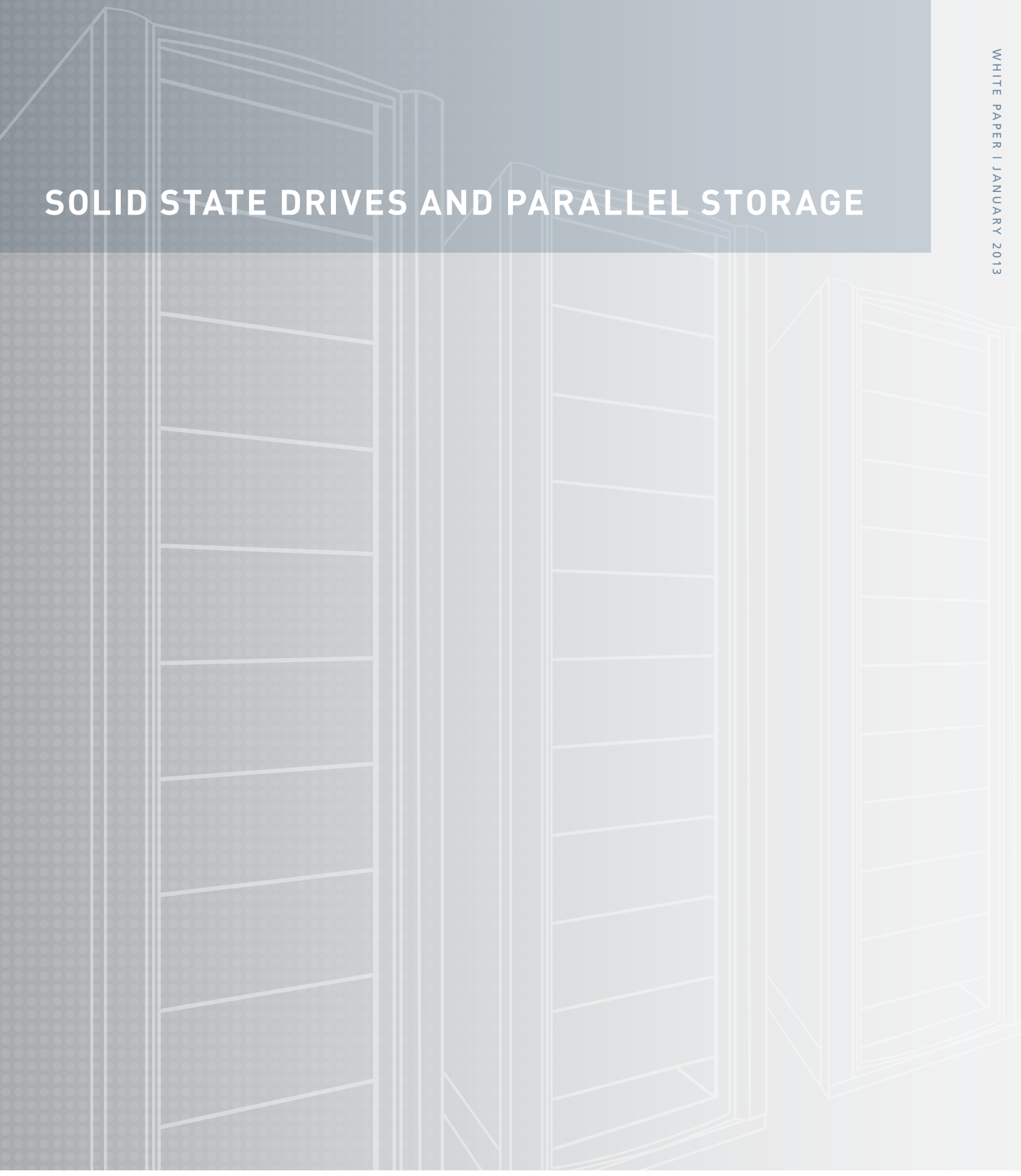
SOLID STATE DRIVES AND PARALLEL STORAGE

WHITE PAPER | JANUARY 2013



panasas®

1.888.PANASAS | www.panasas.com

OVERVIEW

Solid State Drives (SSDs) have been touted for some time as a disruptive technology in the storage industry. In laptop and desktop computers, SSDs mean longer battery life, shortened boot time and increased application performance. For servers running database applications, SSDs mean higher I/O performance, reduced I/O response time, and reduced data center power consumption.

While SSDs still represent a luxury upgrade for laptops, the value of SSDs in servers as part of an overall storage infrastructure, in light of the orders of magnitude increase in SSD-driven small file, transactional performance, is looked at from the perspective of both $/GB (a traditional way of looking at HDDs), and $/IOPS.

