 16.

Conversely, because Flash SuperCharger always performs long linear writes, and thus is not random IOPS bound, it has the capability of writing Raid-5 at the composite linear speed of the devices. Thus, in the 256GB Marvel Example, eight drive performance of SuperCharger extrapolates to 312,000 IOPS, fully six times faster than is possible with the same drive set operated Raid-10 with a 3081 controller, as well as three times faster than the random read speed of the same device.

There will be convergence of these values over time. The 6GB SATA-3 replacement for the 3081 is the 9211-8i. This will handle about 290,000 IOPS according to LSI's spec sheets. However, we would urge caution as the 290k number is probably a composite for reads and writes. The practical limit shown to date by testing has been 200,000 read or write, which is 100,000 Raid-10 writes.

6. The Limits of Communications Controllers in Storage Appliances

Just as Raid controllers impose performance limits on throughput, so communications controllers impose throughput limits when an external storage appliance is used to deliver data to one or many independent servers.

For instance, 10 gigabyte iSCSI sounds incredibly fast until you realize that this is only 1.2 gigabytes per second of theoretical throughput, or a maximum possible throughput of 300,000 4KB IOPS per second. Here, the practical throughput is less than the theoretical, both because of cramming and the significant overhead of TCP processing. That said, intelligent iSCSI cards may improve throughput.

Even the newest Fiber Channel technologies are not that much faster than iSCSI, and it is not until one gets to a technology such as Infiniband that one gets exponents of performance improvement.

Similarly, to some degree, designers can improve performance by using multiple channels. But care should be taken to assure that any plumbing does not throttle intrinsically fast throughput. And conversely, care should also be taken on the server end. A one gigabyte card can deliver 25,000 4KB IOPS to a target server, at a rate equivalent to a hundred 15,000 rpm SAS drives, but if a 100 megabit card is used, throughput will be throttled to a relatively useless level.

7. Understanding Flash Media Options

Unlike hard disks, Flash media has a finite number of times that it can be overwritten. This number of overwrites is based upon the quality of the Flash, and the cost varies significantly with the number of overwrites permitted. The following table provides information on the various grades of flash:

Flash Memory Types				
Acronym	**Description**	**Price per Gigabyte (9/26/10)**	**Erase Cycle Warranted Life**	**7-Year Maximum Overwrites per Day**
SLC	Single Level Cell	6.05	50,000	19.57
eMLC	extended life Multi Level Cell	?4.00	30,000	19.57
MLC	Multi Level Cell (high grade)	1.45	5,000	1.96
MLC	Multi Level Cell (low grade)	1.00	3,000	1.17
TLC	Three (or four) Level Cell (high gr.)	0.97	3,000	1.17
TLC	Three (or four) Level Cell (low gr)	0.97	1,500	0.59

The price shown for eMLC is a guess as this is a proprietary product available only from Micron. All other prices come from http://DRAMExchange.com. In considering which memory type to use, there are two primary considerations.

The first will be how much writing of new data you expect to do. Overwriting even all the data on your Flash storage system once a day for seven years is a daunting prospect. The vast majority of Enterprise systems update only 5% to 50% of a system's data in a day, for the simple reason that they need access to huge volumes of historical data as well. Completed orders rarely change, but today's do.

Similarly, to the extent that you maximize the advantages of SuperCharger and build larger storage arrays with much lower cost "commercial grade" media rather than "Enterprise" media, the tendency of your customers will be to put more of their data into flash arrays, rather than on high performance hard disks or even slower drives, just because of the lower cost. Just as 256GB SSDs wear out much more slowly than 32GB, so bigger stores tend to wear out more slowly than small intensively used stores.

Finally, as we will see in the coming pages, the Raid-5 wear advantage and data locality both have a significant influence on actual wear and performance, as do other factors which can reduce the effective wear on media.

Using SuperCharger, it is possible today to build a 33 terabyte array with 72 drives in a 4u case. Similarly, the media for such a system will cost under $4 per addressable gigabyte. To build the same sized array with 15k rpm SAS drives would require about 1,300 37GB SAS drives, which would cost about $20 per addressable gigabyte, even without short stroking, and would also require two to four racks of storage space and power while still running slower.

But we must consider more than just new data written to media. The second concern will be the correct wear co-efficient in use on your flash media. On first generation flash drives, wear coefficients were extremely bad.

8. A Small Mail Server and Other Examples of Real World of Wear and Free Space

Any discussion of performance and durability must ultimately begin with real world examples. EasyCo has been delivering servers with its SuperCharger capability since July of 2007. Conversely, the current generation controller chips are just a few months old. Accordingly, we will share some of our long term examples.

For several years, we ran a commercial mail server for about 400 users spread over numerous end-user entities. This Linux server was built using two 32GB CF 133x flash cards, in part because reasonably priced SSDs didn't even exist at the time. Each card was mirrored to a 7200 rpm hard disk, and then the two mirrors were striped. 4KB Random Read IOPS for this device were about 7,000. Random writes without SuperCharger were about 16, but with SuperCharger were about 7,000 as well. Given that 2/3rds of the activity of this system were writes or deletes, the write performance mattered, and produced comfortable response for IMAP users. The system read and wrote approximately 60 times faster than it did before Flash media was installed.

Gross storage on this device was 64 billion bytes (59.4GB), of which 10% was set aside as mandatory dedicated free space, yielding 53.4GB of logically addressable space. On any typical day, this server had a low point of 36GB of actual email data, and a high point of approximately 48GB of consumed space. That said, in several instances, it totally ran out of space because of excess spam or extended weekends.

On any given day, the low point typically consisted of 30GB of storage used for IMAP service, and 6GB of undelivered POP3 mail. The server received 12GB of new mail on a typical day, typically writing this to media, reading the same for download, and then deleting it. In the case of IMAP, something similar happened in that trash was purged after 7 days.

Statistics show that the wear coefficient for this machine was a surprisingly low 1.3 even after several years of operation rather than the theoretical 3.4 which might have been projected. A 1.3 wear coefficient means that when the flash cells are updated, 77% of the data written is new data while only 23% is rewritten information. As 27% of the drives were getting updated on a daily basis, and the storage media was 5,000 erase cycle MLC, the media had a projected wear life of 50.4 years, even though the average the average free space over time was only 29% (42GB/59GB).

This machine was retired in January of this year and replaced by a larger machine with about twice the IOPS rate and size, using left over SSDs with a 40MB/sec linear write speed. The new system is split with about 60% mail and the rest as database files. The wear coefficient of this system over eight months has been <2.0 even though it had about 37% free space, and should have had a coefficient of 2.7. There is suggestion, which will be discussed later, that locality can improve with time in the MFT environment.

Recently, we had a call from a customer with a two year old system used to store a distribution database. They were getting massive numbers of error messages that they didn't understand. A quick check of the machine showed that they were out of addressable disk space. In spite of filling their disk, they had a lifetime wear coefficient of just 1.4, the equivalent of 72% free space, even though their system had had only 15% free space for most of the time, because we didn't teach them how to take advantage of dynamic free space. This company was updating half of its system daily, with large numbers of updates to today's records, which only represented a quarter percent of the storage, as they maintain several years of history on line. Accordingly, they had very high locality, and had a projected life of 30 years for their media.

Some manufacturers tout SLC and eMLC. But here are servers that do significant work every day without durability or performance concerns. The reason this is so is because the software (Flash SuperCharger) has been optimized to minimize wear while optimizing random write

performance at approximately 95% of the linear bandwidth of the system. In the next few sections, we will expand upon wear and performance issues.

9. Understanding Wear and the Necessity of Free Space

We asserted above that Flash media has a finite number of erase cycles available to it. But this is only half of the problem. The bigger problem is that Flash Media is managed in chunks, called erase blocks, that are minimally a half megabyte long in the case of MLC memory, but that can become several megabytes long when a number of these are tied together in parallel to increase speed.

For instance, the first generation jMicron chip called the 604 had 8MB erase blocks. While it had a high linear write speed of well over 100MB per second, and a reasonable 4KB random read speed of about 4,200 IOPS, it had a random write speed of only about 15 random writes per second. The way the jMicron worked was that when a 4KB update was required, the entire 8MB stripe had to be rewritten. When a 4KB update occurred, this had to be merged with 2047 other 4KB blocks of data which were rewritten from an existing erase block. Thus, when all data written to a server was in 4KB random blocks, if the drive media was to last for seven years, the write rate would have to be $<^{9}/_{100}\%$ of the drive per day. That is not to say that the 604 was a bad product. It was typical of its time, and was explicitly designed for use in workstations where the average write block is very large and approaches 100KB, reducing the wear problem by a factor of 25. It's just that the 604 was totally unsuitable for use with servers unless used with our Flash SuperCharger product.

Second and third generation SSD controllers have become much more sophisticated since then and can temporarily use the free space on a drive to accumulate groups of changes before committing these as a composite update in the same way the 604 did. Basically, there are three such methods. One is to use a ring buffer. A second is to leave blocks or parts of blocks empty and fill as one goes, using back referencing so that the disk properly points to the current copy of a sector, rather than older copies. The third, which SuperCharger uses, is to create long linear regions of space, and to maintain a general table of references, avoiding the problems of back-linking in Flash.

When use is made of free space in this way, then the wear coefficient is fundamentally determined by free space (with the exceptions discussed below). For instance. If there is 7% free space, the wear coefficient will theoretically be 1/.07 or 14.2, and if there is free space of 27% (as is explicitly the case with Enterprise class SSDs), then the wear coefficient is 3.7. But rather than thinking of coefficients, it is often easier to think of the percentages themselves.

