

Demartek SSD Deployment Guide 2012-Q2

Overview

Solid state storage is transforming the entire computing industry. Solid state storage has completely revolutionized consumer electronics products, replacing spinning disk drives in virtually every category of consumer devices. This same enthusiasm for solid state storage has moved into the data center. Because of this interest, Demartek has produced this *Demartek SSD Deployment Guide 2012-Q2*, one in a series of technology deployment guides. This guide can be found on our website in our [SSD Zone](#) or by searching the Internet for “Demartek SSD Deployment Guide” using any well-known Internet search engine.

Audience

This guide is designed for managers and technical professionals within IT departments who are exploring the possible benefits of SSD technology or who are looking for actual deployment examples of SSD storage solutions.

Objectives of this Guide

The *Demartek SSD Deployment Guide* is designed to provide basic information about SSDs and practical guidance for planning and deploying SSD technology and products. This is primarily a technical document, including information about the types of solid state media that are available, form factors, deployment examples, and more.

This guide is intended to be used as a reference and is divided into sections including SSD technology areas, and specific vendor products. There are screen shots and information from actual deployments of these products. The work was performed in the Demartek lab in Arvada, Colorado, USA.

About Demartek

Demartek has its own lab, and the vast majority of the research work we do involves running hardware and software solutions in our lab with our staff. The Demartek lab is equipped with servers, network infrastructure, and storage, and supports 1 Gb Ethernet, 10 Gb Ethernet, iSCSI, Fibre Channel, FCoE, SSDs, and a variety of other technologies. Public evaluations of SSD solutions are available in our [SSD Zone](#) on our website.

Demartek produces highlight videos of public evaluations and deployment guides. Links to these videos are available on our web site and are posted on the [Demartek channel](#) on YouTube.

Table of Contents

Overview	i
Audience	i
Objectives of this Guide.....	i
About Demartek	i
Table of Contents	ii
Introduction	5
Demartek Deployment Guides.....	5
Demartek Lab Notes.....	5
Basic Storage Architectures	6
Direct Attached Storage (DAS).....	6
Network Attached Storage (NAS).....	6
Storage Area Network (SAN)	6
Unified Storage	7
Demartek Storage Interface Comparison	7
SSD Technology Overview	8
DRAM SSD Technology.....	8
NAND Flash Technology	8
NAND Flash Characteristics	9
NAND Flash and NOR Flash.....	9
NAND Flash Types: SLC and MLC.....	10
NAND Flash Endurance	11
<i>Enterprise MLC (eMLC)</i>	11
<i>NAND Flash Die Size Reduction</i>	11
<i>JEDEC Endurance Standards</i>	12
<i>Over-Provisioning</i>	13
<i>Remaining Life</i>	13
NAND Flash Performance	14
<i>Preconditioning</i>	14
NAND Flash Cost	15
NAND Flash Media Sanitization	16
SSD Form Factors	17
SSD-specific Form Factor	18

mSATA	18
μSSD.....	18
2.5-inch PCIe®	18
Disk Drive Form Factor	19
<i>Dimensions</i>	19
<i>Hybrid Drive</i>	19
<i>Interfaces</i>	20
PCIe Card Form Factor	21
<i>Dimensions</i>	22
<i>Capacities</i>	22
<i>Power</i>	22
DIMM Form Factor	23
Traditional Disk Arrays	24
<i>Demartek Comments on SSD Adoption</i>	24
All-flash Arrays	25
<i>Performance</i>	25
<i>Host Interfaces</i>	25
<i>Power Consumption</i>	25
Caching Appliances.....	26
Data Placement.....	27
SSDs as Primary Storage	27
<i>Automated Tiering Solutions</i>	27
<i>Chunk Size</i>	28
SSD Caching.....	29
<i>SSD Caching Workloads</i>	30
<i>Write-through Cache vs. Write-back Cache</i>	30
<i>Read/Write “Flip”</i>	31
<i>Multiple SSD Caching Solutions</i>	31
SSD Caching vs. Tiering – Demartek Opinion.....	31
Operating System Behavior with NAND Flash SSDs	32
Trim.....	32
UNMAP	32
Garbage Collection in External Storage Systems.....	33
Test Environment for this Guide	34

Workload Tests	34
<i>Synthetic Workloads</i>	34
<i>Real-World Workloads</i>	34
Vendor Products Tested for this Guide	34
Server Specifications	35
Network Infrastructure	35
FlashSoft SE SSD Caching Software	36
Managing SSD Cache with FlashSoft SE	37
Tests run with FlashSoft	38
<i>IOmeter on Windows</i>	38
<i>TPC-C like Workload and Different Cache Sizes</i>	39
Nimbus Data S-Class All Flash Array	40
Tests run with Nimbus Data	41
OCZ PCIe and SATA SSDs	42
Tests run with OCZ SSDs	43
<i>Configuration 1 – PCIe SSD vs. 8Gb Fibre Channel Storage</i>	43
<i>Configuration 2 – Drive Form Factor</i>	44
SMART Storage Systems SSDs	46
Tests run with SMART Storage Systems XceedStor 500S SSDs	47
<i>TPC-C like Workload Results</i>	47
<i>Exchange Jetstress Results</i>	47
Legal and Trademarks	48

Introduction

Solid state storage (SSS) has captured the interest of many in the Information Technology (IT) field due to the promise of substantially increased application and system performance. However, this technology is available in a wide variety of types and form factors, which can be confusing, or at least require some explanation.

The *Demartek SSD Deployment Guide* provides an explanation of the types of solid state drive (SSD) technology available today and examples of how they can be deployed in the enterprise. The primary focus of this guide is for enterprise SSD technologies, although this guide also discusses consumer SSD technologies.

One could make a case that the terms “solid state storage” and “solid state drive” do not have the exact same technical meaning. However, through common use, these two terms have come to have the same basic meaning and are used interchangeably throughout this document.

Demartek Deployment Guides

This guide is part of a series of **Demartek Deployment Guides** that provide technical background and deployment guidance for a variety of information technologies. These guides are available on the Demartek website (www.demartek.com) and can be found by searching for “**Demartek Deployment Guide**” in any major Internet search engine.

The most current version of the *Demartek SSD Deployment Guide* is available at www.demartek.com/Demartek_SSD_Deployment_Guide.html on the Demartek website.

Demartek Lab Notes

To be notified when new Deployment Guides and lab validation reports become available, you can subscribe to our free monthly newsletter, [Demartek Lab Notes](#), available on our website.

Basic Storage Architectures

SSD technology is used primarily in a computer storage context, although other contexts are sometimes appropriate. As background, we provide an overview of the basic storage architectures in use today. SSDs can be deployed in any of these storage architectures. In subsequent sections we discuss different form factors used for SSDs and specific ways to use SSD technology with respect to data placement, including SSD caching and automated storage tiering using SSDs.

Direct Attached Storage (DAS)

Direct Attached Storage (DAS) is probably the most well-known form of computer storage. In a DAS implementation, the host computer has a private connection to the storage, and almost always has exclusive ownership of the storage. The host computer accesses the storage in a “block” fashion, which means that it directly addresses blocks on the storage device. This implementation is relatively simple and usually inexpensive, depending on the storage technology selected. Potential disadvantages are that the distance between the host computer and the storage are frequently short, such as inside a computer chassis or within a rack or adjacent rack. Some DAS implementations require that the host computer be taken offline when adding or removing storage devices, such as a boot drive directly connected to a motherboard storage interface. SATA is a common DAS interface. SSDs can be used in DAS storage architectures.

Network Attached Storage (NAS)

Network Attached Storage (NAS) devices, also known as file servers, share their storage resources with clients on the network in the form of “file shares” or “mount points.” The clients use network file access protocols such as CIFS/SMB or NFS to request files from the file server. The file server then uses block protocols to access its internal storage in order to satisfy the requests. Because NAS operates on a network, the storage can be very far away from the clients. Many NAS solutions provide advanced features such as snapshot technologies, global namespace, SSD caching and more.

Storage Area Network (SAN)

SAN architecture provides a way to use block access methods over a network such as Ethernet or Fibre Channel to provide storage for host computers. The storage in a SAN is not owned by one server but is accessible by all of the servers on the network. This SAN storage can be carved into logical storage pools or volumes that can be assigned to particular host servers. These logical volumes are independent of the geometries or components of the storage hardware. The storage appears to host servers and applications in the same way that DAS storage appears, but because SAN storage uses a network, storage can be a long distance away from the host servers.

SAN architectures use block Small Computer System Interface (SCSI) protocol for sending and receiving storage data over their respective networks. Fibre Channel (FC) SANs implement the SCSI protocol within the FC frames. Internet SCSI (iSCSI) SANs implement the same SCSI protocol within TCP/IP packets. Fibre Channel over Ethernet (FCoE) is a newer interface that encapsulates the Fibre Channel protocol within Ethernet packets using a relatively new technology called Data Center Bridging (DCB). DCB is a set of enhancements to traditional Ethernet, and is currently implemented with some 10GbE infrastructure. InfiniBand can also be used as a SAN interface by using SCSI RDMA Protocol (SRP). Because each of these technologies allow applications to access storage using the same underlying SCSI command protocol, it is possible to

use all of these technologies in the same enterprise, or to move from one to the other. Generally speaking, applications running on a host server cannot tell the difference between Fibre Channel SAN storage, FCoE SAN storage, and iSCSI SAN storage. In fact, applications generally cannot tell the difference between DAS storage and SAN storage. SSDs can be used in any of these SAN storage architectures.

There is more to choosing a storage system than selecting the host interface. Regardless of the type of interface, several other factors need to be considered, including the number and type of disk drives (including SSDs), management software, advanced features, support from the vendor, and several other factors. Advanced features of modern storage systems may include various forms of replication, thin provisioning, compression, data de-duplication, SSD caching, automated storage tiering, and others.

Unified Storage

Unified storage combines NAS and SAN technologies into a single, integrated solution. These unified storage solutions provide both block and file access to the shared storage environment. These often provide simplified management by combining the management of all storage, regardless of the transport, or “plumbing,” into a single management console.

Demartek Storage Interface Comparison

We have compiled a reference page on our website that compares several interfaces used for storage applications. This page is a vendor-neutral page that provides technical information regarding these storage interfaces. We update this page periodically. The storage interfaces in this comparison include:

- ◆ FC – Fibre Channel
- ◆ FCoE – Fibre Channel over Ethernet
- ◆ IB – Infiniband
- ◆ iSCSI – Internet Small Computer System Interface
- ◆ PCIe – PCI Express
- ◆ SAS – Serial Attached SCSI
- ◆ SATA – Serial ATA
- ◆ USB – Universal Serial Bus

The contents of this page include the following sections:

- ◆ Acronyms
- ◆ Storage Networking Interface Comparison Table
- ◆ Transfer Rate, Bits vs. Bytes, and Encoding Schemes
- ◆ History
- ◆ Roadmaps
- ◆ Cables: Fiber Optics and Copper
- ◆ Connector Types
- ◆ PCI Express (PCIe)

This page can be found at the following location or by entering “Storage Interface Comparison” in any of the major Internet search engines:

- ◆ www.demartek.com/Demartek_Interface_Comparison.html

SSD Technology Overview

Solid state storage devices are computer storage devices that use memory technology for the storage media rather than traditional magnetic media such as hard disk drives (HDD) or tape drives. These SSS devices can be made with either DRAM technology or Flash memory technology, or sometimes with both. These devices appear to the host operating system as storage devices, and can be used in any storage architecture, such as DAS, NAS, or SAN.

SSD devices have no moving parts and offer very fast performance. In the enterprise marketplace, these storage devices can sustain very high transaction rates and support large numbers of users.

DRAM SSD Technology

DRAM can be used in storage devices to deliver the highest rates of input/output (I/O). These DRAM SSD systems are expensive and found in industries that must maintain extremely high transaction rates and very low latencies such as financial, telecom, e-commerce, and others.

DRAM SSDs use the same type of memory that is found in enterprise servers. Because DRAM stores information only when it has electrical power, these DRAM SSD systems usually provide battery backup, some flash memory, a hard disk for data backup or combinations of these technologies to insure that no data is lost. In addition, like in enterprise servers, DRAM SSDs usually provide ECC-protected memory which delivers additional data protection. Just like in enterprise servers, this type of memory is more expensive than the non-ECC memory usually found in desktop and laptop computers.

Storage administrators often measure storage performance in I/Os per second (IOPS), and DRAM-based SSDs can deliver hundreds of thousands or sometimes millions of IOPS in one system. In addition, latencies, or round-trip times for these DRAM SSD systems are often measured in microseconds, which are much faster than HDD-based storage systems.

DRAM SSD systems can be used as an external cache in front of other storage systems to accelerate performance of existing storage systems. DRAM is also often used as a cache within large storage systems.

NAND Flash Technology

NAND flash memory was invented by Toshiba in 1987, making this year the 25th anniversary of NAND flash. NAND flash memory can be used in storage devices and systems to deliver very high I/O rates. NAND flash provides very high IOPS, generally in the range of tens of thousands to hundreds of thousands of IOPS, depending on the implementation. These rates are not quite as high as DRAM-based systems, but are generally much higher than HDD-based systems. Latencies are also very good, often measured in microseconds or very low milliseconds. NAND flash is quiet, consumes low amounts of electric power, has low-weight, and produces less heat than HDDs.

NAND flash SSD systems, like DRAM SSD systems, can be used as an external cache in front of other storage systems, including DAS, NAS and SAN storage systems. NAND flash can also be used as a cache within larger storage systems. We discuss SSD caching in more detail in the [data placement section](#) below.

NAND Flash Characteristics

NAND Flash and NOR Flash

NAND flash is a type of EEPROM (electrically erasable programmable read-only memory), along with NOR flash. Both of these types of memory are non-volatile which means that they can retain data even if the power is turned off. Data is stored by electrically erasing and “reprogramming” data areas on the flash media. Individual bytes can be addressed (read and written) with NOR flash, but it contains fewer bits per square inch, so it is physically larger than NAND flash. NAND flash is more dense (more bits per square inch) but the data can only be addressed in blocks, or pages, typically 4KB or 8KB today.