The mixed workloads associated with technical computing applications in High Performance Computing (HPC) environments require a new way of thinking about storage to fully realize the benefits of SSDs for high performance storage infrastructures. This white paper examines the underlying SSD technology and how Panasas ActiveStor® 14 fully leverages that technology without introducing excessive cost or otherwise compromising the needs of HPC customers.

What is an SSD?

SSD is an umbrella term for a device that behaves like a traditional Hard Disk Drive (HDD), but uses memory technology instead of a magnetic medium as its method of recording data. It is interesting to note that this is actually not a new concept, as SSD-based products have been around since the 1970s.

For the most part, SSDs are offered in the same form factors as traditional HDDs. This allows for easy, drop-in replacements into existing storage infrastructure. SSDs are available with either DRAM or NAND flash used as the underlying media. Given the higher density of NAND flash relative to DRAM, flash-based SSDs are now more prevalent, despite the fact that DRAM-based SSDs offer more performance. For the purpose of this white paper, we will focus on NAND flash-based SSDs.

WHAT'S ALL THE HYPE ABOUT?

The game-changing event that made SSDs such a hot topic in recent years has simply been the reduction of cost over time. With the rise in the use of flash in USB memory sticks, smartphones and tablets, the price of NAND flash memory has come down significantly in the last 10 years. SSDs were extremely expensive in the 1970s through the early 2000s in terms of $/GB, and mainly seen in government and military applications. Now, a consumer grade SSD can be easily obtained in retail channels at less than a dollar per GB. This, of course, gives storage vendors expanded component options in order to create faster solutions at costs suitable for almost all pricing tiers.

WHY ARE SSDs FASTER THAN HDDs?

Many people commonly believe that SSDs are always faster than HDDs. This is not always true. In reality, it depends on the workload.

Data on an HDD is stored in concentric tracks on platters (the recording media). An actuator arm with a read/write head moves on top of the platter to perform the actual read or write operation, moving from track to track, much like a DVD/Blu-ray drive.

The analogy to a DVD/Blu-ray drive does not stop there. A movie on a DVD or Blu-ray is really just a large file on the disc. Playing a movie from the disc or an HDD is a large block, sequential read operation. There is not a lot of movement on the part of the actuator arm as the motor spins the platter as the movie plays. The mechanical design lends itself to this kind of streaming, sequential file access, making the HDD quite capable of delivering good sequential read or write performance.

Now, imagine if there are multiple small files being accessed simultaneously. In order to keep up with all of the read/write requests in this case, the actuator arm will have to move away from the current file, across the platter to the next correct track, land on the correct block of the next file, read/write that data, and then move on to the next file and repeat the process all over again.

HDD technology was simply not designed to accommodate this type of small block, random workload. HDD manufacturers, over the years, have done as much as possible to increase the performance of HDDs for this type of workload, mainly by increasing the spin rate of the motor and reducing the size of HDD platters. Enter SSDs.

SSDs are not limited by the mechanical movements of an actuator arm. Instead, there are multiple channels inside the SSD; each channel operates like an HDD's actuator. Thus, the multiple channels operate independently, allowing multiple files to be accessed at the same time. An SSD is better suited for small block, random workloads, where there are lots of concurrent requests for data. An HDD cannot physically do this.

© Intel

Why are SSDs Faster than HDDs?

Does this mean SSDs are going to replace HDDs? In a word: No.

HDD manufacturers have been quite good at optimizing their products for $/GB. Compared to an SSD, an HDD simply provides more capacity at a lower price. SSD manufacturers, on the other hand, have been focused on optimizing for dollar per random I/Os per second (IOPS).

The chart below clearly shows that HDDs and SSDs have distinct uses in the big picture of any storage environment: HDDs for applications where sequential performance is important and the capacity requirement is large; and SSDs for applications where random performance is important and capacity requirements are relatively small.

CHARACTERISTIC	HARD DISC DRIVE	SOLID STATE DRIVE
OPTIMAL TRANSFER SIZE	Large Block (16KB+)	Small Block (4-8KB)
OPTIMAL WORKLOAD	Sequential Read OR Write	Random Read AND Write
PERFORMANCE (IOPS)	100's	10,000's+
AVAILABLE CAPACITY*	4TB (3.5")	1.6TB (2.5")
$/GB (ENTERPRISE GRADE)*	~ $0.10/GB	~ $1.60/GB

HOW SSDs WORK

Inside all SSDs are two major components: multiple NAND flash memory chips and a controller. The number of memory chips determines the capacity of the drive. The controller, being the "brain" of the SSD, has the responsibility of making the collection of NAND flash chips look like a fast HDD to the host system.