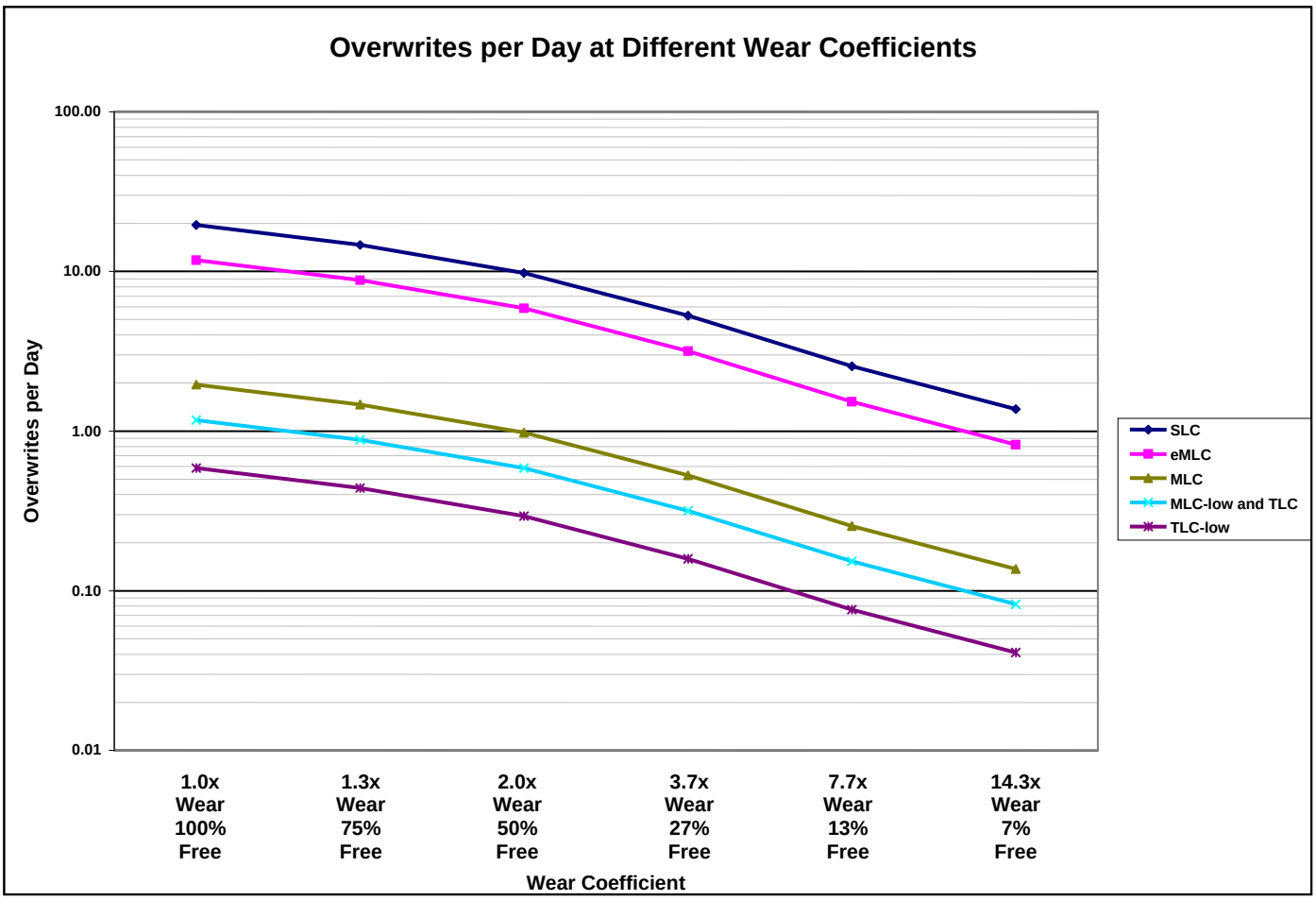
For instance, in the table of Flash Media types, above, we specified that an SSD built with 5,000 cycle MLC and with a 1.0 wear coefficient could be expected to last for seven years as long as it was not overwritten more than 1.96 times in an average day. To come up with a realistic number based upon free space, we can take the enterprise wear coefficient and apply it in the form 1.96 x 0.27. Thus Enterprise drives (i.e. with 27% or better free space) with 5,000 erase cycle media can be expected to last seven years as long as the average amount of random writes per day is <52% of the gross disk space. (Note that we say "gross space" here. If the drive with 128GB of is sold as having 100 billion bytes of addressable space, then the the <52% of gross space becomes <71% of addressable space.)

There is a problem with this, as you can see. What if you are using an enterprise drive with MLC and expect 100% overwrites to your logical (addressable) space in an average day? Here, if you have 27% free space, your overwrites per day would be three times the gross space and 3.8 times the addressable space. However, if your logical space is reduced to 66% of the gross disk size, you will find that the equation balances: 128*0.66/.34 = ~256.

To assure that you don't over-wear your SSDs, the ideal solution would be a dial-a-yield feature from the manufacturers. But no SSD controller manufacturer now includes such a capability. However, you can do so yourself simply by wiping the SSD set clean and then creating a

partition/volume that is smaller than the reported available space. For instance, if you have 128GB of flash, whether sold as a 128 billion byte drive, a 120, or 100, as long as your partition or volume is 84 gigabytes, you will have a theoretical wear coefficient of 2.0, and will be able to overwrite the entire media of the drive once a day with the expectation that it lasts for seven years.

Below, we have prepared a table that shows overwrite capacity of different media. The principle one you should pay attention to is the olive line, which represents good quality current generation MLC media. MLC media is three to five dollars less per gross gigabyte than either eMLC or SLC media. It is normally suitable for most mass storage situations.



While the methodology discussed above is generally true, it should be taken with a grain of salt. On the one hand, because of our statistical methods and long blocks, SuperCharger will always produce a wear yield superior to the theoretical. Given the certainty of our design, we have chosen to be highly transparent about our methods. The same cannot be said with certainty by all controller makers. Some of the methods such as ring buffers are likely to require double writing of data, and thus halve the yield from free space. Double writing may also be a problem for some relative referencing models. Many of the designs are by-gosh-and-by-golly, and theoretical wear cannot be summarized or tested. But if you build with SuperCharger, you don't have to worry about these internals at all because SuperCharger always writes full erase blocks on erase block boundaries, and thus operates on all drives in the same manner, generally bypassing these alternate but contingent methods entirely.

10. SuperCharger Raid-5: Four Times the Intrinsic Media Life of Raid-10

Having talked generally about wear and free space in the last section, we want you to now understand why SuperCharger Raid-5 intrinsically has four times the wear life of the same drive set configured Raid-10 even before other wear and performance advantages of SuperCharger are considered. We will look at this advantage two ways: first as a function of free space, and second as a function of daily write volume, before talking about inherent design consequences.

Let's begin with the assumption that we are building a system with 24 drives. If we use a 24 drive set of 200GB enterprise drives, we will have 6,144GB of intrinsic gross space, 4,800GB of gross apparent disk space when using Enterprise drives, and 2,400GB of addressable space in a Raid-10 paired drive configuration. We will also have 27% free space, for a wear coefficient of 3.7.

Similarly, if we take a similar 24 drives set to their 7% gross space setting of 256GB, after we set aside one drive for parity, we will end up with 23 x 256 = 5,888GB of gross addressable space, and 5,888 x 0.27 = 4,298 of net (user) addressable space after set aside of mandatory free space.

But if we only need 2,400GB of addressable space, we could configure the system as having only 2,400GB of addressable space, but with the remaining 3,488GB of the 5,888GB as mandatory free space, producing total free space of 59.25%. Given that free space more than doubles, the wear coefficient drops from 3.7 to 1.7, and the drive set can be expected to last twice as long.

This, however, understates wear advantage by a factor of two: because data is being written twice to pairs of drives, the Raid-10 set is accepting almost twice as many writes as the Raid-5 set is accepting. Because same-sized SuperCharger systems have more than twice the free space but only half the functional writes of Raid-10, the wear advantage of SuperCharger Raid-5 is actually four-fold. This becomes clearer when we think about data activity.

Let's consider this from the perspective of average daily write activity. Let's assume that the drives in question will receive 2TB a day of updates, and that the drives are built with 5,000 erase cycle MLC.

In the case of Raid-10, 2TB a day of data will be received for writing, but assuming the drive is fully used, this will expand to 2TB / .27 = 7.4TB a day of data written per drive. But as this 7.4TB will actually be written to each half of the Raid-10 pair of drives, the 7.4TB a day will actually result in 14.8TB of total writes and rewrites daily.

Now, let's consider the same for the SuperCharger Raid-5 with the same drive count. Here, the incoming data is also 2TB. But the free space is 59.25%. Therefore, the gross data written will be 2TB / .5925 = 3.3755TB. But this does not allow for the parity data which must also be written, so the next step in the expression is 3.3755TB / 23 * 24 = 3.52TB.

If we compare the Raid-10 writes and rewrites of 14.8TB a day with the SuperCharger writes of 3.52TB, we will see that the latter is only 24% of the former. Similarly, we can compute the projected life of the arrays. In the case of the Raid-10, this would be:

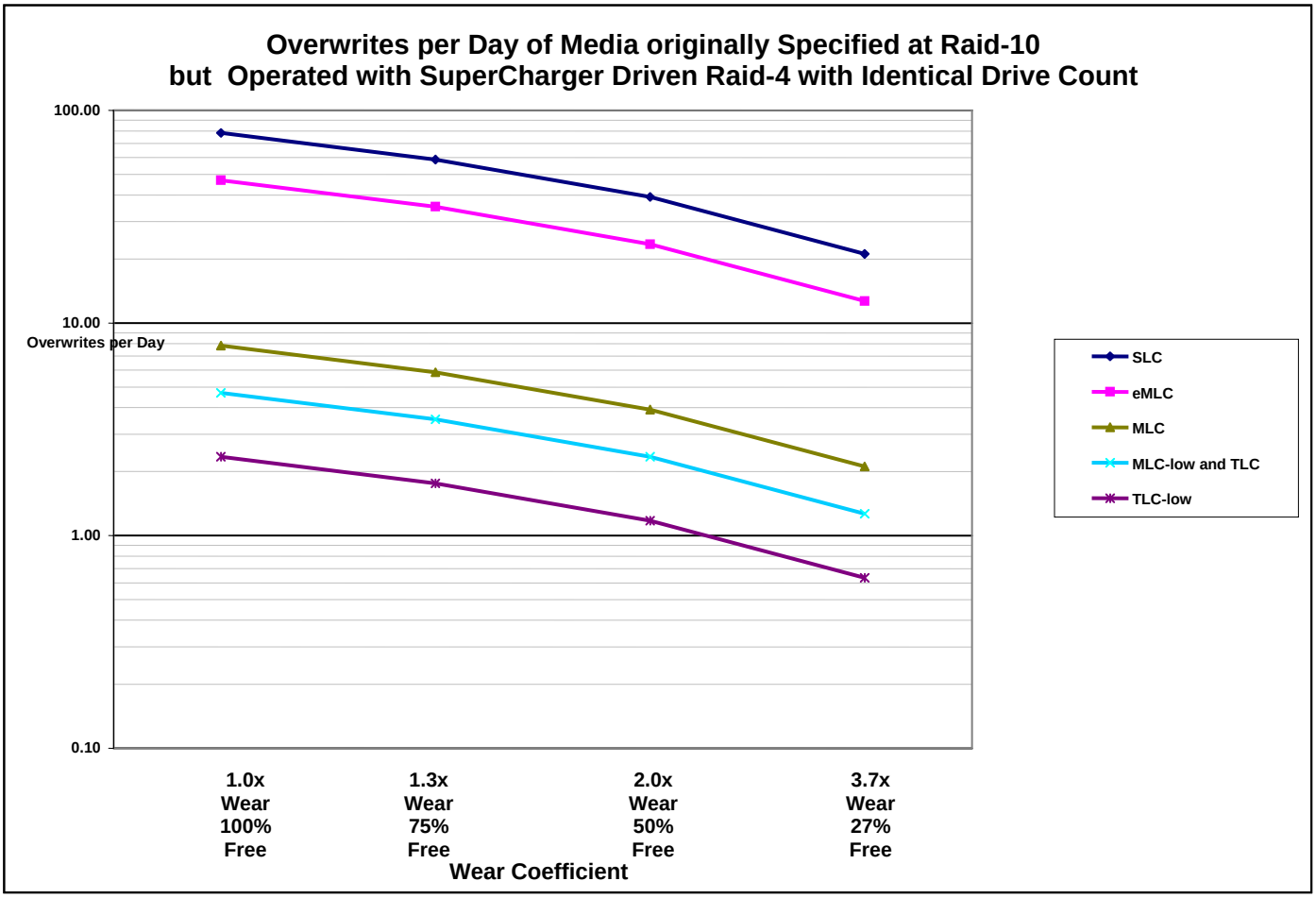
6,144 * 5000 / (14.8 * 1,024) = 2,027 days = 5.55 years

In the case of the SuperCharger Raid-5, it would be:

6,144 * 5000 / (3.52 * 1,024) = 8,522 days = 23.35 years

In specifying media, it is often more convenient to think in terms of drive overwrites per day based upon usable data size. Accordingly, we have modified the table in the last section to reflect the 4x writing improvement in SuperCharger. What we see is that with SuperCharger, even low grade TLC is a candidate for the needs of many servers while high grade 5,000 erase cycle MLC performs almost as well as the significantly more expensive eMLC and SLC. Here, we see MLC as

capable of accepting 2x overwrites. If you had 15TB of space, would you be capable of generating 30TB a day of source updates?



This inherent wear performance can be used in any number of ways. It can be used to reduce design costs and drive count. Similarly, it can be used with drive count preservation for future growth in total storage. Finally, it can be used for further write and wear performance. Here, it is important to remember that while the best case design speed is one point of measurement, in both bare drives and SuperCharger enhanced systems, effective performance will eventually fall to some function of free space. A higher proportion of free space sets a higher floor for worst case performance.

11. Understanding the Difference Between Static and Dynamic Free Space

While SuperCharger's statistical methods are advantageous, what is more generally advantageous is its unique capacity to use dynamic free space rather than being dependent upon static free space as most SSD controllers are.

Fundamentally, most SSD controllers can only use space that has never been touched by a write from the computer. As soon as the space has been touched, it ceases to be available as free space. This is why so many drives run very fast until they have been overwritten once.

Conversely, SuperCharger virtualizes any space that has been physically deleted and overwritten with zeros. You can test that behavior by running an SSD in IOMeter with SuperCharger as the driver. What you will find is that IOMeter reports "absurd" random 4KB write rates of about 3GB per second per drive (about 750,000 IOPS per drive). The reason it does so is because the default test setting for IOMeter is to write all zeros (hex '00').

Dynamic free space is extremely important because unused space on a system is often very large. On relatively young systems, it is likely to be more than half of the addressable space on a system, because most systems are ordered for future, larger, data storage needs, rather than just today's needs. The ability to use such space dramatically improves the overall write speed of a device while reducing media wear.

SuperCharger is not the only company that supports dynamic free space. Any SSD controller manufacturer who supports the Windows trim() command also supports dynamic free space to some degree. The problem with trim() is that it is only supported in Windows 7, is not supported well in Linux, and does not work through Raid sets, because the controllers don't know how to pass through the trim() command.

That said, a few controllers can support the idea of dynamic free space in other ways. For instance, SandForce can compress data. While compression is useless or counter-productive in some areas, because it is meaningless in the context of already compressed documents as well as system level encrypted files, it is useful where totally white space exists. This can be expected to compress 8:1 or perhaps even more efficiently. Even through a Raid controller, one can attain this advantage by filling up almost all of the visible free space with nulls (hex '00') and then deleting the same. The problem here is time. Filling the 30% of unallocated space on such a drive is a matter of physically writing lots of zeros. If you have a 200GB drive, collecting this free space is a 10 or 20 minute operation. But when SuperCharger's space virtualization is used, a 15 to 30 second process of writing virtual data achieves the same result with a gain of up to 14% more free space per empty sector.

12. Understanding the Impact of Locality on Wear

If you test some SSDs with CrystalDiskMark and use the default settings, you may sometimes see some totally crazy results. These devices will report that they are writing 40,000 4KB random IOPS a second. Yet, when you test these on a larger test bed you will find the drives do only a couple of thousand writes a second. This is an extreme example of locality: the drive has 128MB of RAM cache, and the area written to is 100MB of space. Thus, all the writes end up in RAM and are leisurely written to flash with what is often more than 100% write efficiency.

While this is an extreme and distorting example, locality is a practical fact that needs to be taken into account when designing systems and considering their actual speed and life. An extreme but practical example of locality is counter files, which update the same sector again and again. Here, because all free space based systems use time to their advantage, it doesn't matter whether one has one big efficient write to an erase block or a hundred updates to one sector in the block. In either case, wear is the same. The same principle also holds true in things like orders files. Today's data ends up in a concentrated location and is re-accessed and updated over and over in the fulfillment process, while the records of hundreds of yesterdays sit unchanged.

Technically, wear is not the ratio between free space and all space but rather between free space and active space. As we have seen in the examples of real world systems, locality practically increases the efficiency of free space by 10 to 20 percentage points, so that a free space level of 27% has practical wear and performance equivalent to 37 to 47%. A as our real world examples showed, sometimes it can be even more extreme.

With sophisticated algorithms, it is possible to increase this advantage even more. For instance, as SuperCharger does not rely upon physical locality at all, it is possible to take old data, perhaps unchanged for 30 days, and consolidate this into write blocks that are perfectly full and that may remain unchanged for years. This increases the average density of used erase blocks and amplifies/concentrates the effect of locality by increasing the number of blocks that are totally or near-totally empty.

We'll admit that we don't understand why, but the implication of our real world examples is that SuperCharger already naturally builds locality in its space reclamation process. The examples show two year old systems with exceptionally low coefficients of 1.3 and 1.4 in spite of having only 30% and 15% respective free space, while a much younger system using the same software has already declined to a coefficient of 2 even though it should have a value of 2.7. Such values indicate that our active process of empty erase block building, discussed in the next section, slowly concentrates inactive information together.

But while this process is possible with SuperCharger, it can't happen with non-linearized methods. When confronted with linear data, these pulse, using time to scatter data wherever it will fit, but eventually need to recollect it into a physical format so that new free space can be manufactured.

13. Understanding the Impact of Time and Wear on Speed

Optimum random write speed in an SSD controller or in SuperCharger happens when the device is able to write to a totally free erase block without having to incorporate pre-existing information from that block. In other words, the device runs faster when it only writes new data and totally avoids rewriting old data.

Conversely, in any free space environment, the worst speed will be the best speed times the percentage of available free space. From here, there are two intermediate speed possibilities.

The first is the simple improvement from locality and other factors that improve wear and effectively increase free space. For instance, when our real world examples reported wear coefficients of 1.3, 1.4 and 2.0, they were indicating that the average apparent speed was 77%, 71%, and 50% respectively of the best speed of the system.

There is one further exception that increases speed. This is to pre-manufacture fully empty space whenever the system has a quiescent or relatively quiescent period. All it does is to separate housekeeping from space utilization as new random writes. This does not increase wear. Done intelligently, it actually reduces wear by applying cleanup only to those erase blocks which have above average free space.

This is a surprisingly fast process. If a system has 30% free space on a 128GB Flash drive, the more efficiently free half of that space can normally be built into totally free space in about five minutes, and a series of drives will accumulate free space in a like proportionate way.

Thus, it is fair to say that unless a system is saturated for long periods of time, it will function at best possible speed, and because of locality, statistical technique, and other factors, it will never perform as bad as its free space quota would suggest.

These observations will be true of better made SSD controllers as well, though there are, in fact, some controllers which, when run in random IO mode, will fall to their saturated performance level and stay there.

14. The SuperCharger Test Program

In the next several sections, you will see a number of grids of results from our testing. This testing is based upon EasyCo's own proprietary test program. The results shown have been trimmed for clarity. The entire test tests the full range between 512 bytes and 4 megabytes in doubling size steps.

This test was originally constructed for internal development use because the various tests generally available have deficiencies which distort the results, or make them difficult to obtain.

For instance, many test suites will test using all null (hex '00') data, or other repetitive data. Some drive manufacturers indeed report these superior results, even though they overstate practical performance by a factor of two to four. (We don't normally discuss our trim() functionality. This writes at about 3 gigabytes a second on a single drive – about 750,000 IOPS.) Accordingly, this test uses only random data.

Similarly, this test tests, by default, at a 4GB sample size. Some tests have a default locality of only 100MB. In such an environment, drives with on-board RAM return impossible results, and all drives tend to improve performance when there is high locality of data. While it is important to also test in larger sizes, whole-drive tests tend to skew reporting in favor of smaller drives, and understate the relative performance of large drives. 4GB seems to be a useful medium avoiding both extremes.