As a result, NOR flash is typically found where small amounts of data, measured in megabits (Mb), need to be stored and possibly changed, such as low-cost mobile phones and small consumer appliances. It can also be found on desktop and server motherboards for portions of the BIOS, small industrial devices, televisions, and consumer gaming devices.

NAND flash is used where larger quantities of non-volatile memory are required, such as smart phones, tablet computers, notebook computers, consumer and enterprise SSDs, USB flash drives, high-end televisions, and others. Some devices may have both NOR flash and NAND flash in the same device. This guide focuses on SSDs, which are primarily a NAND flash type of application.

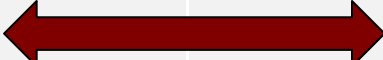




NAND Flash Types: SLC and MLC

NAND flash is available in single-level cell (SLC) and multi-level cell (MLC) varieties. SLC flash stores one bit per cell, and provides faster performance and longer endurance than MLC flash. MLC flash stores two or more bits per cell, and as a result, has higher storage capacity than SLC flash. MLC flash is available in 2-bit per cell, 3-bit per cell and 4-bit per cell varieties. Sometimes, 3-bit per cell MLC flash is called TLC for triple-level-cell.

The page size for SLC is typically 4KB. The page size for MLC today is either 4KB or 8KB. As the die size continues to shrink to 20nm and below, these page sizes will most likely double.

As with many other technologies found in Information Technology, there are tradeoffs to be made, and NAND flash types are no different. As there are many vendors producing many SSD products, the chart below has no absolute numbers, but provides a relative, sliding scale across the categories.

Table 1 - SLC and MLC Flash Characteristics

	SLC	MLC-2	MLC-3	MLC-4
Bits per cell	1	2	3	4
Performance	Fastest			Slowest
Endurance	Longest			Shortest
Capacity	Smallest			Largest
Error Prob.	Lowest			Highest
Price per GB	Highest			Lowest
Applications	Enterprise	Enterprise / Consumer	Consumer	Consumer

NAND Flash Endurance

NAND flash cells wear out due to the way that writes occur on NAND flash. NAND flash devices use the “program-erase cycle” in order to accomplish writes. This is also sometimes known as the “write cycle.” Technically, this “wear-out” means that these cells lose their ability to hold a charge for a long enough period of time to be useful. When data needs to be written to a NAND flash page, the page must be empty. If there is data on that page, the low-level flash controller must read the existing data from the page, erase the page, merge the new data with the existing data and then write the page. All of this takes place on the device at a lower level than what the user can see.

Different grades of NAND flash have different life expectancies. These life expectancies are based on the number of bits per cell and the die size of the NAND flash technology. These life expectancies are expressed in write cycles, or number of program-erase cycles per bit.

- ◆ SLC: Usually 100,000 write cycles
- ◆ MLC-2: 3,000 – 10,000 write cycles (also see “Enterprise MLC” below)
- ◆ MLC-3: 300 – 3,000 write cycles

For enterprise applications, where the number of write operations is very high (such as OLTP database applications, email applications and others), SLC flash provides the endurance and performance that meets enterprise requirements. For consumer applications (such as USB flash drives, digital cameras, etc.), capacity is often the most important characteristic, so various forms of MLC flash are suitable.

Enterprise SSD vendors usually provide a TBW (total bytes written or terabytes written) value per day that the device can sustain, to provide an indication of the expected life of the SSD for a given application workload and SSD capacity. This may also be expressed as the number of times the full capacity can be written per day during the expected life, or warranty period, of the device.

SSDs intended for consumer applications may provide an estimated total terabytes written over the expected life of the device, such as five-years. A daily write limit could be calculated from the lifetime figure. For example, if a 128 GB consumer SSD is rated for 80 terabytes of writes over five years, then the average daily write limit would be approximately 44 GB of new data, or 34% of the total capacity of the device.

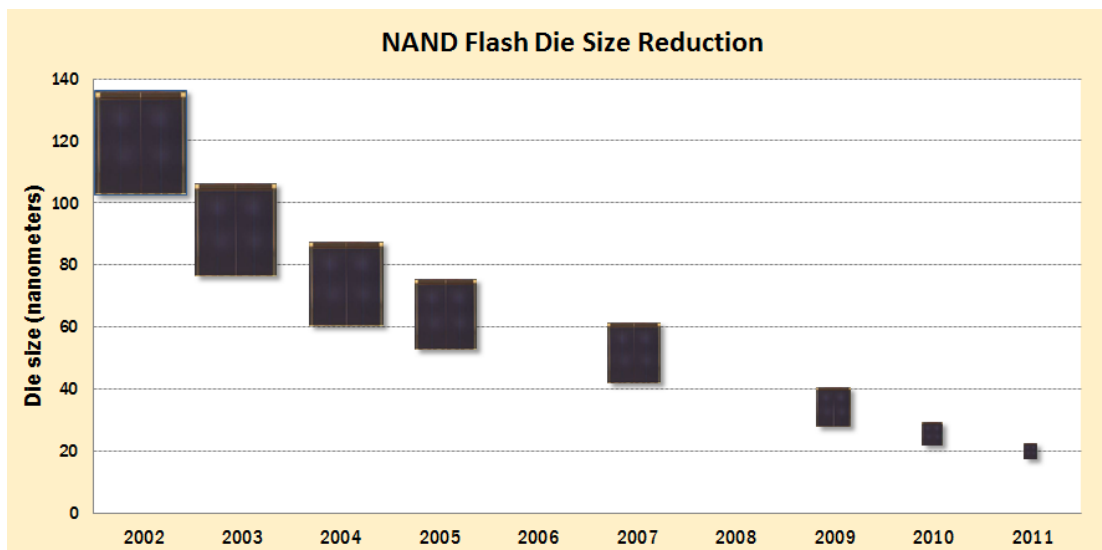
Enterprise MLC (eMLC)

Recently, a new type of NAND flash was introduced called “Enterprise MLC” or simply “eMLC.” By using advanced low-level flash controllers and other techniques, some SSD manufacturers have taken regular MLC-2 flash and extended the number of write cycles by a factor of approximately ten, resulting in a lifetime of 20,000 – 30,000 write cycles per bit. A more technically correct name for today’s eMLC is probably “Endurance MLC”, as only the endurance characteristics of MLC-2 are typically enhanced. The performance is generally no better than regular MLC-2. This type of eMLC flash is more suitable for enterprise applications than regular MLC-2.

NAND Flash Die Size Reduction

Over the last few years, NAND flash die sizes have shrunk. This die size reduction has advantages and disadvantages that are related to the physics of NAND flash. As the die size shrinks, capacities increase, but endurance decreases. For MLC flash, as the die size shrinks, the number of write

cycles moves to the lower end of the ranges shown above. For consumer applications, endurance becomes relatively less important as density and capacity increase. Also, as the die size shrinks, the standard flash page size will increase.



JEDEC Endurance Standards

The [Joint Electron Devices Engineering Council](#) (JEDEC) develops standards for the microelectronics industry, and specifically for memory technologies. JEDEC developed two standards relating to testing SSDs for endurance, JESD218A and JESD219. In these standards, SSD endurance classes and requirements are provided.

The two general categories of SSDs are:

- ◆ **Client** – SSDs that would be installed in client computers and similar devices that primarily serve a single user.
- ◆ **Enterprise** – SSDs that would be installed in servers that host enterprise applications or serve multiple users.

Table 2 – JEDEC SSD Endurance Classes and Requirements

Application Class and Workload	Active Use (power on)	Retention Use (power off)	Functional Failure Rqmt. (FFR)	UBER
Client	40° C/104° F 8 hrs/day	30° C/86° F 1 year	≤3%	≤10 ⁻¹⁵
Enterprise	55° C/131° F 24 hrs/day	40° C/104° F 3 months	≤3%	≤10 ⁻¹⁶

Just as there are desktop and enterprise HDDs, there are also two types of SSDs. One of the differences between client and enterprise devices is the number of hours per day of operation for

which they are designed and rated. Client SSDs (and desktop HDDs) are only expected to be operational for approximately 8 hours per day. Enterprise SSDs (and enterprise HDDs) are designed to be operated 24 hours per day. In addition, the maximum operating temperatures for the two types of SSDs are shown in Table 2.

Shelf-life, or “retention use,” for SSDs refers to the length of time and temperature conditions for which the SSD must be able to retain data when the SSD has been removed from the host computer and no power is applied to the device. The time periods specified in the table may seem relatively short compared to other types of storage devices. This is because NAND flash SSDs are designed to be online high-performance devices and not long-term, offline archive devices.

UBER is the uncorrectable bit error rate. This is a measure of the rate of occurrence of data corruption that is equal to the number of data errors per bit read, after applying any specified error-correction method. These values are similar to those found in HDDs.

$$UBER = \frac{\text{number of data errors}}{\text{number of bits read}}$$

Over-Provisioning

Most SSD products provide more NAND flash capacity than is advertised. This is done so that the low-level flash controller can manage the number of writes occurring, attempting to spread the writes evenly across all of the available flash memory, extending the overall life of the product. The additional NAND flash capacity is used and remapped as needed. This mapping of logical to physical blocks on the NAND media is performed in order to achieve the best NAND performance and endurance. The amount of extra, unadvertised NAND flash capacity varies by product and can be up to 10%, 20%, 30% or more additional capacity. Larger amounts of additional capacity allocated to over-provisioning result in longer overall life of the product. Some SSD products indicate the amount of over-provisioning capacity available.

Some SSD products allow the user to adjust the amount of over-provisioning capacity. The total capacity is divided into a user accessible capacity and the over-provisioning capacity. Increasing one of these types of capacity decreases the other.

Remaining Life

Most SSDs provide a way to interrogate the device to determine the amount of remaining life in the SSD. SSDs provide data in the Self-Monitoring, Analysis, and Reporting Technology (S.M.A.R.T.) information that indicate the amount of remaining life of the device. However, different vendors report this media wear indicator (MWI) or “wearout” data differently. Many of these vendors provide a utility program to interrogate the remaining life of the device.

NAND Flash Performance

NAND flash SSDs have much higher performance for most operations than HDDs. Most NAND flash SSDs can have 10,000 – 250,000 IOPS per device, depending on the type of NAND flash and the form factor. Some of the newest implementations of NAND flash SSDs may get as high as 1 million IOPS. By contrast, enterprise HDDs typically provide approximately 100-200 IOPS per device, and desktop HDDs typically provide less than 100 IOPS per device.

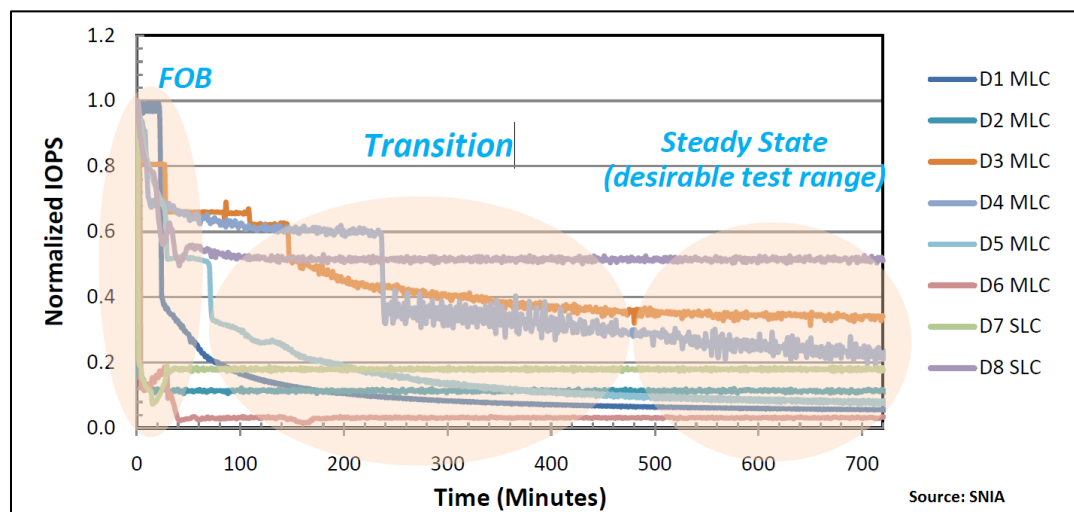
With basic NAND flash, writes are generally slower than reads, so NAND flash media is said to be “asymmetrical” with respect to read and write speeds. Modern enterprise NAND flash controllers can mitigate this disparity using advanced flash management techniques to improve the write speed of NAND flash.

Preconditioning

One of the observations many have seen (including us in our lab) is the difference in performance between “Fresh Out of the Box” (FOB) and sometime later when the SSD has reached “steady state” performance. When SSDs are new from the factory, they exhibit very high performance levels, primarily because all of the flash pages are empty, so the normal program-erase cycle has nothing to erase, enabling all of the initial writes to be performed relatively quickly.

The [Storage Networking Industry Association](#) (SNIA) has published its Solid State Storage (SSS) [Performance Test Specification](#) (PTS) in two versions, Enterprise and Client. These documents define a set of comprehensive test suites and methodologies to measure the performance characteristics of SSDs.

The performance patterns for different SSDs as they move from FOB through the transition period to steady state performance have been observed. The SNIA SSS PTS defines “pre-conditioning” tests that move an SSD from FOB to steady state. The goal is to make sure that performance measurements that users take for various workloads are taken after the device has settled into its steady state performance, rather than the short-lived elevated performance levels found with fresh out of the box SSDs. The steps needed to pre-condition an SSD are described in the PTS documents.



NAND Flash Cost

When viewed simply by price per gigabyte, SSD technologies seem a bit expensive. However, there are other metrics that should be considered when examining SSDs. The chart below provides some of these metrics. These figures are for individual HDD or SSD devices.

	\$/GB	\$/IOPS	IOPS/watt
SSD (SLC)	\$6 - \$40	\$0.003 - \$0.18	1000 - 15000
SSD (MLC)	\$0.95 - \$7	\$0.004 - \$0.05	1000 - 15000
HDD (enterprise)	\$0.50 - \$2	\$1 - \$3	10 - 30
HDD (desktop)	\$0.07 - \$0.37	\$1 - \$4	10 - 40

Prices (US\$) for “factory new” products sampled in early April 2012 and are subject to change.

When considering price per IOPS or IOPS per watt, SSDs provide significantly greater value than HDDs.

The summary of this chart is:

- ◆ SSDs are *dollars per gigabyte* and *pennies per IOPS*.
- ◆ HDDs are *pennies per gigabyte* and *dollars per IOPS*.

Pricing notes:

- ◆ Prices at the lower end of the HDD desktop price range are for the slower “green” drives.
- ◆ HDD pricing is 5% - 15% higher than in 2011 due to the Thailand flooding that occurred in late summer and autumn of 2011. These higher prices are expected to remain through most or all of 2012 or until full production resumes.
- ◆ Seagate suffered little or no damage to their factories in Thailand, while Western Digital suffered serious damage. However, the disk drive supply chain for both companies was affected. Product shortages are expected to have a more limited impact on high-end enterprise HDDs.
- ◆ SSD prices include disk-drive form factors and PCIe SSD form factors.
- ◆ Prices for eMLC SSD devices are at the upper end of the MLC price range.
- ◆ PCIe SSDs tend to have higher price per GB than the disk drive form factor, but also have higher performance.

NAND Flash Media Sanitization

There are many reasons to want to securely erase data from NAND flash SSD devices. In many environments, there are regulations that compel organizations to completely remove data deemed to be sensitive from storage devices before repurposing that equipment elsewhere within the organization, returning leased equipment, or selling or donating that equipment. Even for those organizations not compelled by regulation to clear or purge data from storage devices, it is usually prudent to insure that residual data is actually removed from these storage devices. There are too many examples of data breaches that have become publicly known simply because of sloppy or careless practices. Public embarrassment has been the least of the problems encountered by these organizations that did not properly erase data from storage devices.

The process of completely removing data from storage media is sometimes known as clearing or purging the devices. The goal is to thoroughly sanitize the storage media so that no data thought to be deleted is able to be recovered. Some organizations refer to “data remanence” as residual data that remains on storage media after insufficient purging procedures were taken that are appropriate for that storage media type.

For SATA interface storage devices, including HDDs and SSDs, there is a “security feature set” that can be implemented by the manufacturer. Included in this security feature set are commands known as “security erase” that are intended to erase all user data from all locations of the device. For NAND flash SSDs, this includes all of the over-provisioning areas (spare pages) of the device. Implementations of the ATA secure erase command vary by SSD manufacturer.

Some SSDs encrypt data as it is written to the flash media. One way of securely erasing these SSDs is for the secure erase commands to simply zero out (“zeroize”) the encryption key, which renders any data stored on the device as encrypted with no key for decryption. This process takes very little time for the device to complete. This feature is available primarily for SATA interface SSDs, but some SAS interface SSDs are beginning to implement this feature.

Some SSDs do not encrypt data as it is written, and for these, there does not seem to be consensus on implementing a secure erase function. Certainly a brute force method could be used to overwrite all the data pages with certain data patterns, but any such method would have to insure that it also overwrote all the over-provisioning areas.

For SSDs that use some of the newer SAS, PCIe and related interfaces, there is work that remains to be completed. The industry standards bodies and individual SSD vendors are working on this issue. Various governmental and commercial agencies are evaluating the methods available in SSD products today to determine which methods are acceptable. Demartek is following this issue and will provide additional commentary on this topic as it becomes available.

The US National Institute of Standards and Technologies (NIST) published *NIST Special Publication 800-88* in September 2006 that addresses storage media sanitization for a variety of storage media types, including NAND flash devices. This publication addresses data purging as well as media destruction for situations where simply purging the data does not provide adequate security. This publication is available at http://csrc.nist.gov/publications/nistpubs/800-88/NISTSP800-88_rev1.pdf.

SSD Form Factors

SSDs can be implemented in several form factors. Although either DRAM or NAND flash could be used for these, NAND flash is the most common type of SSD technology used today. These form factors include:

- ◆ SSD-specific – Newer form factors specifically designed for SSDs.
- ◆ Disk drive – Any of the 3.5-inch, 2.5-inch, 1.8-inch, or other sizes of disk drives.
- ◆ PCIe card – A PCI-Express® card with SSD technology mounted directly on it.
- ◆ Memory slot – NAND flash can be mounted on 240-pin DIMMs along with a SATA or other storage interface to provide storage in unused DIMM sockets.

In addition to the individual device form factors, there are storage system solutions that incorporate SSD technology in them. These include:

- ◆ Traditional disk arrays – Use SSDs in place of some of the HDDs normally provided.
- ◆ All-flash arrays – Use only SSDs and have no HDDs.
- ◆ Caching appliances – Operate as an external cache between servers and storage, and are available for SAN or NAS deployments.

SSD-specific Form Factor

mSATA

In 2009, the [Serial ATA International Organization](#) (SATA-IO) introduced the mini-SATA (mSATA) interface connector. This interface was originally intended for small form factor HDDs and SSDs that would be used in portable devices. Currently, only SSDs are available in this form factor. Although the mSATA interface resembles a mini-PCIe interface, these are different from each other and not compatible with each other. The mSATA interface currently supports 1.5 Gb/s and 3.0 Gb/s transfer rates. Some newer devices support 6.0 Gb/s. Motherboards began to appear with mSATA ports (or connectors) in late 2011. The physical dimensions of typical mSATA devices are included with the disk drive form factor (Table 3) below.



μSSD

In 2011, a new drive form factor was introduced that is specific for SSDs. The SATA-IO introduced SATA μSSD specification for embedded SSDs. These devices do not have the traditional SATA interface connector, but use a single ball grid array (BGA) package that can be surface mounted directly on a system motherboard. These SATA μSSD devices are intended for mobile platforms such as tablets and Ultrabooks™, and consume less electric power than traditional SATA interface devices – as low as 10 mW in slumber mode. These devices support the 6 Gb/s SATA III interface, although individual devices may not be capable of these transfer rates. These devices appear to the operating system as a standard SATA SSD.



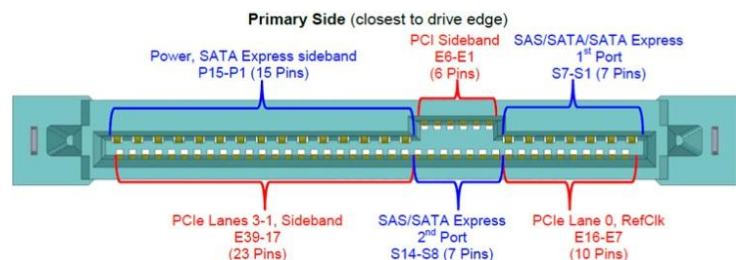
2.5-inch PCIe®

The [SSD Small Form Factor \(SFF\) Working Group](#) has developed a PCI-Express® interface that fits into the existing 2.5-inch disk drive form factor (see Table 3 below for disk drive form factors) with a combined PCIe/SAS/SATA interface that is backward compatible with SAS interface drives.

This SSD-specific form factor combines a high-performance PCIe interface with a hot-swappable 2.5-inch form factor. It uses combination connector known as SFF-8639.



One of the first shipping products to use this SSD-specific form factor was announced in March 2012. This device uses SLC flash, and provides a little less than the 2 GB/s maximum bandwidth supported in the PCIe 2.0 x4 configuration of this product.



Disk Drive Form Factor

The disk drive form factor is the traditional size and shape that have been used for hard disk drives for many years. SSDs have been built using these same sizes, and some have been built using some of the same width and depth but are thinner than traditional HDDs. Note that there can be more than one thickness, or height, for some of these drive form factors. Products from different vendors may have slight variations from the sizes specified below. The most common SSD drive type available today is the 2.5-inch form factor.

Dimensions

Table 3 – Drive Form Factor Dimensions

	mSATA	1.8-inch	2.5-inch	3.5-inch
Width (in/mm)	1.18/30.0	2.13/54.0	2.76/70.1	4.0/101.6
Depth (in/mm)	2/50.8	3.1/78.74	3.955/100.45	5.76/146.52
Height (in/mm)	0.191/4.85	0.315/8.0 0.197/5.0	0.591/15.0 0.374/9.5 0.276/7.0	1.0/25.4

Note that the traditional term for expressing the width of disk drives is not the actual measurement. For example, “3.5-inch” drives are actually 4 inches wide.

Hybrid Drive

Hybrid hard disk drives (H-HDD) are hard disk drives with a small NAND flash SSD imbedded in the drive package. This design combines the speed of SSDs with the capacity of HDDs, and these devices are generally targeted at the consumer market, including laptop computers, desktop computers and desktop gaming systems. These drives are available with a standard SATA interface.

These hybrid drives improve boot times by pinning some of the boot sectors into the SSD, allowing the operating system to start quickly and get the users to an operational state very quickly, regardless of the operating system.

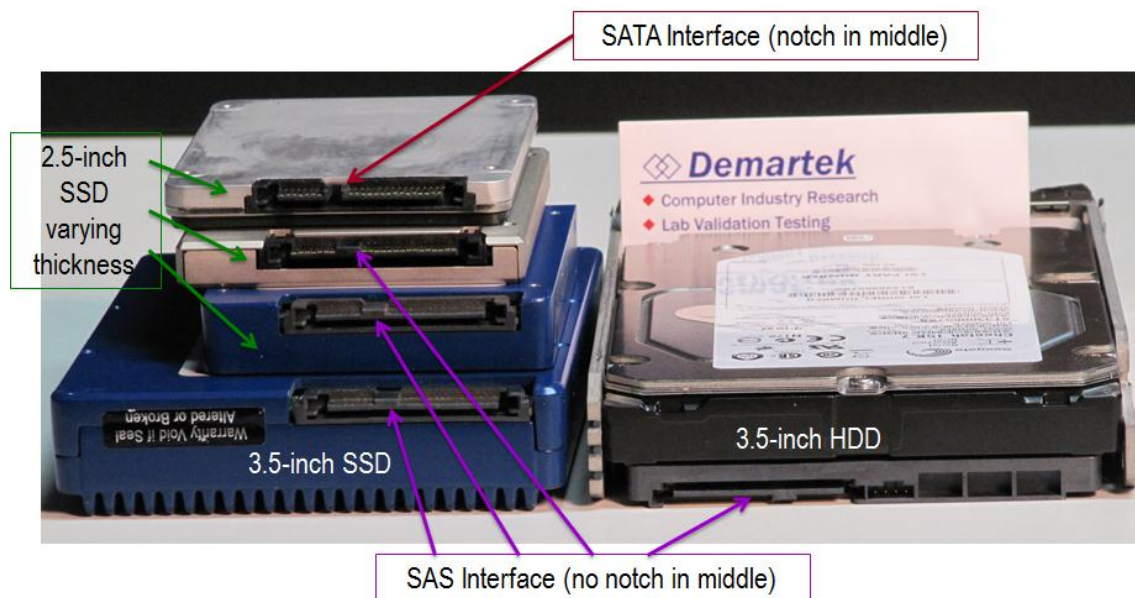
Several desktop computer operating systems can automatically detect an H-HDD. In these cases, the operating system will automatically cache frequently accessed data into the SSD, allowing the HDD to spin down and conserve power, which can extend the time that a system can run on a battery.

Interfaces

SSDs available today primarily have either a SAS or SATA interface. Older desktop class SSDs have IDE/PATA interfaces, and some older enterprise SSDs have a Fibre Channel (FC) interface. We expect enterprise SSDs in the future to either have SAS or possibly the 2.5-inch PCIe interface described previously.

The majority of the drive-form factor SSDs available today use on-board flash controllers that communicate to the outside world with a SATA interface chip. We expect that over time, newer SSD flash controllers, especially but not limited to those that manage enterprise SSDs, will embrace SAS and PCIe interface chips, and may eventually become more common than today's SATA interface flash controllers. We expect to see adoption of 12Gb/s SAS components in some SSDs by late 2012 or in 2013.

Several SSDs and one HDD are shown in the photo below for comparison of their interfaces.



PCIe Card Form Factor

PCI-Express (PCIe) stands for Peripheral Component Interconnect Express and is the computer industry standard for the I/O bus for computers introduced in the last few years. The first version of the PCIe specification, 1.0a, was introduced in 2003. Version 2.0 was introduced in 2007, and version 3.0 was introduced in 2010. These versions are often identified by their generation (“gen 1”, “gen 2”, etc.). It can take a year or two between the time the specification is introduced and the general availability of computer systems and devices using those specification versions. The PCIe specifications are developed and maintained by the PCI-SIG (PCI Special Interest Group).

PCIe I/O operations are measured in gigatransfers per second (GT/s), and the total bandwidth is determined by the number of lanes used, such as “x1”, “x4”, etc. These are generally spoken as “by 1”, “by 4”, etc. The following table provides bandwidth for various numbers of lanes for PCIe 1.0-3.0.