This test only accurately reports results for new or reset drives. SNIA is developing a new protocol to test the performance of Flash media once all the free space has been filled up. EasyCo is participating in that standards process and has developed a scripting procedure and modules to consistently test towards that end. The impact of free space on performance and wear life will be discussed in greater detail below.

EasyCo's test is provided with all product deliveries and can be used to test all devices on a machine, including Hard Disks and USB sticks, with or without SuperCharger software. The source for Linux is also available upon request.

15. A Detailed Examination of Single Drive Performance

Let's begin our performance review by looking at the comparative performance of typical Enterprise and workstation class media.

The first drive compared in this case will be an OCZ Vertex 2 with a SandForce 1200 chipset, 120 billion bytes of addressable space, and with 13% dedicated free space. At the time this document was prepared, the price of the Vertex 2 was $289.

The SandForce 1200 commercial grade and 1500 Enterprise grade (called the Vertex Pro by OCZ) are identical except in three specific respects. The 1500 supports installation of a super-capacitor to assure that data buffered in the drive's memory is written to flash in the event of power loss. Next, the 1200 operates at the same random write speed as the 1500 for the first five minutes and then is crippled to not more than 10000 IOPS. This can be observed on the table below in the anomaly of a 40 thread 4KB write number that is significantly below the 10 thread number when it should normally be equal or better. The last difference is that the 1200 defaults to 7% or 13% free space, while the 1500 defaults to 27% free space. SandForce supports dial-a-yield free space selection at the drive manufacturer level. At the time of writing, the SandForce 1500 based Vertex Pro was priced at $529.

These model differences are largely irrelevant to SuperCharger as SuperCharger relies upon the linear rather than random write speed of the device, and as SuperCharger's FIFO writing method and optional TPC support largely negates the advantage of a super-capacitor.

The following is an abbreviated random read and write performance overview for this drive, performed using EasyCo's standard benchmark test methodology, previously discussed. Henceforth, we will refer the many brands of SSD based upon the SandForce chipset as the SandForce SSD.

SandForce Single SSD														
Random Read Tests							Random Write Tests							
Block Size	1 Thread		10 Threads		40 Threads		Block Size	1 Thread		10 Threads		40 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,238	16.5	22,838	89.2	22,836	89.2	4K	12,880	50.3	21,927	85.6	13,510	52.7	4K
8K	3,982	31.1	14,903	116.4	14,943	116.7	8K	10,494	81.9	11,010	86.0	11,454	89.4	8K
16K	3,522	55.0	9,064	141.6	9,060	141.5	16K	5,618	87.7	5,652	88.3	5,404	84.4	16K
32K	2,729	85.2	5,203	162.5	5,218	163.0	32K	2,815	87.9	2,818	88.0	2,703	84.4	32K
128K	1,141	142.6	1,560	195.0	1,563	195.4	128K	650	81.2	653	81.7	665	83.1	128K
512K	361	180.6	408	204.0	410	205.2	512K	146	73.2	137	68.9	130	65.1	512K
2M	97	195.5	103	206.7	106	212.0	2M	35	70.3	38	76.1	44	89.5	2M

As can be observed, the SandForce SSD is a high performance device. It can execute up to 25,000 4KB random reads a second. More importantly, the Enterprise model can reliably execute up to 20,000 4KB random writes per second, while the commercial grade model is crippled at 10,000.[2] These are impressive performance numbers. Each SandForce SSD has the random read/write performance of 50 to 100 15,000 rpm SAS drives. Each is also significantly faster than the workstation SSD next considered.

[2] SandForce advertises 'up to 60,000 random writes per second.' This is only true when the data to be written is either compressible or is duplicate data processed by their dedupe engine. While these are useful features that can improve performance even in a SuperCharger environment, we believe that all testing should be performed with random data.

Now, let's look at the same SandForce SSD run with SuperCharger. As will be noted, the performance characteristics are about the same. That's because the primary limiting factor in SandForce's performance is actually it's +/- 90MB/sec linear write speed.[3] However, you should observe that the SuperCharger drive with its linear writing is not impacted by the write IOPS caps placed by SandForce on the 1200 series controller.

SandForce Single SSD with SuperCharger														
Random Read Tests							Random Write Tests							
Block Size	1 Thread		10 Threads		40 Threads		Block Size	1 Thread		10 Threads		40 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,183	16.3	22,873	89.3	22,843	89.2	4K	18,563	72.5	18,422	71.9	18,793	73.4	4K
8K	3,924	30.6	14,942	116.7	14,932	116.6	8K	11,137	87.0	10,030	78.3	9,717	75.9	8K
16K	3,463	54.1	8,977	140.2	8,952	139.8	16K	4,487	70.1	4,489	70.1	4,958	77.4	16K
32K	2,673	83.5	5,201	162.5	5,203	162.6	32K	2,501	78.1	2,325	72.6	2,430	75.9	32K
128K	1,129	141.1	1,574	196.8	1,579	197.4	128K	612	76.5	607	75.9	615	76.9	128K
512K	362	181.2	410	205.3	415	207.8	512K	156	78.0	168	84.1	161	80.9	512K
2M	98	197.3	104	209.1	104	209.7	2M	42	84.7	42	84.0	43	86.1	2M

Now, let's take a look at a 32GB A-Data S596. This uses an Indilinx Barefoot controller, has 32 billion bytes of addressable space, and has 7% dedicated free space. At the time of this article, the 128GB version of this drive was priced at $240, less than half the cost of the SandForce 1500 though just about 20% less than the SandForce 1200. This drive will in future be referred to as the Barefoot SSD.

BareFoot Single SSD														
Random Read Tests							Random Write Tests							
Block Size	1 Thread		10 Threads		40 Threads		Block Size	1 Thread		10 Threads		40 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	5,030	19.6	15,830	61.8	15,858	61.9	4K	2,357	9.2	2,360	9.2	2,354	9.1	4K
8K	3,910	30.5	10,516	82.1	10,538	82.3	8K	2,243	17.5	2,245	17.5	2,229	17.4	8K
16K	2,727	42.6	6,235	97.4	6,265	97.8	16K	6,713	104.8	6,718	104.9	6,711	104.8	16K
32K	2,222	69.4	4,493	140.4	4,498	140.5	32K	3,306	102.8	3,290	102.8	3,305	103.2	32K
128K	1,006	125.7	1,460	182.5	1,467	183.4	128K	835	104.3	821	102.7	822	102.7	128K
512K	278	139.3	396	198.3	415	207.6	512K	204	102.4	207	103.5	211	105.5	512K
2M	85	171.3	102	205.5	112	225.7	2M	50	100.1	51	102.7	54	109.1	2M

As you can observe, the random read speed of the Barefoot is about a third slower than the SandForce, though the single thread speed is somewhat superior. Some workstation-grade SSDs have single thread random read speeds that are dramatically better than those of Enterprise devices. This can be important when designing for end-of-day routines. Conversely, the random write performance of the Barefoot SSD is intrinsically only about 17% of the SandForce SSD. Now, let's look at the Barefoot when managed by Flash SuperCharger software:

[3] Recall that we earlier mentioned that SandForce supports both lower grade and higher grade MLC, and that the higher grade has a linear write speed of 120 to 130 megabytes per second. Had the latter been used, all the write results would have improved 33% to 44%. Thus, SuperCharger with these faster drives would have delivered 27,000 4KB IOPS, and all other results would have been proportionately higher.

BareFoot Single SSD w/ SuperCharger														
Random Read Tests							Random Write Tests							
Block Size	1 Thread		10 Threads		40 Threads		Block Size	1 Thread		10 Threads		40 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,921	19.2	15,848	61.9	15,842	61.8	4K	27,834	108.7	28,044	109.5	28,003	109.3	4K
8K	3,929	30.6	10,808	84.4	10,821	84.5	8K	13,720	107.1	13,268	103.6	12,953	101.1	8K
16K	2,964	46.3	6,918	108.0	6,926	108.2	16K	6,481	101.2	6,383	99.7	6,441	100.6	16K
32K	2,245	70.1	4,586	143.3	4,590	143.4	32K	3,293	102.9	3,220	100.6	3,221	100.6	32K
128K	982	122.9	1,464	183.0	1,474	184.2	128K	798	99.7	800	100.0	801	100.1	128K
512K	302	151.3	399	199.8	408	204.3	512K	205	102.5	202	101.0	201	100.6	512K
2M	85	170.7	100	200.5	104	209.5	2M	50	100.0	50	100.7	51	103.5	2M

With SuperCharger, the 4KB random write performance of the Barefoot increases approximately twelve-fold, placing it way above the SandForce 1200. Indeed, it is actually 30% above that of the SandForce Enterprise (1500) SSD. In this case, the limiting factor of the drive is the linear write speed, which is a tad below 110MB/sec. Indilinx drives with larger form factors have higher linear performance. For instance, a 128GB Barefoot has a linear write speed of 130MB/sec due to different Flash memory topology, and hence will yield write numbers 15% better than those shown: about 33,000 4KB random writes a second. Some Workstation SSDs using other controllers can deliver linear writes in excess of 215MB/sec, and correspondingly will have SuperCharger random write speeds double those shown.

This information has been included only to show how SuperCharger best improves single SSDs used, typically, in workstation environments: by turning sows ears into silk purses. For instance, SuperCharger can turn poor performing SSDs with good linear speed into lions, and can make SDHC cards viable high speed storage media for netbooks and laptops.

16. A Detailed Examination of Raid-5, Raid-10 and SuperCharger Performance

Having looked at the raw performance numbers for both a typical workstation SSD (Barefoot) and an Enterprise SSD (SandForce), lets look at the same in the context of an 8 drive set of devices. Here, we will look at three sets of numbers for each controller brand, not just two. We will look at a hardware-only eight drive Raid-5. Next, we will look at an eight drive Raid-10 set. Finally, we will look at a Raid-5 set with SuperCharger as the driver for that set.

We begin with a set of eight SandForce SSDs configured for use as a Raid-5 set but without SuperCharger. While this configuration will support 840GB of addressable space, as you will see, the random write performance is terribly poor and asymmetric, with random writes occurring at less than a fifth the speed of random reads. The single thread random write performance is particularly bad.

SandForce Eight SSD Raid-5 WITHOUT SuperCharger

Random Read Tests							Random Write Tests							
Block Size	1 Thread		10 Threads		40 Threads		Block Size	1 Thread		10 Threads		40 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,739	18.5	42,657	166.6	80,271	313.5	4K	335	1.3	8,686	33.9	15,296	59.7	4K
8K	4,354	34.0	37,909	296.1	78,709	615.0	8K	331	2.5	6,419	50.1	11,855	92.6	8K
16K	3,776	59.0	29,048	453.8	58,674	916.7	16K	335	5.2	4,462	69.7	8,891	138.9	16K
32K	2,835	88.6	19,658	614.3	35,456	1,108.0	32K	334	10.4	3,120	97.5	5,147	160.8	32K
128K	1,704	213.0	7,286	910.8	10,132	1,266.5	128K	303	37.8	1,444	180.5	1,960	245.0	128K
512K	1,302	651.4	2,654	1,327.2	2,459	1,229.6	512K	264	132.0	764	382.2	900	450.0	512K
2M	407	815.1	626	1,253.5	629	1,259.3	2M	161	323.5	279	559.7	265	530.1	2M

You will note that the 4KB random read speed here is 80,000 IOPS, which is less than the 95,000 predicted in the discussion on controllers, while the 4KB random write rate infers 125,000 IOPS. This is due to the thread-count limitation, as can be seen in the Barefoot Raid-5 test below, run with 100 rather than 40 threads. In repeated testing, we have not been able to get the LSI 3081 above 96,000 random read IOPS on any drive set.

Now, let's look at a Raid-10 table, which is clearly more impressive.

SandForce Eight SSD Raid-10

Random Read Tests							Random Write Tests							
Block Size	1 Thread		10 Threads		40 Threads		Block Size	1 Thread		10 Threads		40 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,342	16.9	39,951	156.0	80,004	312.5	4K	10,940	42.7	36,191	141.3	37,410	146.1	4K
8K	4,004	31.2	35,850	280.0	77,681	606.8	8K	9,003	70.3	31,278	244.3	31,883	249.0	8K
16K	3,536	55.2	28,868	451.0	55,968	874.5	16K	6,877	107.4	18,294	285.8	22,201	346.8	16K
32K	2,739	85.5	20,544	642.0	34,935	1,091.7	32K	4,485	140.1	9,756	304.8	10,908	340.9	32K
128K	1,754	219.3	8,502	1,062.7	10,613	1,326.6	128K	1,728	216.0	2,726	340.7	2,631	328.9	128K
512K	1,463	731.8	2,775	1,387.6	2,960	1,480.0	512K	578	289.2	602	301.2	561	280.6	512K
2M	360	721.7	730	1,460.7	788	1,576.7	2M	154	309.0	143	287.7	158	316.7	2M

Again, we see random reads maxing out at 80,000, but random writes have risen from the 5:1 penalty of Raid-5 to just a 2:1 differential here. It should be noted that the 37,000 random write value is below the expected IOPS limit of the Raid controller, but also about what would be expected from the SandForce drive on a calculated business. Please remember that the Enterprise grade SandForce 1500 will not deliver much more in the way of random write iops

because of the controller. What has not been integrated into the above is the loss of performance when the drive has been saturated.

Now, let's look at the drive set configured Raid-5 but with the SuperCharger driver.

SandForce Eight SSD Raid-5 with SuperCharger														
Random Read Tests								Random Write Tests						
Block Size	1 Thread		10 Threads		40 Threads		Block Size	1 Thread		10 Threads		40 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,598	17.9	41,692	162.8	80,259	313.5	4K	149,515	584.0	131,643	514.2	143,338	559.9	4K
8K	4,197	32.7	35,722	279.0	59,529	465.0	8K	73,998	584.0	72,778	568.5	75,695	591.3	8K
16K	3,611	56.4	27,541	430.3	44,310	692.3	16K	32,810	512.6	32,084	501.3	31,732	495.8	16K
32K	2,772	86.6	18,558	579.9	29,931	935.3	32K	15,868	495.8	16,397	512.4	16,586	518.3	32K
128K	1,522	190.2	7,010	876.9	9,390	1,173.8	128K	4,054	506.8	4,191	523.9	4,145	518.2	128K
512K	967	483.5	2,363	1,181.8	2,116	1,058.3	512K	1,070	535.1	1,092	546.0	1,038	519.0	512K
2M	303	607.3	627	1,255.5	618	1,236.5	2M	258	517.5	259	519.0	261	522.3	2M

What is most obvious is that while the random read numbers don't change, the random write performance has jumped dramatically above the Raid-10 level, from 37,000 to 143,000. This is true in spite of the low 90MB linear write speed of the SandForce devices.

Now, let's look at the same results using an eight drive set of Barefoots. As before, we start with a Raid-5 drive set operated using standard Linux software. The one anomaly in this test is that it was run with a top end setting of 100 threads, and so the read performance jumps somewhat.

BareFoot Eight SSD Raid-5 WITHOUT SuperCharger														
Random Read Tests								Random Write Tests						
Block Size	1 Thread		10 Threads		100 Threads		Block Size	1 Thread		10 Threads		100 Threads		Block Size
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec		IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,888	19.0	37,190	145.2	95,291	372.2	4K	312	1.2	3,858	15.0	5,659	22.1	4K
8K	3,837	29.9	27,924	218.1	76,799	599.9	8K	307	2.4	3,004	23.4	4,140	32.3	8K
16K	2,657	41.5	18,486	288.8	46,388	724.8	16K	304	4.7	2,620	40.9	4,927	76.9	16K
32K	2,068	64.6	15,309	478.4	34,820	1,088.1	32K	271	8.4	1,991	62.2	1,260	39.3	32K
128K	1,608	201.1	7,555	944.4	11,967	1,495.9	128K	230	28.8	1,068	133.5	488	61.0	128K
512K	1,318	659.0	2,996	1,498.1	3,012	1,506.0	512K	209	104.5	496	248.3	274	137.0	512K
2M	469	939.3	786	1,573.0	790	1,580.5	2M	116	233.3	161	322.5	98	197.5	2M

The random write performance of this drive set is both good and bad. On the one hand, it is only about 6% of the random read speed. On the other, it is better than the expected single drive write speeds. The explanation is that the reads are proportionately so fast they improve overall efficiency.

Now for the Raid-10 test. Here, we clearly observe the design anomaly in the Barefoot. While it's 4k and 8k random write speeds are less than those of the SandForce, its 16k and above numbers match the SandForce. This implies that the Barefoot has a core block size of 16k, while others use different block sizes.

BareFoot Eight SSD Raid-10														
Random Read Tests								Random Write Tests						
Block Size	1 Thread		10 Threads		40 Threads			Block Size	1 Thread		10 Threads		40 Threads		
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec			IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,751	18.5	39,270	153.4	82,864	323.6		4K	8,433	32.9	8,520	33.2	8,669	33.8	4K
8K	3,729	29.1	28,502	222.6	59,511	464.9		8K	8,214	64.1	8,354	65.2	8,346	65.2	8K
16K	2,638	41.2	19,747	308.5	36,809	575.1		16K	5,912	92.3	21,687	338.8	21,412	334.5	16K
32K	2,072	64.7	16,080	502.5	28,071	877.2		32K	3,874	121.0	10,654	332.9	10,771	336.6	32K
128K	1,437	179.6	6,838	854.8	10,137	1,267.2		128K	2,360	295.0	2,681	335.1	2,599	324.9	128K
512K	1,171	585.5	2,493	1,246.9	2,793	1,396.9		512K	621	310.7	626	313.3	616	310.3	512K
2M	275	550.3	701	1,403.5	766	1,533.1		2M	152	305.3	158	317.3	160	321.3	2M

Finally, let's look at the same eight drive set run Raid-5 with a SuperCharger driver. Again, we see random reads pegged at 80,000. But we see 4KB random writes at 160,000. And had we been using a 128GB drive instead of a 32GB drive, we would have seen 189,000, as proven by other tests.

BareFoot Eight SSD Raid-5 with SuperCharger														
Random Read Tests								Random Write Tests						
Block Size	1 Thread		10 Threads		40 Threads			Block Size	1 Thread		10 Threads		40 Threads		
	IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec			IOPS	MB/sec	IOPS	MB/sec	IOPS	MB/sec	
4K	4,664	18.2	39,269	154.7	81,101	316.8		4K	142,775	557.7	155,731	608.3	160,070	625.2	4K
8K	3,828	29.9	29,525	230.6	60,524	472.8		8K	79,319	619.9	78,588	613.9	80,022	625.1	8K
16K	2,854	44.6	20,282	316.9	40,117	626.8		16K	40,372	630.8	40,014	625.2	40,734	636.4	16K
32K	2,184	68.2	14,328	447.7	26,285	821.4		32K	20,433	638.5	20,118	628.7	20,548	642.1	32K
128K	1,402	175.2	6,245	780.7	9,130	1,141.3		128K	5,044	630.6	5,048	631.1	5,184	648.0	128K
512K	1,012	506.1	2,704	1,352.0	2,717	1,358.9		512K	1,299	649.5	1,328	664.1	1,347	673.8	512K
2M	368	737.2	718	1,437.3	724	1,448.1		2M	339	679.7	347	694.1	342	685.7	2M

17. Comparative Costs and Performance of Raid-10 and Flash SuperCharger

Having covered a great many subjects, we now come to the point where we begin to consolidate all these results into a meaningful set of numbers and relationships.

Here, we begin by taking the table of drives and the test results from the last several sections and consolidating them. In the process, we choose to ignore simple Raid-5 performance as drive driven Raid-5 is just not performance-viable.

What we do instead is to project the single drive performance and cost numbers into a composite table that compares the cost and performance of Raid-10 with Flash SuperCharger driven Raid 5.

Here, we will begin with a very simple methodology. We will compute 4KB random reads by multiplying the single drive value by eight. We will similarly compute Raid-10 random writes by multiplying the manufacturer's value or independently tested value reported by third parties by four. Conversely, we will compute the random write rate of SuperCharger by taking the tested values and adjusting proportionately based upon reported linear write speed.