Table 4 – PCI-Express Maximum Transfer Rates

	GT/s	Encoding	x1	x2	x4	x8	x16
PCIe 1.x	2.5	8b/10b	250 MB/s	500 MB/s	1 GB/s	2 GB/s	4 GB/s
PCIe 2.x	5	8b/10b	500 MB/s	1 GB/s	2 GB/s	4 GB/s	8 GB/s
PCIe 3.x	8	128b/130b	1 GB/s	2 GB/s	4 GB/s	8 GB/s	16 GB/s

Although the GT/s rate of PCIe 3.0 is not double that of PCIe 2.0, the encoding schemes are different. PCIe 2.0 uses an 8b/10b encoding scheme, which has a 20% overhead on the data transfer rate. PCIe 3.0 uses a 128b/130b encoding scheme, which has a 1.5% overhead on the data transfer rate. By using an encoding scheme with a significantly lower overhead, the actual data transfer rate for PCIe 3.0 is doubled over PCIe 2.0. Additional information about encoding schemes is available at www.demartek.com/Demartek_Interface_Comparison.html on the Demartek website.

Many of the PCIe SSDs require either PCIe 1.0 x8 or PCIe 2.0 x4 slots. Some require PCIe 2.0 x8 slots. The larger capacity PCIe SSDs require a PCIe 2.0 x16 slot to achieve full bandwidth.

PCIe 2.0 servers typically support a total of 20 lanes of PCIe 2.0 for each processor socket.

On March 6, 2012, the major server vendors announced their PCIe 3.0 servers. These servers support 40 PCIe 3.0 lanes per processor socket (Intel® Xeon E5-2600), so a dual-processor server would have 80 lanes of PCIe 3.0 for I/O operations, doubling both the speed and total number of available lanes from the previous generation of servers. These servers have various combinations of PCIe 3.0 x4, x8 and x16 slots. We expect to see announcements of PCIe 3.0 SSDs beginning this calendar year.

Dimensions

The PCIe form factor is an excellent way to deploy NAND flash SSD technology inside a host server or storage system. Several companies are now producing PCIe SSDs. PCIe cards can be various physical sizes, including combinations of the following:

- ◆ Height
 - ◆ Full height (“standard height”): 4.20 inches (106.7mm)
 - ◆ Half height (“low profile”): 2.536 inches (64.4mm)
- ◆ Length
 - ◆ Full length: 12.283 inches (312mm)
 - ◆ Half length: 6.6 inches (167.65mm)

PCI-Express cards are also available in a mini PCIe form factor. This is a special form factor for PCIe that is approximately 30mm x 51mm or 30mm x 26.5mm, designed for laptop and notebook computers, and equivalent to a single-lane (x1) PCIe slot. A variety of devices including WiFi modules, WAN modules, video/audio decoders, SSDs, and other devices are available in this form factor. The mini PCIe form factor should not be confused with the mSATA interface, as these are not electrically compatible.

Capacities

Full-size PCIe SSDs are available with up to 5 TB per card, although cards of this size are quite expensive and cost far more than the price of large enterprise servers. It is not uncommon, however, to find capacities between 300 GB and 1.2 TB for PCIe SSDs for more reasonable prices.

Mini PCIe cards are available with up to 64 GB of capacity today.

Power

For PCIe 1.x and 2.0, PCIe cards used for I/O in servers can draw up to 25 watts of power from the PCIe bus. Some of the larger PCIe SSDs may require additional power to operate in their highest performance modes, and may need to obtain additional power from elsewhere in the system. Some rack servers only have power cables for existing components and do not have extra power cables available for some of these PCIe SSDs.

PCIe 3.0 may provide additional power draw for devices directly from the bus.

DIMM Form Factor

SSDs have been built onto 240-pin DIMMs that fit into DDR3 memory slots. These look very similar to regular memory DIMMs, but have a SATA data interface port mounted on the DIMM on or near the top. These SSDs get their power from the DIMM socket, but transfer their data through the SATA port. These use the JEDEC MO-269 form factor.

These SSDs are currently available in capacities up to 480 GB per DIMM socket.

This design allows SSDs to be installed into servers with large numbers of DIMM sockets that may not be populated with memory DIMMs, potentially providing large amounts of storage in existing servers. This also allows small storage arrays to be constructed in one rack unit (1RU) form factors.

Traditional Disk Arrays

Traditional disk arrays are available from most of the server and storage vendors with SSD technology as part of the storage system. Most of these can accept drive-form factor SSDs in place of some of the HDDs normally provided in the system. These SSDs can be used for direct storage and tiering, or can be used as an SSD cache. Some of these systems use PCIe SSD cards in their internal controllers as an SSD cache to improve performance. See the [data placement section](#) below for details on SSD caching and tiering.

By adding SSD technology to storage systems, these storage arrays can have multiple types of storage devices, each with different performance and capacity characteristics. Enterprise storage systems typically have:

- ◆ SSDs – Very high performance with low capacity per device.
- ◆ 15K and 10K RPM HDDs – High performance with moderate capacity per device.
- ◆ 7200 RPM HDDs – Moderate performance with large capacity per device.

Relative to the other types of storage described above, 15K RPM HDDs and 10K RPM HDDs have similar performance characteristics, and are sometimes grouped together.

When adding SSDs to enterprise storage arrays, it often requires only 3% - 10% of the total capacity to be comprised of SSDs in order to get significant performance improvements. We have seen the typical “sweet spot” for noticeable performance improvements with 3% - 5% of the total capacity in SSDs.

A few years ago when SSDs were first being introduced to enterprise storage systems, there were limits to the number of SSDs that could be supported by the storage array vendor. This was because, at that time, 15K RPM HDDs, introduced in 2001, were the highest performing devices available. The array controllers were designed to handle large numbers of 15K RPM HDDs with some additional headroom. When SSDs were introduced, which have significantly higher performance, the old designs were not sufficient for large quantities of SSDs, and the controllers became the bottleneck. During the last two or three years the storage array vendors have been re-architecting their controller designs to accommodate larger numbers of SSDs.

Prices of SSDs, while still higher per gigabyte than enterprise HDDs, have been dropping. Some of the storage array vendors are hinting that, over time, they expect to sell arrays with larger numbers of SSDs and fewer numbers of 15K RPM HDDs. These array vendors are currently proposing configurations of combinations of SSDs and 7200 RPM “nearline” large capacity drives that provide larger total capacity, higher overall performance, and lower cost than an equivalent array populated only with 15K RPM HDDs.

Demartek Comments on SSD Adoption

For the past three years, while speaking at conferences and other events, we have been predicting that by this calendar year (2012), a shift would take place. When enterprise users consider “tier-1” storage, they will first think of SSDs and not 15K RPM HDDs. This does not necessarily mean that 15K RPM HDDs will be eliminated this year, but simply that a major shift in end-user thinking is occurring.

All-flash Arrays

There are several all-flash arrays on the market today, one of which we tested for this guide. These arrays have only SSDs and no HDDs. They provide exceptional performance and some models have capacities up to 500 TB.

Most of these all-flash arrays are available today from younger or start-up vendors. The larger, more established server and storage vendors have noticed the traction that these all-flash arrays are gaining and it would not be unreasonable to expect announcements of all-flash arrays by the larger vendors, perhaps this calendar year.

Many of these all-flash arrays have some of the same advanced features that are available in the more traditional designs, such as thin-provisioning, compression, de-duplication, snapshots, and others.

Performance

Not surprisingly, all-flash arrays provide very high performance. Most of these products advertise 100K IOPS up to 1M IOPS or more for dozens of TB of capacity. They also advertise high bandwidth, such as 4 GB/s, 7 GB/s, or more, depending on the number of type of host interfaces available.

Also, because all-flash arrays do not have HDDs, their average latencies can be much lower than any storage system that depends heavily on HDDs. Some applications are more sensitive to latency than to overall performance, and for these, all-flash arrays may be a good choice.

Host Interfaces

Because the all-flash arrays are designed for high-performance and low latency, they often have high-speed host interfaces, such as 8 Gb/s Fibre Channel and 10 Gb/s Ethernet/iSCSI. Some even have 40 Gb/s (QDR) or 56 Gb/s (FDR) Infiniband host interfaces. It would be reasonable to expect storage systems to support faster host interfaces, such as 16 Gb/s Fibre Channel and 40 Gb/s, Ethernet in the not too distant future.

Power Consumption

Power consumption for all-flash arrays is considerably lower than for arrays full of HDDs. Some of these all-flash arrays consume only 500 watts for 10 TB of capacity. Many others consume well under 1000 watts for 10 TB – 20 TB of capacity.

Caching Appliances

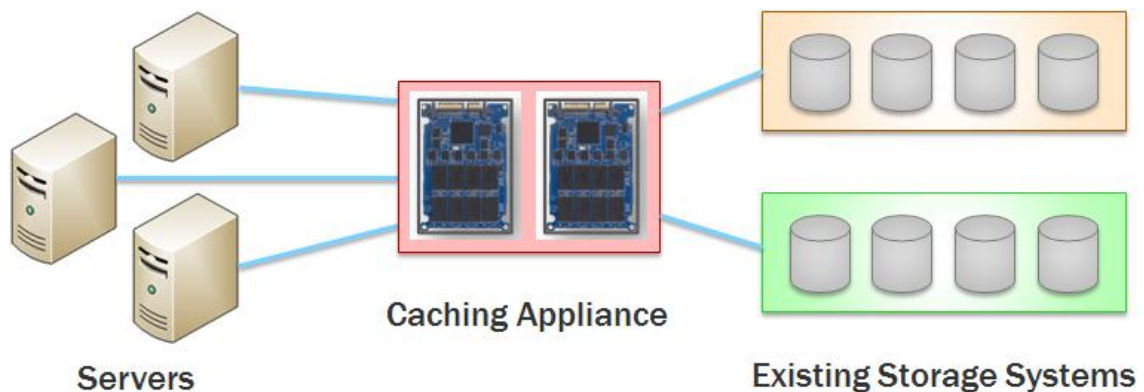
Caching appliances are another category of SSD storage. These external appliances may use NAND flash or DRAM plus NAND flash, and are designed to accelerate storage from one or more external storage systems and provide this acceleration for more than one host server. These appliances use their SSD storage as an external cache to provide the acceleration. In many cases, reads are accelerated and writes are passed through (“write-through”) to the back-end storage. In some cases, these accelerators may also accelerate writes (“write-back”).

These appliances sit in the network, either Fibre Channel or Ethernet, and accelerate storage performance for SAN (“block”) storage or NAS (“file”) storage. There are separate appliances for file and block storage. The block accelerator appliances typically work with Fibre Channel SAN storage or iSCSI SAN storage. The file accelerator appliances work with NFS and/or CIFS/SMB file traffic.

For the block accelerator appliances, specific LUNs from one or more back-end storage systems are assigned to be accelerated. For file accelerator appliances, specific “shares” or “mount points” are assigned to be accelerated. Some of the file accelerator appliances support the concept of global namespaces.

The diagram below shows the placement of the caching appliance in the network. In many cases, these appliances can be configured in pairs for redundancy so that there is no single point of failure.

In most cases, no changes are required to the back-end storage systems in order to take advantage of the caching appliances.



Data Placement

Once a decision has been made to incorporate SSD technology into an enterprise, the next major decision focuses on how to use this SSD technology, or more specifically, how to place data on these devices. There are two general ways to use SSDs: for primary storage or for caching. Both methods are effective at improving I/O performance. Some solutions allow an SSD to be used for both methods at the same time, allocating a portion of the available capacity to each method. Both of these two methods can work with any SSD form factor.

The overall goal of SSD solutions, both primary storage and caching, is to improve the performance of the storage system or systems, which are generally the slowest parts of a computing environment. If the SSD solution achieves this, it reduces the bottleneck in the storage system, which helps to expose the next bottleneck, which could be in the network, CPU or elsewhere.

SSDs as Primary Storage

When SSDs are used as primary storage, the following observations can be made:

- ◆ The administrator (or user) decides which data to place on the SSDs.
- ◆ The administrator decides when to place the data on the SSDs.
- ◆ The administrator is responsible for adjusting applications and backup configurations in order to find the data at its new location on the SSDs.
- ◆ Only the applications whose data resides on the SSDs gain a performance improvement.
- ◆ The performance gains are immediate.

In order for the user to determine which data should be placed on the SSDs, some sort of decision or ranking process must be performed. This can include ranking the most performance-sensitive applications, considering the amount of high-speed storage capacity needed, and determining the effects of the improved performance to the overall business or operation.

Once the data has been identified, the time to place it on the SSDs must be determined. Certainly the timing of the installation and configuration process must be planned. A more important consideration is the long-term set of workload I/O patterns. For example, it could be determined that the data for one application should be placed on the SSDs because it is the most performance-sensitive most of the time. However, perhaps a month-end process would benefit from having improved performance for a day or two, so for the period of that process, its data should be placed on the SSDs and the other application's data should be moved elsewhere.

If there were only two applications in the enterprise, swapping their usage of SSDs might not be too painful. However, most enterprises have many more than two applications. With a large number of applications, how can the ideal data placement be determined?

Automated Tiering Solutions

Automated tiering solutions observe the I/O patterns over a period of time, and can determine which data should be placed on SSDs and when to place that data on the SSDs in order to obtain the optimum performance benefit. These solutions are available from several sources, some operating as software in the host servers, while others operate within a storage system. The minimum sampling time for these solutions typically ranges from ten minutes to one hour, and for some solutions, the sampling time can be customized. It is generally recommended to run the

initial sampling for at least one day and usually several days in order to obtain a reasonably good sampling of the I/O activity before enabling any of the data movement operations.

These data movement operations are generally policy-based, with the policies set by the administrator. Some of these solutions provide default or “starter” policies that can be used by those new to this concept. Most of the solutions allow the policies to be highly customized.

It helps if the administrator has some idea of the I/O patterns for each of the application workloads. Knowledge of the I/O patterns can help in determining how to construct the storage tiers and the amounts of capacity to place in these tiers of storage. The data sampling features of these solutions will help identify the I/O patterns, and these results can be compared to what you may already know about your workloads. For example, one application might perform 90% of its I/O on 5% of the data, and another application might perform 100% of its I/O evenly across 100% of the data. Different tiering policies may be appropriate for different I/O activity patterns.

Policies can be set to observe I/O patterns during certain times, and ignore I/O patterns during other times. Policies can be set to lock certain volumes in certain tiers for specific time periods. Policies can be organized by application type, RAID group type, or nearly any other criteria that the administrator chooses. Policies can also be set to exclude certain data types or volumes from the SSD tier.

Some of these tiering solutions provide exactly two tiers: the SSD tier and the non-SSD tier. Others provide three or more tiers so that SSDs, 15K/10K RPM HDDs, and 7200 RPM HDDs can each have their own tier. As we mentioned in the [traditional disk array section of this report](#), the relative difference in performance between 15K RPM HDDs and 10K RPM HDDs is small compared to the difference between them and SSDs or 7200 RPM HDDs.

The other consideration for automated tiering solutions is to determine the optimal time to perform data movement. This might be during the night, on weekends, etc. This data movement will consume some I/O resources within the storage system as it moves data from one storage tier to another in the background.

Many of the automated tiering solution providers have tools that can help in planning by predicting the effectiveness of their tiering solution. Also, many of these vendors offer professional services to help plan and deploy their solutions.

Chunk Size

Automated tiering solutions perform their analysis of I/O patterns and data movement in chunk sizes. Chunk sizes are the amount of data moved at one time. There is a difference of opinion as to the optimum chunk size, and this varies by tiering solution. We have seen chunk sizes of 512 KB, 2 MB, 4 MB, 8 MB, 42 MB, 1 GB, etc. Some solutions use a variable chunk size and others allow the administrator to specify a chunk size within a given range. As of this writing, we have not compared the efficiency of different chunk sizes in our lab, so we can't speculate on which size is optimal.

SSD Caching

When SSDs are used as a cache, the following observations can be made:

- ◆ The caching solution places a *copy* of “hot” data into the cache.
- ◆ The caching solution decides when to place the copy of the hot data into the cache.
- ◆ Multiple applications can gain a performance benefit.
- ◆ The aggregate performance gains occur over time as the cache “warms-up.”
- ◆ Applications do not need to be modified to take advantage of the SSD cache.
- ◆ Some caching solutions only cache reads, others cache both reads and writes.
- ◆ Management of a caching solution is relatively simple.

Caching solutions are available as host-based software, packaged with RAID controllers, exist as external caching appliances, or are found inside of large storage systems. The host-based caching solutions can generally take advantage of any SSD form factor available to that host server, while caching solutions supplied with hardware tend to use only the SSD form factors supported by that hardware.

Because caching solutions move a copy of hot blocks into their cache, no modification or re-configuration of applications is required. If the caching solution is a host-based solution, then generally no changes are required to any back-end SAN or NAS storage. If the caching solution is found inside the storage system, then except for the installation of the caching solution, no other changes to the storage system are generally required.

Different caching solutions employ different caching algorithms in order to determine which data is “hot,” and when to place a copy of that hot data into the cache. Some caching solutions do not cache I/Os for very large blocks on the assumption that these are backup jobs, video streaming applications, or other applications where the likelihood of a cache “hit” is low. Similarly, some of these caching solutions do not cache multiple sequential I/O patterns for the same reason. Some caching solutions may pre-fetch data near “hot” blocks in the anticipation that these pre-fetched blocks will become hot.

Because caching solutions observe all I/O traffic within their assigned purview, they are not dependent on knowledge of specific applications – they place copies of any data that becomes hot into their cache, can react to repeated accesses to the same blocks in real-time, and accelerate those accesses immediately. These solutions can also flush items out of cache in real-time as other blocks become “hot.”

When caching solutions are deployed and as the cache warms, the remaining I/O activity on the back-end HDDs is reduced, allowing the HDDs to perform better.

It’s important to understand that the objective of SSD caching is to enable a balanced server-storage system. Caching can eliminate the I/O bottleneck that slows overall system performance below the capabilities of CPU, memory and storage infrastructure. In testing a caching solution, when the performance gains stop increasing due to CPU saturation (100% utilization), the goal of removing the I/O bottleneck has been achieved. When the CPU has reached 100% utilization, providing a cache on a larger or faster SSD will not increase system performance, as faster I/O won’t improve the performance of a fully saturated CPU.

SSD Caching Workloads

SSD caching is most effective with “cache-friendly” workloads. Workloads are said to be “cache friendly” if they have “hot spots,” where there are repeated accesses to many of the same data blocks. Some workloads are more cache friendly than others. For example, many OLTP database workloads are cache friendly because many of the same records are being accessed more than once in a relatively short time period. Also, the indexes to databases are accessed frequently and are good candidates for SSD caching. Read-intensive web server applications also are good candidates for SSD caching, as they often have many “hits” on the same pages as topics become popular. File server workloads can be cache friendly because, at a minimum, the table of contents is frequently accessed, and some data files are accessed more frequently than others. For these workloads, only the “hot” data areas need to fit into the cache, not the entire set of data. If an entire set of data can fit into the cache, then, of course it will show very strong performance acceleration.

On the other hand, some workloads are not very cache friendly and do not show significant performance gains with SSD caching. Any workload that accesses a data set that is larger than the SSD cache, and accesses this data approximately evenly is not a good candidate for SSD caching. We have found that some workload simulation tools, such as Microsoft Exchange Jetstress, tend to be one of these cache unfriendly workloads, as one of its design goals was to make the I/O patterns friendly to relatively slow hard disks, by spreading out the I/O access patterns and not having many hot spots.

For those interested in using Microsoft Exchange Jetstress, we have compiled a comparison of the I/O profiles of Exchange Server 2003 vs. 2007 vs. 2010. This comparison is available at www.demartek.com/Demartek_Exchange_2003_2007_2010_I-O_Comparison_Summary.html on the Demartek website.

Write-through Cache vs. Write-back Cache

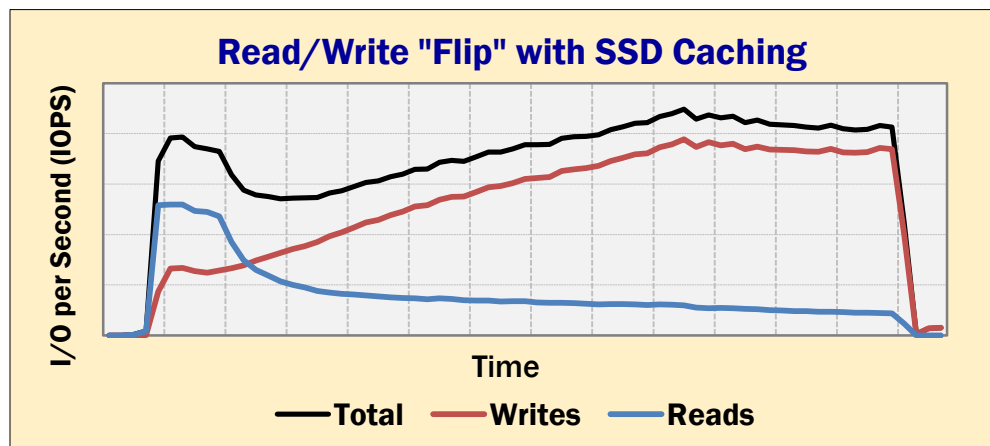
For those SSD caching solutions that cache writes, there are generally two ways these can operate. A write-through cache simply passes writes directly to the back-end storage without attempting to accelerate the write operation. From a performance perspective, writes going through these solutions generally behave as though there were no cache at all, and are subject to the performance characteristics of the back-end storage. These solutions also allow the back-end storage to guarantee the integrity of the data written, because a write-through SSD cache basically does not touch this write data. Write-through cache is common with SSD caching solutions today. Some write-through SSD caching solutions place a copy of the write data into their cache, in the anticipation that a read operation for this same data may occur in the near future. If such a read operation does occur, then the read is accelerated.

Some SSD caching solutions support write-back, which means that the SSD cache will accelerate the write operation by “completing” the write to its own cache. Shortly after it has signaled a completed write, it passes the write to the back-end storage. When an SSD caching solution caches the writes, it must be able to guarantee the integrity of the data that was written, and insure that it eventually gets written on the back-end storage device or system. These solutions often have to take extra steps to insure this data integrity (and protection). For example, a write-back SSD caching solution must be able to properly handle writes that it reported as completed when the back-end storage system fails or is temporarily offline.

Read/Write “Flip”

When testing SSD caching solutions that cache only the reads, we have repeatedly observed an interesting, though expected, phenomenon when testing mixed read/write workloads. Similar results have been observed with various SSD caching solutions, such as server-side, storage-side, etc.

As the SSD warms up and fills with hot “read” data, the back-end storage system is affected positively. For example, when testing read/write workload mix of 70% read and 30% write, before the cache warms up, this 70%/30% read/write mix is seen by the HDDs in the back-end storage system. However, as the SSD cache fills with reads, the read/write mix as seen by the back-end storage system changes over time to become much more write-heavy, because the reads are satisfied by the SSD cache. After the cache is completely full, we have observed a higher total I/O rate on the back-end storage system than without the SSD cache. The chart below shows the read/write “flip” and was taken from one of our public evaluations of a server-side SSD caching solution.



Demartek SSD public evaluation results are available in our [SSD Zone](#) on our website.

Multiple SSD Caching Solutions

There is generally nothing preventing multiple caching solutions from being used together. For example, a host-based or “server-side” SSD caching solution can be combined with a storage system SSD cache. In our lab tests, we have seen higher overall performance using this example than when using only one SSD caching solution.

SSD Caching vs. Tiering – Demartek Opinion

There is debate over which method, caching or tiering, is the most effective. If one were to compare a fully warmed caching solution to a fully 100% automated tiering solution, both would perform well at accelerating the performance-sensitive data. The differences would be:

- ◆ Caching solutions are very simple to manage but tiering solutions introduce management complexity.
- ◆ Caching solutions react immediately in real-time to accelerate performance but tiering solutions have to move data before performance gains are realized.
- ◆ Tiering solutions consume extra back-end storage IOPS but caching solutions do not.

In our opinion, SSD caching is the preferred solution for most situations.

Operating System Behavior with NAND Flash SSDs

When a file is deleted in most operating systems, the table of contents entry for that file is removed from the volume table of contents, but the data blocks for that file are normally untouched. When a new file is saved at a later time, a new entry is created in the volume table of contents, and the file data is stored at any available block address, which may overwrite old data blocks that were previously deleted.

With HDDs this type of file system process poses no special challenges, as the number of steps for reading and writing with HDDs are nearly the same. With NAND flash SSDs, however, this normal file system process can create some interesting challenges. Because writes with NAND flash media can only occur on blank pages, overwriting data blocks requires the completion of a “program-erase” cycle, which requires considerably more steps to complete than the equivalent process on an HDD.

This situation would be improved if the operating system or application could initiate program-erase cycles on NAND flash media when files were deleted, so that when a new file was ready to be written sometime later it would find plenty of erased blocks that were ready to accept writes.

Trim

For SATA-interface NAND flash SSDs, there is a command known as Trim that initiates the program-erase cycle on blocks that are no longer needed by the operating system. This process is sometimes known as “garbage collection,” and runs in the background when the Trim command is issued.

Microsoft Windows 7 and Windows Server 2008 R2 automatically detect SATA-interface NAND flash SSDs and enable the Trim functions for those devices. These operating systems also turn off the automatic background defragmenting for SATA-interface SSDs. Defragmenting SSDs unnecessarily consumes program-erase cycles without providing the same level of performance improvement that defragmenting does for HDDs.

Other operating systems support Trim in more limited fashions. Some Linux operating systems support Trim, but only for specific filesystem types such as EXT4. In these cases, Trim is available but not enabled by default; it must be enabled by the administrator. Apple OS X Lion 10.7 supports Trim but only for Apple-branded SSDs.

For those operating systems that do not support Trim, the SSD vendors provide separate utilities that can initiate the garbage collection as a separate step that can be run in the background.

UNMAP

Trim works well for SATA-interface SSDs that the operating system can detect as SATA-interface devices, or for those non-SATA devices that implement Trim in their driver. However, there is no Trim command in the SAS/SCSI protocol. For SAS/SCSI interface devices, there is an equivalent command to Trim known as UNMAP. UNMAP is currently supported by the newer SAS SSDs, but UNMAP support has not yet made its way higher into the storage stack of most operating systems, RAID controllers, etc.

Garbage Collection in External Storage Systems

External storage systems do not generally present individual physical devices to the host operating systems, so the operating systems have no guaranteed way of knowing that any given device is a NAND flash SSD. These external storage systems manage garbage collection directly by monitoring activity on the SSDs and determining the appropriate time to initiate garbage collection.

Test Environment for this Guide

To run the tests required to produce this guide, we used server, networking and storage infrastructure in the Demartek lab in Arvada, Colorado, USA. This infrastructure includes servers, 6Gb SAS RAID controllers, 10Gb Ethernet storage and 8Gb Fibre Channel SAN storage. In addition, the vendors supplied their solutions to Demartek to be tested in the Demartek lab.

Workload Tests

We used a combination of synthetic and real-world workloads to test the SSD technologies that we configured for this Deployment Guide. The purpose of all of these workloads was to generate I/O activity on the various SSD solutions and observe the performance.

In some cases, we compared the SSD performance to one or more HDD configurations. In other cases, we provided multiple workload results for the same SSD configuration. In one of the SSD caching tests, we altered the amount of SSD cache available to the workload to show the effects on performance of different amounts of SSD cache.

Synthetic Workloads

Synthetic workload generators allow specific I/O patterns to be generated. Parameters for these tools include read/write mix, random or sequential, block size and queue depth (number of simultaneous I/O requests issued). These tools provide performance results for very specific I/O profiles.

- ◆ IOmeter – An open-source I/O load generator
- ◆ VDBench – An open source I/O load generator
- ◆

Real-World Workloads

Real-world workload tools show performance of real applications. This is accomplished by running the same I/O engine as the actual application (Exchange Jetstress) or a tool that simulates the real applications.