Finally, we compute the costs of storage. Here, costs for "Enterprise" devices used in a Raid-10 environment are the drive price times two divided by the manufacturer's addressable storage. We use the same methodology for commercial drives, but we adjust for the additional free space needed to bring these to a level of 27% free. As discussed in the wear section, we can do this ourselves without reference to the limits of the manufacturers. We apply the same methodologies to the SuperCharger Raid set, adjusting for seven rather than four write surfaces, and then add in the cost of the SuperCharger royalty. This works fine for the "commercial" drives but does not work for the enterprise drives. Here, we have to double-dip on free space, and this is reflected in a lower differential between Enterprise drives used Raid-10 and those used with SuperCharger.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets

	Single Drive Statistics					Eight Drive Raid-10			Raid-5 with SuperCharger		
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	4KB Read IOPS	4KB Write IOPS	Cost per Gig 27% Free	4KB Read IOPS	4KB Write IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	252,000	17,280	17.92	252,000	247,273
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	432,000	162,000	16.46	432,000	400,000
SandForce 1500	100	529	22,836	20,000	90	10.58	164,419	72,000	9.10	164,419	143,338
SandForce 1200	128	289	22,836	10,500	90	6.01	120,000	37,410	4.36	120,000	143,338
Barefoot 32GB	32	92	15,858	2,354	110	7.65	114,178	8,669	5.32	114,178	160,051
Barefoot 128GB	128	240	18,741	2,782	130	4.99	134,937	10,015	3.76	134,937	189,091
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	95,846	9,540	4.79	95,846	85,818
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	432,000	108,000	4.34	432,000	203,636
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	432,000	162,000	4.18	432,000	312,727

What this shows is that SuperCharger Generally costs less than Raid-10 per addressable gigabyte and that it also always random writes significantly faster in no load situations. This is generally true, if only because SuperCharger will always, in the case of eight drive sets, write to seven devices rather than just to four pairs.

But we next have to take the step of integrating Raid controller performance limitations into our results. This changes relative performance in a significant way. It drives down peak random reads to 95,000 IOPS from an eight drive set, while driving down peak Raid-10 random writes to just 47,500. We notice for the first time that all the drives have random read rates in excess of 95,000 while the Raid-10 random writes no longer seem as widely variable as before.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets Adjusted for Current SATA-2 Raid Controller Limitations											
	Single Drive Statistics					Eight Drive Raid-10			Raid-5 with SuperCharger		
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	4KB Read IOPS	4KB Write IOPS	Cost per Gig 27% Free	4KB Read IOPS	4KB Write IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	95,000	17,280	17.92	95,000	247,273
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	95,000	47,500	16.46	95,000	400,000
SandForce 1500	100	529	22,836	20,000	90	10.58	95,000	47,500	9.10	95,000	143,338
SandForce 1200	128	289	22,836	10,500	90	6.01	95,000	37,410	4.36	95,000	143,338
Barefoot 32GB	32	92	15,858	2,354	110	7.65	95,000	8,669	5.32	95,000	160,051
Barefoot 128GB	128	240	18,741	2,782	130	4.99	95,000	10,015	3.76	95,000	189,091
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	95,000	9,540	4.79	95,000	85,818
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	95,000	47,500	4.34	95,000	203,636
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	95,000	47,500	4.18	95,000	312,727

We also see that the random write performance spread between SuperCharger and Raid-10 increases significantly.

But we must next integrate worst case wear coefficients – with their assumption of 27% free space - into the values. This does not change random read values, but it changes random write performance significantly for most devices but not for all. For instance, both the Marvel and the SandForce 1500 numbers were previously capped by controller limitations, and their apparent performance does not decline as much. That said, in every case SuperCharger significantly outperforms Raid-10 in random writes.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets Adjusted for Current SATA-2 Raid Controller Limitations and Worst Case Wear											
	Single Drive Statistics					Eight Drive Raid-10			Raid-5 with SuperCharger		
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	4KB Read IOPS	4KB Write IOPS	Cost per Gig 27% Free	4KB Read IOPS	4KB Write IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	95,000	13,200	17.92	95,000	66,764
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	95,000	47,500	16.46	95,000	108,000
SandForce 1500	100	529	22,836	20,000	90	10.58	95,000	20,201	9.10	95,000	38,701
SandForce 1200	128	289	22,836	10,500	90	6.01	95,000	10,101	4.36	95,000	38,701
Barefoot 32GB	32	92	15,858	2,354	110	7.65	95,000	2,341	5.32	95,000	43,214
Barefoot 128GB	128	240	18,741	2,782	130	4.99	95,000	3,005	3.76	95,000	51,055
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	95,000	2,862	4.79	95,000	23,171
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	95,000	32,400	4.34	95,000	54,982
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	95,000	47,500	4.18	95,000	84,436

Finally, in the last table, we will recognize that most systems are not going to run in a saturated environment all the time, just as most will not run in a fairy tale unloaded environment. There is going to be a lot of data, but conversely there will also be locality, and SuperCharger as well as most drive controllers can construct totally free and empty erase blocks in anticipation of need … at least some of the time. Accordingly, we split the difference and assume 60% free space, and a 1.66 average wear coefficient. The general observation here is that SSDs driven by SuperCharger perform more random writes than they can perform random reads in a like period of time.

Conversely, Raid-10 systems have random write rates that are, at best, half the random read rate.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets Adjusted for Current SATA-2 Raid Controller Limitations and Average Case Wear Equivalent to 60% Free Space											
	Single Drive Statistics					Eight Drive Raid-10			Raid-5 with SuperCharger		
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	4KB Read IOPS	Average Case 4KB Write IOPS	Cost per Gig 27% Free	4KB Read IOPS	Average Case 4KB Write IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	95,000	13,200	17.92	95,000	148,364
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	95,000	47,500	16.46	95,000	240,000
SandForce 1500	100	529	22,836	20,000	90	10.58	95,000	44,892	9.10	95,000	86,003
SandForce 1200	128	289	22,836	10,500	90	6.01	95,000	22,446	4.36	95,000	86,003
Barefoot 32GB	32	92	15,858	2,354	110	7.65	95,000	5,201	5.32	95,000	96,031
Barefoot 128GB	128	240	18,741	2,782	130	4.99	95,000	6,677	3.76	95,000	113,455
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	95,000	6,360	4.79	95,000	51,491
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	95,000	47,500	4.34	95,000	122,182
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	95,000	47,500	4.18	95,000	187,636

Finally, let's look forward into 2011 and the availability of a wider range of SATA-3 Flash SSDs. The table below assumes the use of LSI's SATA-3 controller with SATA-3 SSDs of performance similar to those available today. Again, here, we see symmetry between SuperCharger and random reads, while random writes are, again, at best half the random read rate.

Comparison of Server and Workstation Raid Sets Adjusted for SATA-3 Raid Controller Limitations and Average Case Wear Equivalent to 60% Free Space											
	Single Drive Statistics					Eight Drive Raid-10			Raid-5 with SuperCharger		
Model	Size GB	Price	4kb Random Reads	4kb Random Writes	Linear Write mb/sec	Cost per Gig 27% Free	4kb Read IOPS	Average Case 4kb Write IOPS	Cost per Gig 27% Free	4kb Read IOPS	Average Case 4kb Write IOPS
Intel x25-E 32gb	64	699	35,000	4,800	170	21.84	200,000	13,200	17.92	200,000	148,364
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	200,000	97,200	16.46	200,000	240,000
SandForce 1500	100	529	22,836	20,000	90	10.58	164,419	43,200	9.10	164,419	86,003
SandForce 1200	128	289	22,836	10,500	90	6.01	120,000	22,446	4.36	120,000	86,003
Barefoot 32gb	32	92	15,858	2,354	110	7.65	114,178	5,201	5.32	114,178	96,031
Barefoot 128gb	128	240	18,741	2,782	130	4.99	134,937	6,009	3.76	134,937	113,455
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	95,846	5,724	4.79	95,846	51,491
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	200,000	64,800	4.34	200,000	122,182
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	200,000	97,200	4.18	200,000	187,636

18. Normalization of Performance Results to a 70/30 Read/Write Mix

We can now return to the table presented in Section 1 of this document. In this section, we will take the five tables presented in the prior section and reduce the read and write numbers to a single value based upon the assumption that 70% of IOPS will be random reads while 30% will be random writes. This is useful because the read/write performance of drive sets can vary dramatically from controller to controller. The formula for such normalization is,

$$R / (0.7 + 0.3 * R / W)$$

Where R = Random Read IOPS
W = Random Write IOPS

First, for form's sake, we will begin with a table which is totally useless: the normalization of raw data before adjusting for either Raid controller performance or wear based adjustments to performance. This is useless because it clouds rather than clarifies the performance question.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets

	Single Drive Statistics					Eight Drive Raid-10		Raid-5 with SuperCharger	
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	49,655	17.92	250,563
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	288,000	16.46	421,875
SandForce 1500	100	529	22,836	20,000	90	10.58	118,707	9.10	157,471
SandForce 1200	128	289	22,836	10,500	90	6.01	72,189	4.36	126,162
Barefoot 32GB	32	92	15,858	2,354	110	7.65	24,548	5.32	124,919
Barefoot 128GB	128	240	18,741	2,782	130	4.99	28,456	3.76	147,620
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	25,807	4.79	92,600
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	227,368	4.34	323,250
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	288,000	4.18	387,646

We next consider integration of Raid controller limitations, which is useful because it remains relatively easy to test drives in an unloaded scenario.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets Adjusted for Current SATA-2 Raid Controller Limitations

	Single Drive Statistics					Eight Drive Raid-10		Raid-5 with SuperCharger	
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	40,437	17.92	116,528
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	73,077	16.46	123,177
SandForce 1500	100	529	22,836	20,000	90	10.58	73,077	9.10	105,693
SandForce 1200	128	289	22,836	10,500	90	6.01	64,987	4.36	105,693
Barefoot 32GB	32	92	15,858	2,354	110	7.65	23,824	5.32	108,192
Barefoot 128GB	128	240	18,741	2,782	130	4.99	26,793	3.76	111,670
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	25,763	4.79	92,046
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	73,077	4.34	113,101
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	73,077	4.18	120,081

Next, we will look at the same results computed for worst case performance. Here, we will note that most configurations fall, but that the Marvel/Micron Raid-10 variants fall very little or remain unchanged because they are the devices most impacted by controller limitations. Never the less, we feel that this case is unrealistic. Only a very small percentage of systems are likely to be saturated 7x24 and have no locality what so ever.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets Adjusted for Current SATA-2 Raid Controller Limitations and Worst Case Wear

	Single Drive Statistics					Eight Drive Raid-10		Raid-5 with SuperCharger	
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	33,227	17.92	84,304
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	73,077	16.46	98,559
SandForce 1500	100	529	22,836	20,000	90	10.58	45,007	9.10	66,137
SandForce 1200	128	289	22,836	10,500	90	6.01	26,976	4.36	66,137
Barefoot 32GB	32	92	15,858	2,354	110	7.65	7,378	5.32	69,878
Barefoot 128GB	128	240	18,741	2,782	130	4.99	9,327	3.76	75,503
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	8,913	4.79	49,223
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	60,141	4.34	77,974
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	73,077	4.18	91,563

Next, we return to the table displayed in Section 1: the assumption that current performance estimates would be about right if we calculated speed on the basis of a typical wear coefficient of 1.66. In making this decision, we can point both to the real world examples that do at least this well, as well as the theory behind locality, and the ability to pre-construct free space whenever the system is not taxed.

Comparison of Raid-10 and SuperCharger Eight Drive Raid Sets Adjusted for Current SATA-2 Raid Controller Limitations and Average Case Wear Equivalent to 60% Free Space

	Single Drive Statistics					Eight Drive Raid-10		Raid-5 with SuperCharger	
Model	Size GB	Price	4KB Random Reads	4KB Random Writes	Linear Write MB/sec	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS	Cost per Gig 27% Free	70/30 Rd/Wt 4KB IOPS
Intel x25-E 32GB	64	699	35,000	4,800	170	21.84	33,227	17.92	106,491
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	73,077	16.46	116,031
SandForce 1500	100	529	22,836	20,000	90	10.58	71,169	9.10	92,109
SandForce 1200	128	289	22,836	10,500	90	6.01	48,230	4.36	92,109
Barefoot 32GB	32	92	15,858	2,354	110	7.65	15,374	5.32	95,307
Barefoot 128GB	128	240	18,741	2,782	130	4.99	19,120	3.76	99,874
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	18,336	4.79	75,788
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	73,077	4.34	101,794
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	73,077	4.18	111,517

Finally, we will look again at what 2011 will bring, presenting a normalized result for the LSI SATA-3 6 gigabit controller. Here, while some less-fast controllers do not improve, we generally see that performance overall should increase by about 40%.

Comparison of Server and Workstation Raid Sets Adjusted for SATA-3 Raid Controller Limitations and Average Case Wear Equivalent to 60% Free Space									
	Single Drive Statistics					Eight Drive Raid-10		Raid-5 with SuperCharger	
Model	Size GB	Price	4kb Random Reads	4kb Random Writes	Linear Write mb/sec	Cost per Gig 27% Free	70/30 Rd/Wt 4kb IOPS	Cost per Gig 27% Free	70/30 Rd/Wt 4kb IOPS
Intel x25-E 32gb	64	699	35,000	4,800	170	21.84	38,128	17.92	181,092
Marvel/Micron P-300	100	999	60,000	45,000	275	19.98	151,828	16.46	210,526
SandForce 1500	100	529	22,836	20,000	90	10.58	89,271	9.10	129,104
SandForce 1200	128	289	22,836	10,500	90	6.01	52,087	4.36	107,278
Barefoot 32gb	32	92	15,858	2,354	110	7.65	15,672	5.32	108,052
Barefoot 128gb	128	240	18,741	2,782	130	4.99	18,145	3.76	127,684
jMicron 616 512	512	1,299	13,312	2,650	59	6.75	16,746	4.79	76,164
Marvel/Micron c-300 128	128	288	60,000	30,000	140	5.99	123,007	4.34	167,916
Marvel/Micron c-300 256	256	549	60,000	45,000	215	5.70	151,828	4.18	196,123

19. Practical Extension of Conclusions into a 24 SSD Drive Set

In the following section, we are going to take the conclusions of the prior pages and build them into a 24 drive Raid set, using both Raid-10 and SuperCharger enhanced Raid-5 technology. Based upon the conclusions of prior sections, we will express the following limitations or restrictions to results:

1. We will assume that the user is always more concerned about durability and total performance. Therefore, any cost savings will be incidental (though substantial) in the quest for a better technology. This specifically means that we will use the principles outlined in Section 10 with 4x wear improvement, and addressable space equality.
2. We will discard all other drives and focus only on the Sandforce 1200/1500 series and the Marvel P-300 and C-300 series. In the case of SandForce, we will show a design based upon eMLC as well as a design based upon MLC. Building with the MLC version of the SandForce may be inappropriate in some cases as it has only 1/4th the wear life of the SuperCharger solution.
3. We will only show only advisable solutions. Therefore, we will only show the results for SuperCharger when manufactured with "commercial" drives, and will only show the Raid-10 results for Enterprise class data even if this might be under-performant as noted in (2) above.
4. We will assume that all systems are built with one LSI SAS-3 (6,000mb) for each eight drives as it has been shown that this will improve the performance of even SATA-2 arrays.
5. We will assume that systems are to be sized to the needs limits of Raid-10 systems. As a result, our tables will show a worst case performance based upon 27% free space for Raid-10 systems, and a worst case performance and cost based upon 59.25% free space for SuperCharger enhanced Raid-5 systems. The user, if desiring lower total cost or more addressable storage may presume a unit cost that can be reduced a further 45% with corresponding halving of random write performance and durability. This translates to a total throughput reduction of 25% to 35%, which is still well above the performance of the Raid-10 equivalent.
6. It needs to be noted that this is a theoretical summary and expansion to 24 drives on three controllers based upon interpolation and piecing of results from extended testing of BareFoot and SandForce drives. No direct testing of these combinations has been effected, though independent testing including testing of 24 drive sets reasonably indicates that the summation is approximately correct.

All these limitations stated, what is most compelling about these results is the broad gulf in 4KB random write performance between the SuperCharger enhanced Raid-5 configuration and the generic Raid-10 configuration. This raw random write performance gap is on the order of 4:1. While not directly explained by the phenomenon, this gap is similar to the wear gap between Raid-10 and SuperCharger enhanced Raid-5.

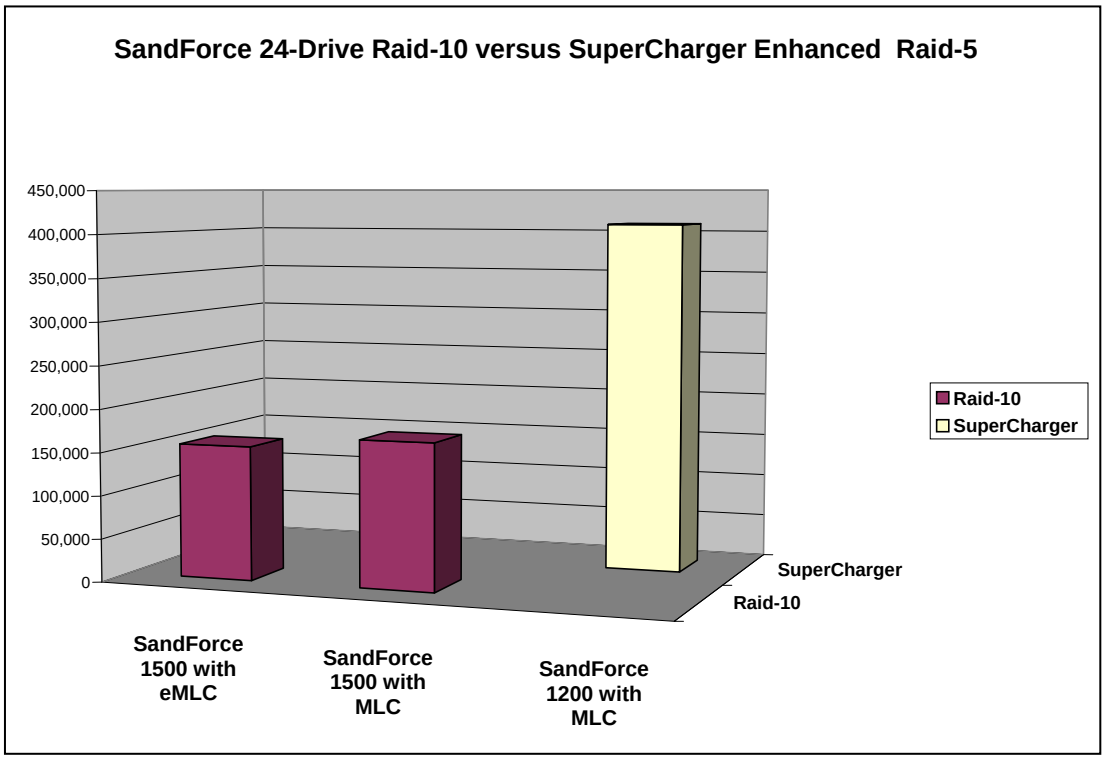
In the first table, we show the projected read and write raw performances, as well as the significantly lower build cost per gigabyte of SuperCharger enhanced solutions even when optimized for speed and durability rather than cost.