- ◆ IOZone – An open source filesystem benchmark tool
- ◆ Microsoft Exchange Jetstress – Exchange Server disk I/O simulation tool
- ◆ Neoload – Webserver stress test tool
- ◆ ROBOCOPY – Windows file copy tool
- ◆ TPC-C like workload – Database workload simulation tool

Vendor Products Tested for this Guide

Specific SSD-related products were tested to produce this deployment guide. A vendor-specific section in this document provides additional information for these products. These were:

- ◆ FlashSoft SE (acquired by SanDisk) SSD Host Caching Software
- ◆ Nimbus Data S-Class all-flash storage array
- ◆ OCZ Enterprise PCIe and disk form factor SSDs
- ◆ SMART Storage Systems Enterprise disk form factor SSDs

Server Specifications

The Demartek servers used to produce this SSD Deployment Guide are listed below.

Table 5 – Demartek Server Specifications

Server	Processor	PCIe	Clock Speed (GHz)	Total Cores	Total Threads	Memory (GB)	Boot Drives
C	Intel Xeon E5345	1.0	2.33	8	8	48	Internal SAS array 4x 15K SAS drives
D	Intel Xeon E5345	1.0	2.33	8	8	48	Internal SAS array 4x 15K SAS drives
J	Intel Xeon X5680	2.0	3.33	12	24	144	Internal SATA SSD
K	Intel Xeon E3-1280	2.0	3.5	4	8	32	Internal SATA SSD

Network Infrastructure

Two switches in the Demartek lab were used for the SSD testing. Various Cat5e and Cat6 cables were used for the 1Gb connections. The 10Gb cabling consisted of copper SFP+ cables and fiber-optic OM3 cables.

Switches

- Dell PowerConnect 2748 – 48x 1Gb ports
- Cisco Nexus 5020 – 40x 10Gb ports

FlashSoft SE SSD Caching Software

[FlashSoft](#) SE is a host-based SSD caching software solution that can use any standard SATA, SAS, or PCIe SSD to accelerate the performance and scalability of the host server on which it is installed. Currently there are versions for Linux and Windows Server 2008 R2.

Versions for VMware and Hyper-V have been publicly discussed.

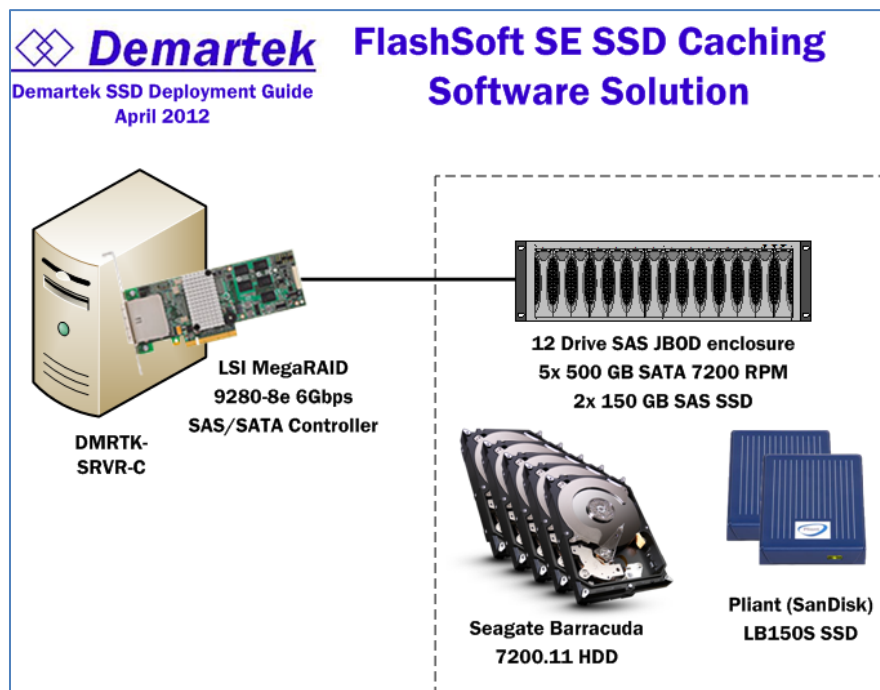


FlashSoft Corporation was acquired by SanDisk on February 15, 2012, and joins the Enterprise Storage Solutions business of SanDisk. According to the [public announcement](#), “SanDisk intends to sell FlashSoft’s products as standalone software, as well as offer these software products in combination with SanDisk’s growing portfolio of SAS, PCIe and SATA enterprise solutions.”

The Flashsoft SE installation process is easy and intuitive for Windows systems, and follows fairly standard practices for Linux systems.

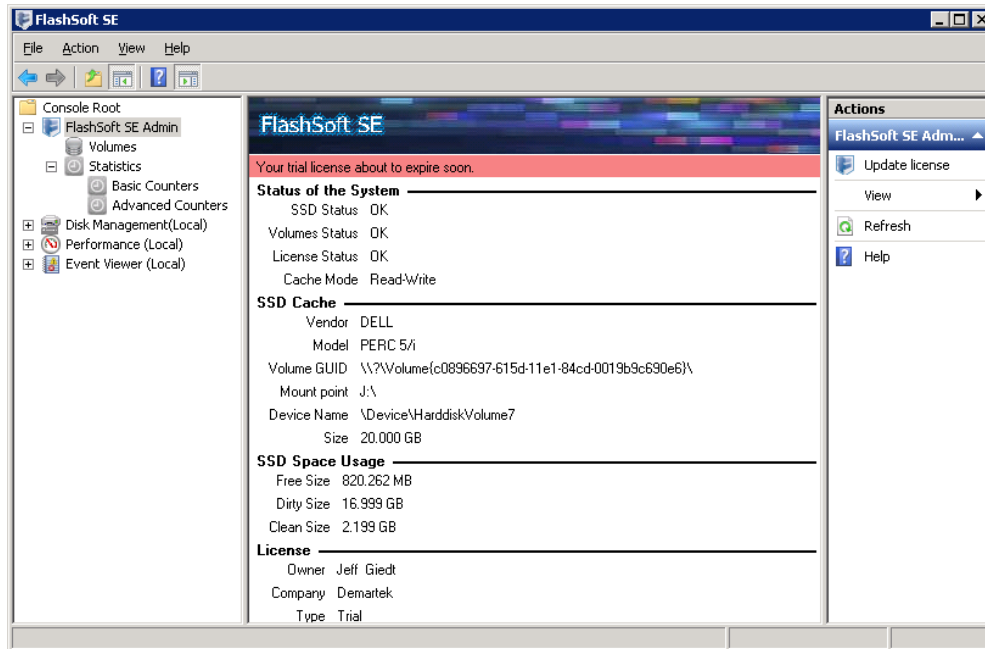
For our configuration, we used an LSI MegaRAID 9280 controller connected to a JBOD enclosure with two Pliant (SanDisk) SAS SSDs (model LB-150S), each having a capacity of 150 GB. We configured the two SSDs as one RAID0 volume group using the MegaRAID management software, so that one device was presented to the operating system. We assigned that LUN to the FlashSoft software as its cache device.

For our tests, we wanted to see the effect of different cache sizes on some of the workloads, so we allocated different volume sizes in the MegaRAID controller software that we presented to the operating system and the FlashSoft software.

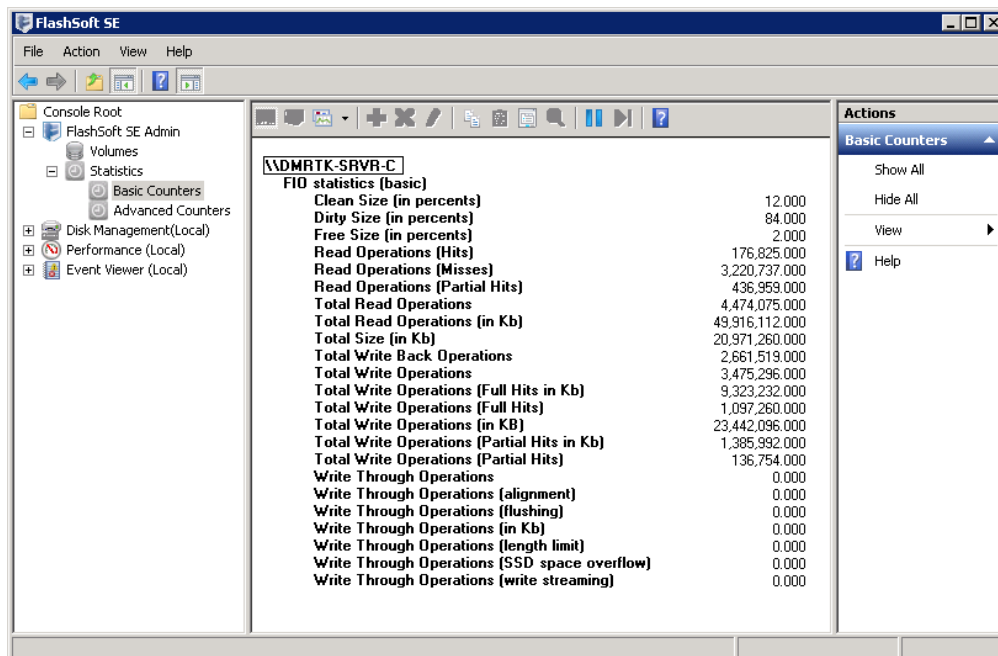


Managing SSD Cache with FlashSoft SE

Managing the SSD Cache with FlashSoft SE is fairly straightforward. The management interface is intuitive and the process to assign the LUN to be accelerated is fairly simple. As expected for an SSD caching solution, very little effort is required to manage this type of solution.



FlashSoft SE provides a good number of cache statistics to provide a complete picture of the caching activity.



Tests run with FlashSoft

We ran two sets of tests using FlashSoft SSD caching:

- ◆ IOmeter on Windows Server 2008 R2
- ◆ TPC-C like workload on Windows Server 2008 R2 with different cache sizes

FlashSoft software is available for Windows and Linux environments. We tested the Windows version for this report. The caching algorithms used are the same for both the Windows and Linux versions of the product, so we would expect similar results for Linux in similar configurations.

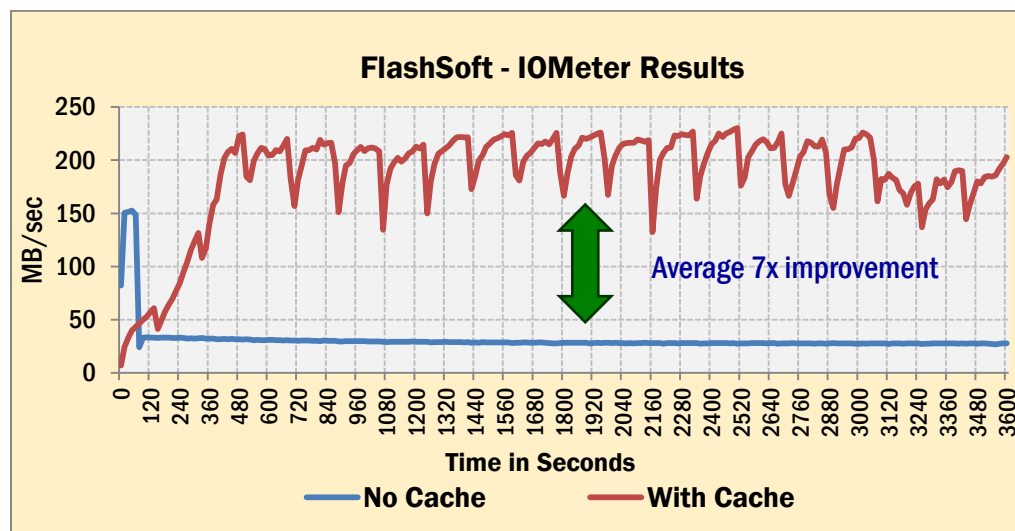
As we described previously in the [SSD caching section](#) above, the performance of SSD caching solutions are highly dependent on the workload. Some workloads have uniformly distributed accesses across the entire data set, while others have hot spots with repeated accesses in the same area.

IOmeter on Windows

IOmeter, when running random workloads, accesses its entire file randomly and approximately evenly, resulting in very few “hot spots” and making it cache “unfriendly.” For the IOmeter tests, we made sure that the cache volume was larger than the IOmeter data file, so that over time, the entire file would be cached. The goal was to see how quickly the cache warmed and the size of the performance improvement with the entire contents of the file potentially residing in the cache.

For this particular test, we used the following IOmeter parameters:

- ◆ Read/write mix: 75% read, 25% write
- ◆ 100% Sequential I/O
- ◆ Block size: 8KB
- ◆ Queue depth: 32



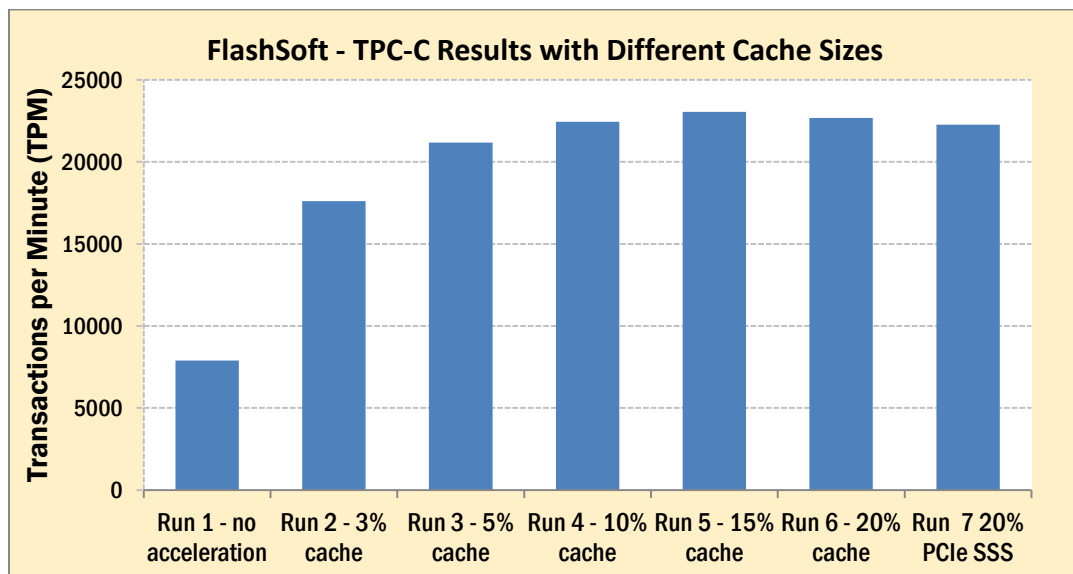
The results follow the classic SSD caching warm-up pattern. For this test, the cache became fully warm at approximately 480 seconds.

TPC-C like Workload and Different Cache Sizes

For the TPC-C tests, we wanted to take this real-world workload and note the effects on performance that different sizes of SSD cache had, relative to the size of the database. For these tests, we configured the TPC-C workload with 400 users and 4500 warehouses, using Microsoft SQL Server 2008 R2. The database was approximately 400 GB.

Each test in this group was run for 7200 seconds, or 2 hours. The first run had no caching enabled, and served as the baseline. For each of the subsequent runs, the amount of SSD available to the cache was increased, the cache was cleared, and then the test repeated with the same parameters. As the amount of SSD cache beyond 5% of the database size increased, the percentage of additional performance gain decreased.

When the cache reached 15% of the database size, or 60 GB, the CPU utilization on the server under test reached 100%, and additional increases in cache size made no further performance improvements. This means that the SSD cache successfully moved the bottleneck from the storage to a different resource, the CPU in this case. Different server and storage configurations may yield different results.



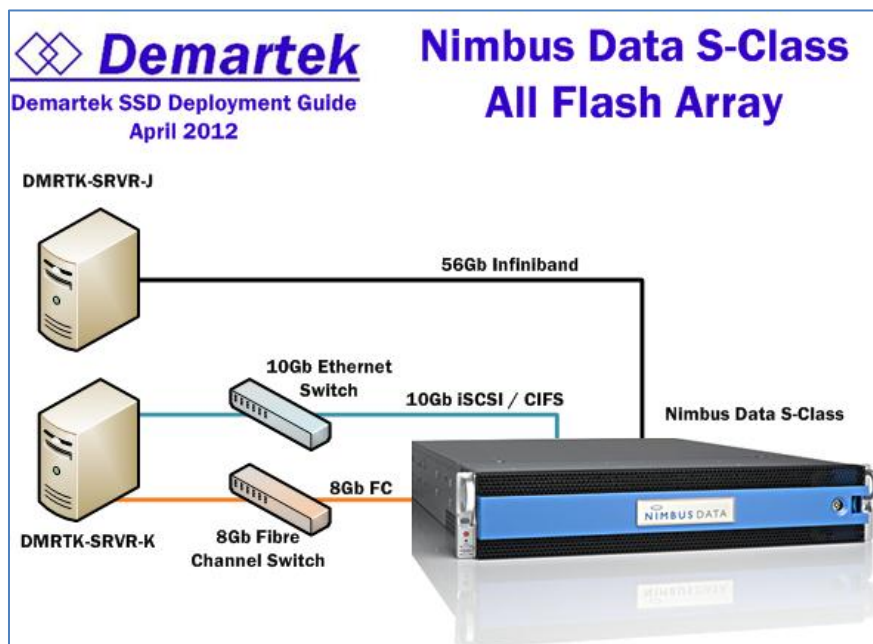
Nimbus Data S-Class All Flash Array

[Nimbus Data](#) introduced its S-Class all flash array in 2010. The 2 rack unit (2U) system we tested includes one shelf of 2.5-inch flash drives. The Nimbus S-Class supports block and file protocols simultaneously, including iSCSI, NFS, and SMB over Gigabit and 10 Gigabit Ethernet, FCP over 4/8Gb Fibre Channel, and native SRP over QDR 40Gb or FDR 56Gb Infiniband. The S-Class can be expanded up to 100 TB of all flash capacity using SAS-connected expansion shelves.



Demartek ran several synthetic and real-world performance tests and some tests individually, while others were combined tests over multiple protocols at the same time. In all cases, the Nimbus Data S-Class provided outstanding performance.

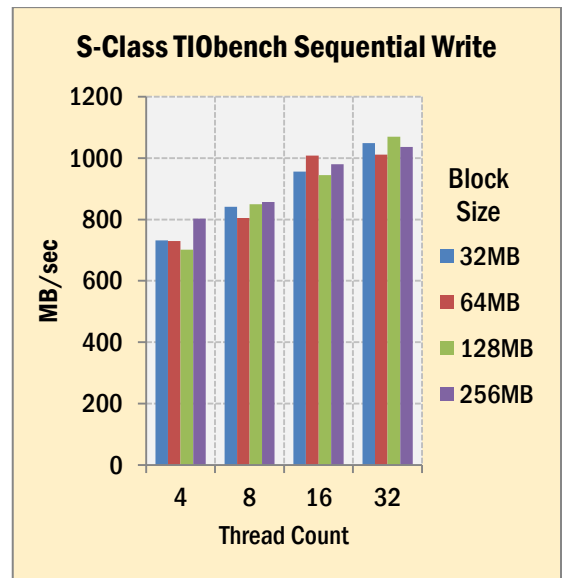
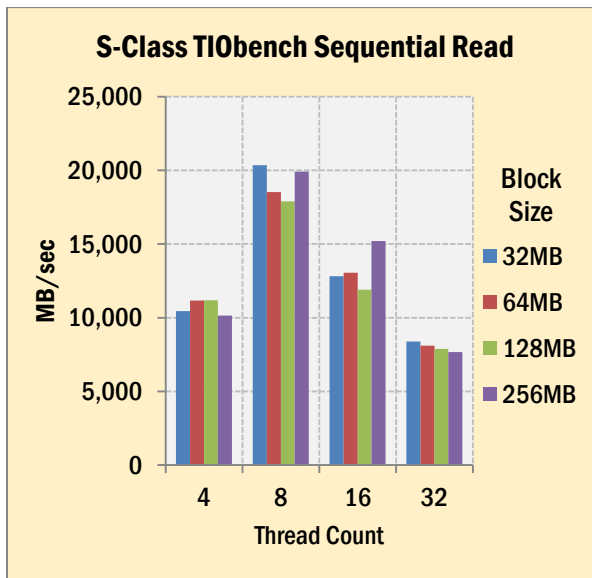
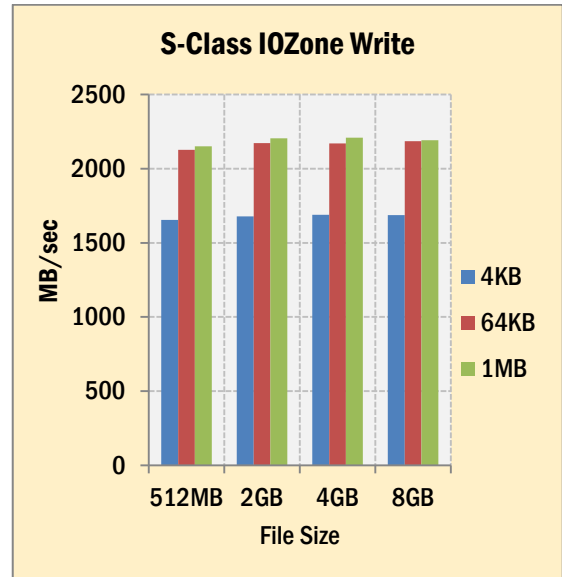
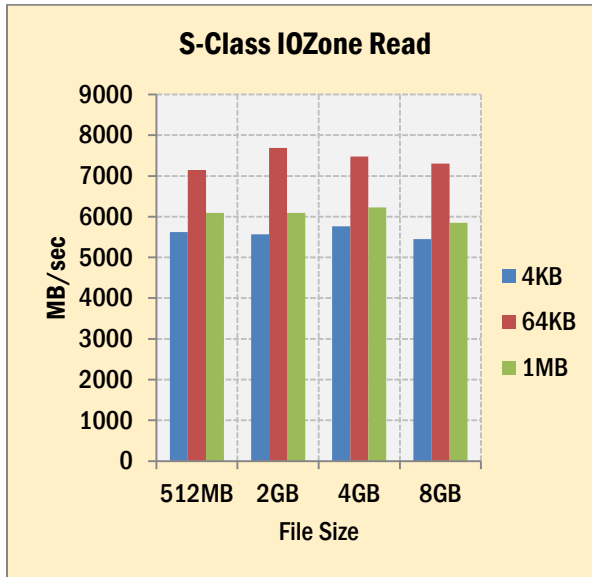
In addition to supporting Windows and Linux operating systems, the Nimbus Data S-Class supports VMware VAAI 5.0, including automatic thin provisioning and erase block reclamation, as well as VMware's automatic SSD features. All Nimbus Data systems are also certified for Citrix XenServer 5.6 and 6.0 and XenDesktop.



The fastest line rate interface for the Nimbus S-Class storage system is the 56 Gb Infiniband interface, so we ran tests with our strongest PCI-Express gen 2.0 server, DMRTK-SRVR-J. The specifications for this server are listed in Table 5 above. We configured this server as a physical server running CentOS 6.1 and let the operating system and applications use all the cores and memory available for maximum performance. For these tests, we installed a Mellanox ConnectX-3 FDR Infiniband host channel adapter (HCA) and connected one port on the host server to one FDR Infiniband interface port on the S-Class system.

Tests run with Nimbus Data

Two tests in particular showed outstanding performance in this Infiniband configuration: IOzone and TIObench.

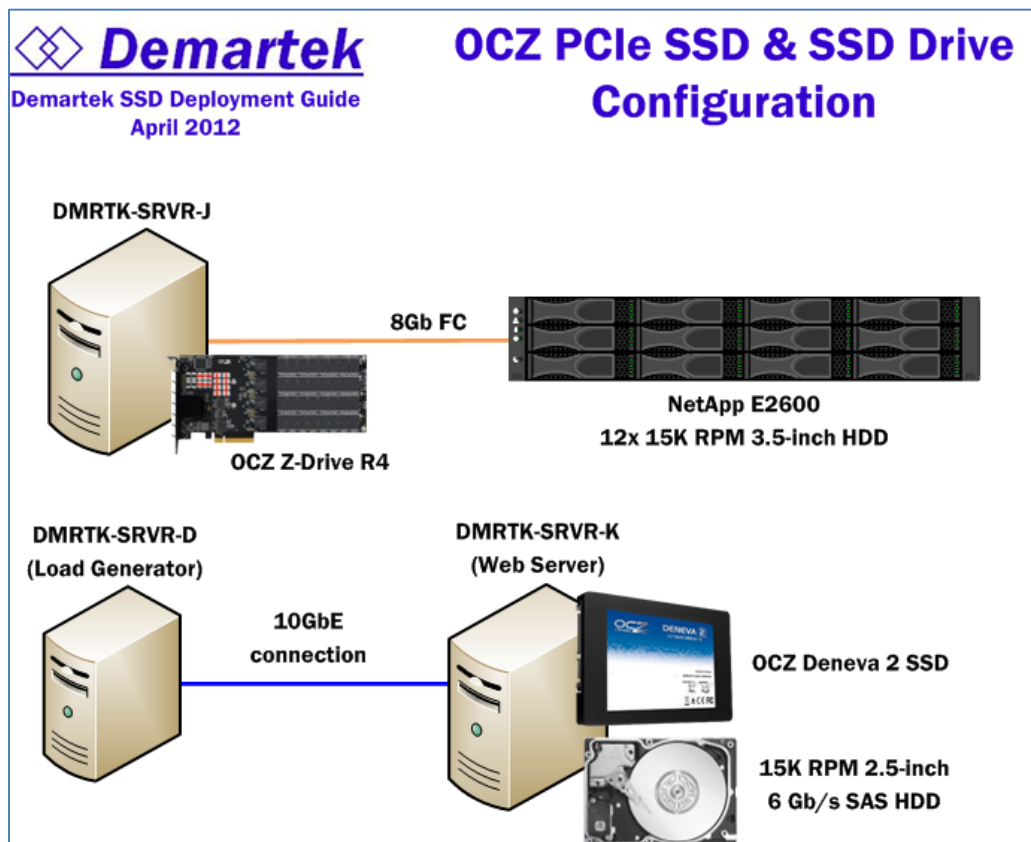


OCZ PCIe and SATA SSDs

OCZ Enterprise manufactures enterprise-class PCIe, SAS, and SATA SSDs. The PCIe form-factor SSDs include the Z-Drive R4 product family. The disk-drive form factor SSDs include the Deneva and Talos product families.



We compared the performance of the OCZ SSDs to HDD configurations in two different sets of configurations. The first configuration compared the OCZ Z-Drive R4 SSD to an enterprise 8Gb Fibre Channel storage system with qty. 12 15K RPM 3.5-inch HDDs. The second configuration compared a single 480 GB, OCZ Deneva 2 SSD to a single 73 GB, 15K RPM 2.5-inch HDD as the storage for a webservice application.