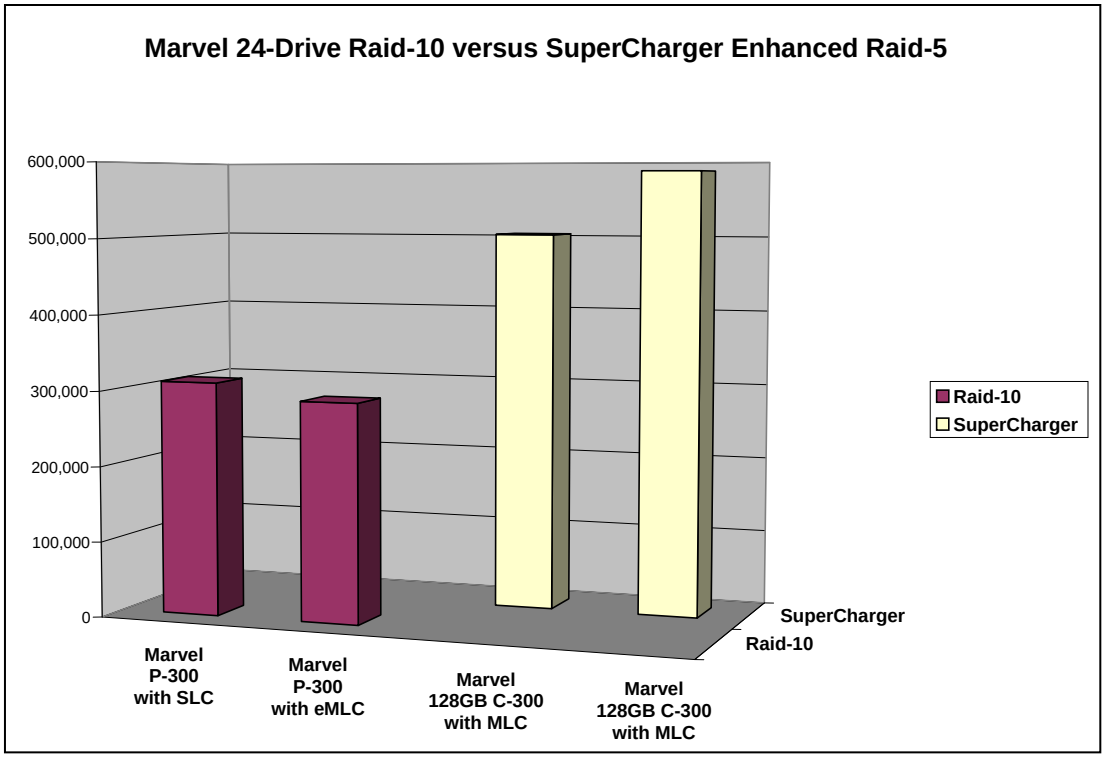
Comparison of Read and Write Performance of 24-drive Raid-10 and SuperCharger Enhanced Systems												
	Single Drive Statistics					Raid-10			Raid-5 with SuperCharger			
Model	Size GB	Price	4kb Random Reads	4kb Random Writes	Linear Write mb/sec	Cost per Gig 27% Free	4kb Read IOPS	Average Case 4kb Write IOPS	Cost per Gig 60% Free	4kb Read IOPS	Average Case 4kb Write IOPS	
SandForce 1500 eMLC	100	880	22,836	20,000	90	17.60	548,064	58,320				
SandForce 1500	100	529	22,836	20,000	90	10.58	548,064	64,800				
SandForce 1200	128	289	22,836	10,500	90				7.79	548,064	258,008	
Marvel/Micron P-300 SLC	100	999	60,000	45,000	275	19.98	600,000	145,800				
Marvel/Micron P-300 eMLC	100	850	60,000	45,000	275	17.00	600,000	131,220				
Marvel/Micron c-300 128	128	288	60,000	30,000	140				7.77	600,000	366,545	
Marvel/Micron c-300 256	256	549	60,000	45,000	215				7.47	600,000	562,909	

In the second table, we normalize the results to a 70/30 read/write mix, in order to show overall performance gain. Again we show significant performance gain as well as a significantly lower cost per addressable gigabyte. What can generally be said is that SuperCharger runs twice as fast for half the cost, and in some cases with dramatically enhanced overall endurance.

Comparison of Composite 70/30 Read/Write Performance of 24-drive Raid-10 and SuperCharger Enhanced Systems									
	Single Drive Statistics					Raid-10		Raid-5 with SuperCharger	
Model	Size GB	Price	4kb Random Reads	4kb Random Writes	Linear Write mb/sec	Cost per Gig 27% Free	70/30 Rd/Wt 4kb IOPS	Cost per Gig 60% Free	70/30 Rd/Wt 4kb IOPS
SandForce 1500 eMLC	100	880	22,836	20,000	90	17.60	155,733		
SandForce 1500	100	529	22,836	20,000	90	10.58	169,295		
SandForce 1200	128	289	22,836	10,500	90			7.79	409,840
Marvel/Micron P-300 SLC	100	999	60,000	45,000	275	19.98	310,147		
Marvel/Micron P-300 eMLC	100	850	60,000	45,000	275	17.00	289,611		
Marvel/Micron c-300 128	128	288	60,000	30,000	140			7.77	503,748
Marvel/Micron c-300 256	256	549	60,000	45,000	215			7.47	588,369

Finally, we present two tables that summarize the 70/30 read/write performance difference for the SandForce and Marvel Chipsets in graphical terms. This said, as of this writing the SandForce options are fully available, while some of the Marvel options are still in a pre-production state. Finally, it should be observed that while SandForce underperforms Marvel in this example, their soon to be released 2000 series should offset the difference.





20. Observations on Larger 72-SSD Drive Sets

It is possible to build larger storage arrays using cases with larger numbers of hot swappable 2.5" drives. This said, the returns of doing so are diminishing, no matter the technology used. Let's consider several of the general limitations.

The first and most obvious of these is the Raid controller itself. Here, the most efficient SAS-3 devices can handle up to 200,000 read IOPS per string, and up to 100,000 Raid-10 random write IOPS. But if we consider tested single drive throughput, we realize that 24 SandForce drives can deliver a soon to increase 600,000 read IOPS, while the Marvel can deliver 1,440,000 theoretical read IOPS. Similarly, though the Raid-10 write channel will be limited to 100,000 write IOPS, with only 27% free space, the 24 drive string of SandForces can deliver a soon to improve 65,000 saturated random writes, while the Marvel P-300 can deliver 144,000 at saturation, exceeding the 100,000 card limit.

SuperCharger has similar issues. While the multi-gigabyte performance of SuperCharger still offers a performance advantage in both linearity and a proportionately higher number of landing surfaces (23 versus the 12 of Raid-10 in this case) over Raid-10, the relative advantage of SuperCharger decreases. There are aggregate limits to what can be pushed through a buss and there are also limits to operating systems such as Linux. Here, EasyCo is already engaged in a project to improve Raid-5 linear write performance.

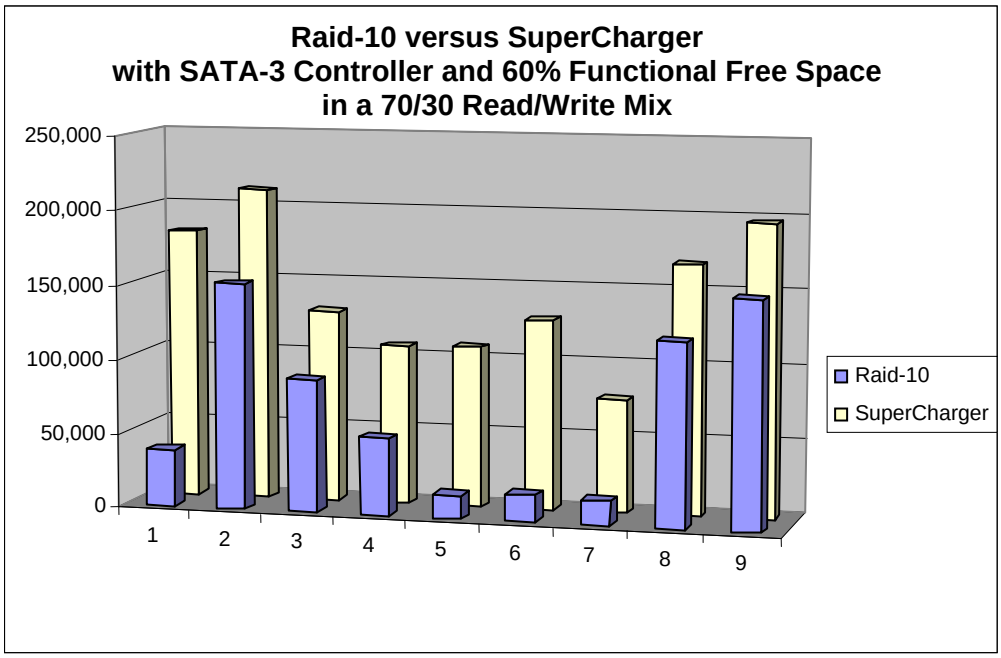
The problems discussed here also apply outside the envelope as well. 600,000 read IOPS at 4KB translates to 25 gigabits a second, well past the threshold of 10 gigabit iSCSI, and traditional Fiber Channel, while very near the practical limit of Infiniband. Accordingly, users should think about cabling, network topologies, and multiple interface cards in order to maximize throughput. In terms of aggregate IO, just as it may make sense to have four rather than three Raid controllers, so it may make more sense to have three 24-drive appliances than to have one 72-drive appliance.

These limitations expressed, SuperCharger here retains three significant advanges:

1. It outperforms Raid-10 due to superior random write performance: the ability to write simultaneously to more logical write surfaces.
2. It retains its media wear advantage, and thus drive durability and reliability advantage of 4:1 on same logical-sized surfaces. See section 10 for a discussion of this.
3. It retains a dramatic cost saving advantage of almost 2:1 on same logical-sized surfaces or 4:1 on space-maximized surfaces.

21. Concluding Thoughts

At the beginning of this paper, we looked at the relative overall performance of an eight drive Raid-10 set and the same drives with SuperCharger enhanced Raid-5 performance. Let's look at the same type of chart again with its real-world wear, but adjusted for anticipated SATA-3 performance:



What this shows, more than anything else is that whatever the media and components used, SuperCharger significantly increases overall performance. Similarly, the capacity to build more addressable storage per dollar of expense can translate to longer media life, as do other design factors, and the capacity to use less expensive media. SuperCharger today permits building 12 terabyte systems that deliver over a half million 4KB IOPS in just a 2u format. Few solutions can make such a claim.

Author: Sam Anderson
sam@easyco.com

EasyCo LLC
220 Stanford Drive
Wallingford
PA 19086 USA
Tel: (+1) 610-237-2000
888-473-7866
Email: sales@EasyCo.com
Web: http://EasyCo.com