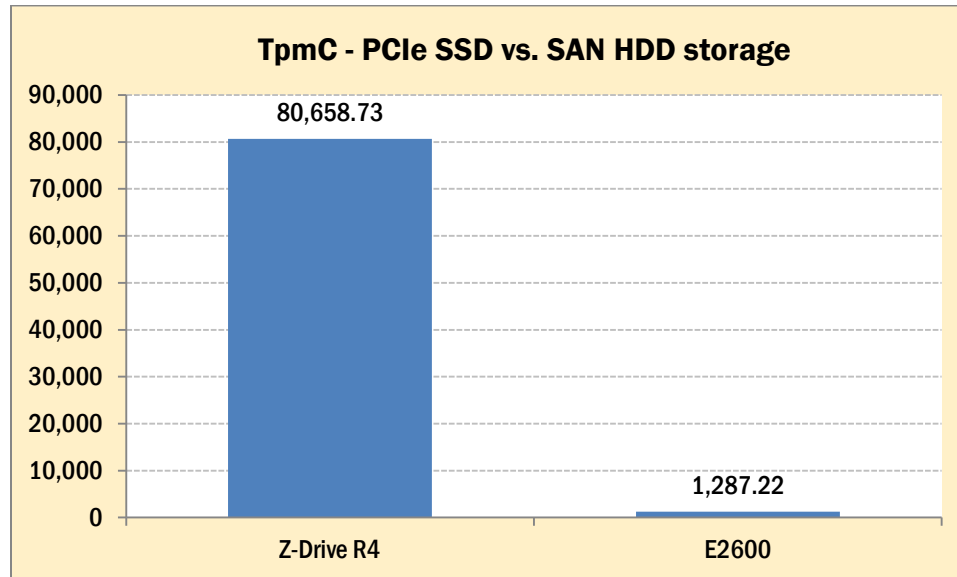


Tests run with OCZ SSDs

Configuration 1 – PCIe SSD vs. 8Gb Fibre Channel Storage

This TPC-C like workload simulates an online transaction processing (OLTP) system. This TPC-C like workload had 100 users and 4500 warehouses. The test was run on the 8 Gb Fibre Channel storage, then repeated on the OCZ Z-drive R4.

This comparison is a single 1.6 TB PCIe SSD vs. 3.6 TB, Qty. 12, 15K RPM HDD storage system.



The Z-Drive R4 outperformed the 12 15K RPM HDDs by a factor of more than 60.

Configuration 2 – Drive Form Factor

For web server applications, key metrics are hits per second, throughput and response time. We ran a read-intensive web server application with the webserver application data stored on a single 15K RPM HDD and then repeated the same test with the webserver application data stored on a single SSD.

The webserver data was randomly and approximately evenly accessed for the duration of the test. Each HTML text page included three graphic images on average, so each web page request resulted in four hits, on average. Total webserver data was approximately 40 GB and included 1.48 million files of web content data.

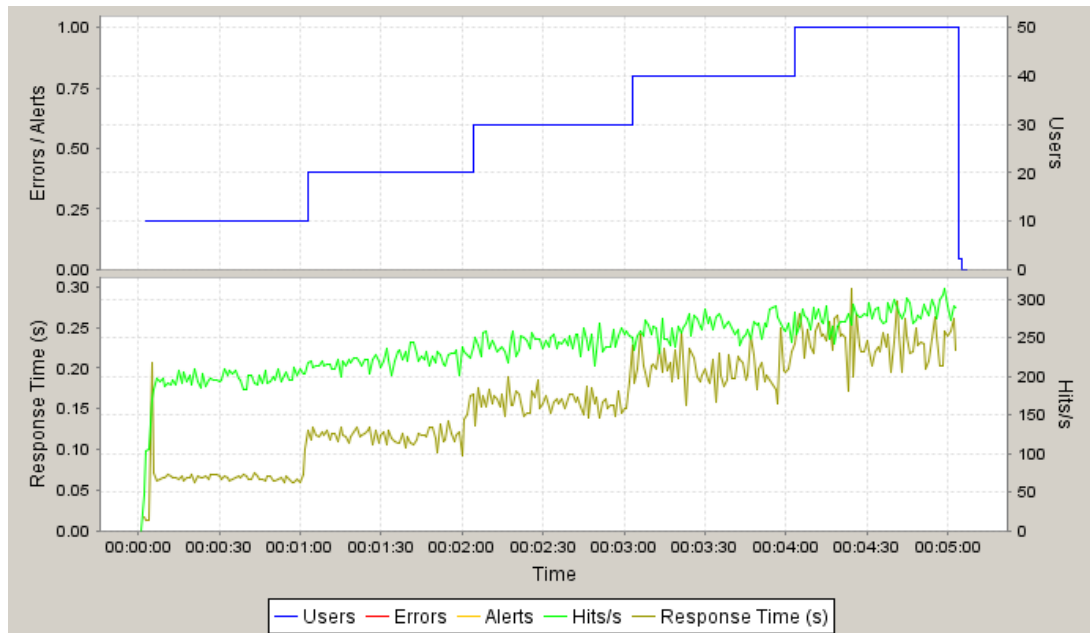
- ◆ 80,000 HTML text pages
- ◆ 1.4 million graphic images (JPEG and PNG)

Table 6 - Webserver Performance HDD vs. SSD

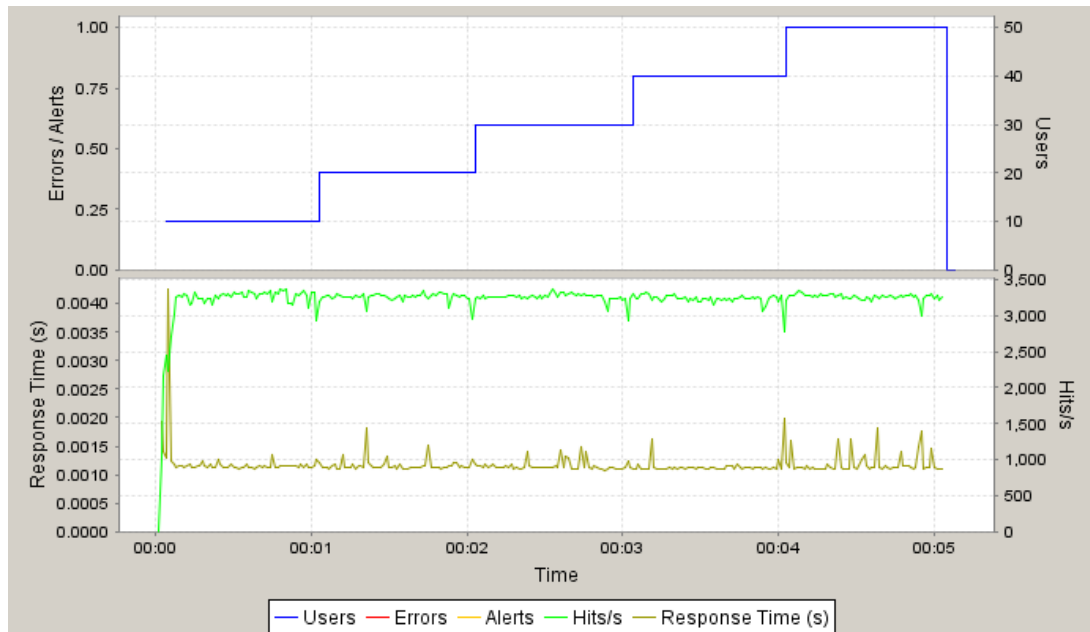
Statistics	HDD	SSD	Difference
Average pages/s	60.1	811.5	+1250%
Average hits/s	241.0	3246.1	+1247%
Total pages	18097	244253	+1250%
Total hits	72537	977062	+1247%
Average Request response time	0.162 s	0.001 s	-99.4%
Average Page response time	0.499 s	0.004 s	-99.2%
Total throughput	2535.19 MB	34096.47 MB	+1245%
Average throughput	67.38 Mb/s	906.22 Mb/s	+1245%
Device capacity	73 GB	480 GB	+557%

The average performance difference was 12x higher performance for the SSD and the response time (latency) was more than 99% lower (better) for the SSD.

Single 15K RPM HDD Results



Single OCZ SSD Results

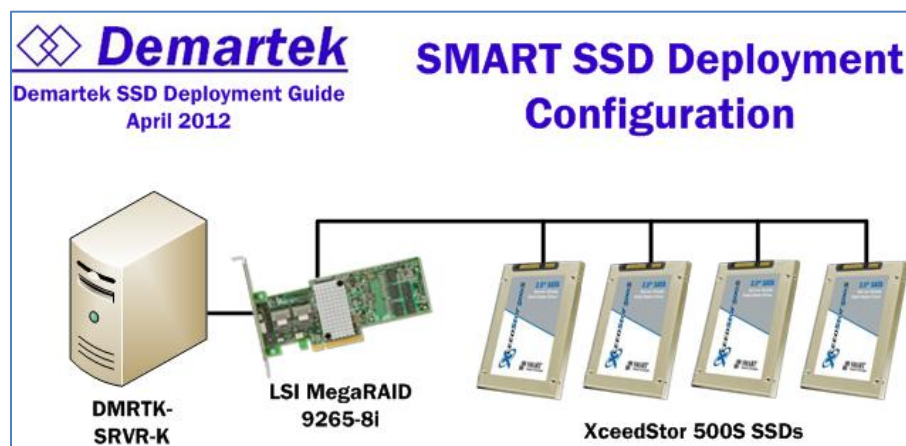


SMART Storage Systems SSDs

[SMART Storage Systems](#) manufactures enterprise-class SAS and SATA SSDs. These SSDs include the Optimus, XceedIOPS2 and XceedStor product families.

The Optimus SAS and XceedIOPS2 SATA SSDs are targeted at applications that use a mixed workload and require an endurance of 7 to 10 random drive writes/day. The XceedStor SSDs are targeted at read intensive applications that require an endurance of less than 1 random drive write/day.

We decided to place heavy enterprise workloads on a group of four SMART Storage System XceedStor 500S SSDs. We ran a TPC-C like workload with different RAID configurations. During one of these TPC-C like workload tests, we also ran Exchange Jetstress on the same drive group at the same time to see if both enterprise workloads could be handled simultaneously.



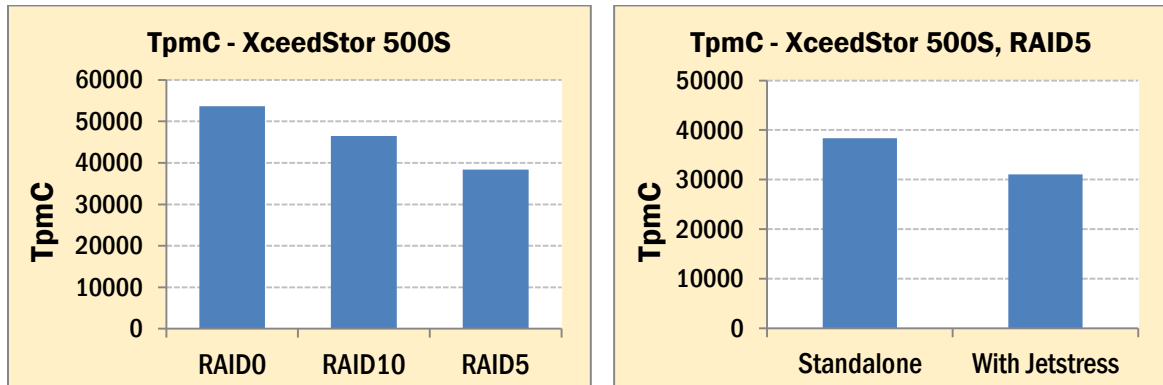
The test results show that these SSDs can handle multiple enterprise workloads at the same time. Traditional best practices often dictate that separate enterprise applications keep their data separate, but with the substantial performance capabilities of these SSDs, perhaps it may be time to re-think storage application workloads.

Tests run with SMART Storage Systems XceedStor 500S SSDs

Each XceedStor 500S SSD has a raw capacity of 240 GB. Three different RAID configurations were used: RAID0 (4 drives), RAID10 (2+2 drives), and RAID5 (3+1).

The TPC-C parameters were 100 connections and 4500 warehouses.

TPC-C like Workload Results



Exchange Jetstress Results

Exchange Jetstress was run simultaneously with the TPC-C like database workload on the same drive group in the RAID5 configuration. The Exchange Jetstress parameters were 200 mailboxes, size=1000 MB, IOPS=0.3, threads=8 and storage groups=2. The Exchange logs were written to separate SSD storage devices.

The Jetstress results show that even while running an OLTP workload on the same drive group, Jetstress received a "PASS." It far exceeded the target IOPS and the latencies were well below the recommended thresholds of 20ms for average database read latency.

Achieved Transactional I/O per Second	8020.012
Target Transactional I/O per Second	60
Initial Database Size (bytes)	229832261632
Final Database Size (bytes)	241928634368
Database Files (Count)	2

MSExchange Database ==> Instances	I/O Database Reads Average Latency (msec)	I/O Database Writes Average Latency (msec)	I/O Database Reads/sec	I/O Database Writes/sec	I/O Database Reads Average Bytes	I/O Database Writes Average Bytes	I/O Log Reads Average Latency (msec)	I/O Log Writes Average Latency (msec)	I/O Log Reads/sec	I/O Log Writes/sec	I/O Log Reads Average Bytes	I/O Log Writes Average Bytes
Instance2376.1	5.049	13.060	1982.395	2041.449	32833.939	34129.318	0.000	0.620	0.000	601.053	0.000	7337.407
Instance2376.2	5.017	13.122	1967.375	2028.793	32833.836	34148.300	0.000	0.620	0.000	600.647	0.000	7353.404

Legal and Trademarks

The most current version of this SSD Deployment Guide is available at:
http://www.demartek.com/Demartek_SSD_Deployment_Guide.html.

Intel and Ultrabook are trademarks of Intel Corporation in the U.S. and/or other countries.

Microsoft, Windows, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

PCI Express and PCIe are registered trademarks of the PCI-SIG.

The stylized “SMART Storage Systems” as well as “SMART Storage Systems”, “Optimus” and “XceedIOPS” are trademarks of SMART Storage Systems.

VMware is a registered trademark and vSphere is a trademark of VMware, Inc.

Demartek is a registered trademark of Demartek, LLC.

All other marks and names mentioned herein may be trademarks of their respective companies.