This is not an easy task. In order accomplish its job; the SSD controller must perform the following tasks:

1. Host interface protocol management
As a very fast analog to an HDD, SSDs must communicate to the host via a storage protocol such as SATA, SAS, or Fibre Channel. There are PCI Express-based SSDs in the market today. However, PCI Express is not a storage interface/protocol yet.

2. Bad block mapping
Just as with magnetic media, NAND flash blocks go bad from time to time. When a bad block is detected during a write operation, SSDs mark the block "bad" and then remap the block from a pool of spare blocks, before retrying the original write operation. When this occurs during a read operation, the SSD controller attempts to recover the data, if possible, before remapping the block.

3. Caching and power-fail protection
It is a common practice for SSDs to use a small amount of DRAM to speed up reads and writes. However, as DRAM is volatile, data meant to be written to the NAND flash can be lost during an unexpected power outage or drive removal. To protect against this, SSDs have a secondary power circuit with either batteries or capacitors to allow for time to flush the cached data. Thus, in addition to running the caching policies, the SSD controller must also monitor the health of the secondary power circuit to ensure its ability to protect data in the cache.

4. Data compression
Some SSD controllers implement data compression. The principal advantage of this is possible endurance improvements for the SSD (more on this later), if the data is compressible. For the SSD controller, implementing compression means managing the statistics and tables for the compression engine.

5. Data encryption
Similar to data compression, some SSD controllers implement data encryption. Unlike data compression, however, encryption has become more necessary as a method to prevent data theft. To the SSD controller, this means the need to manage a crypto engine for all legitimate data traffic in and out of the SSD.

6. Wear leveling
NAND flash can wear out over time. The mechanism for write operations in NAND flash is different than magnetic media. For magnetic media, a write operation can occur over an area that has been previously

*As of December 2012

written to, by simply writing over it. A previously written-to area of NAND flash has to first be erased, before it can store new data; this is referred to as a Program/Erase (PE) cycle. Each block in the NAND flash has a finite amount of P/E cycles before wear-out. Thus, to prevent any single block from wearing out earlier than the rest, the SSD controller must maintain a history of how many times each block has been erased/programmed and spread the writes evenly across all available blocks.

7. Garbage collection
Given that previously written-to blocks must be erased before they are able to receive data again, the SSD controller must, for performance, actively pre-erase blocks so new write commands can always get an empty block. With operating systems support for TRIM (SATA) and UNMAP (SAS) commands, the SSD controller needs to proactively erase the blocks that the operating system deemed to have no valid data.

8. Media scrubbing and error correction
One interesting quirk about NAND flash is the concept of Read Disturb and Write Disturb. As a read or write operation progresses, it is possible for blocks adjacent to the one being accessed to be disturbed, causing an undesired bit flip. This is basically a form of silent data corruption. SSD controllers must proactively look for these bit flips and correct them.

Finally, the controller must aggregate the performance of the flash chips in the SSD to achieve the desired performance for the end user, with no performance degradation stemming from any of the aforementioned tasks.

PANASAS ACTIVESTOR 14

Now that we have covered SSD technology and what it is good for (and not good for), it is worth exploring how and why SSD technology is used in the Panasas flagship storage solution, ActiveStor 14.

The heart of any scale-out storage system is ultimately the parallel file system that runs as part of its storage operating system. For Panasas ActiveStor, that operating system is called PanFS. Unlike most other storage systems, PanFS is an object storage system.

Objects can be best thought of as being at a level of abstraction half way between block storage and file storage. By using objects, PanFS can be extremely smart about how to store file data. Among other things, PanFS can detect small file reads and writes and differentiate them from large file streaming throughput. It can also choose where to store the file system namespace and file attributes (metadata).

Panasas has had great success leveraging its blade architecture, object storage, and integrated parallel file system for high performance computing applications. ActiveStor systems deliver the industry's highest single file system throughput per terabyte of enterprise SATA storage. The design goal for ActiveStor 14 was to maintain ActiveStor's leadership in large file throughput and increase IOPS and small file performance to comparable levels.

As you already know, SSD technology vastly outperforms traditional spinning hard drives when it comes to small block, random I/O performance (especially reads). So as we started designing ActiveStor 14 the question became what is the best approach to leverage this capability to transform the product's performance while keeping the system cost effective. The PanFS allows ActiveStor 14 to intelligently leverage both SSD and Enterprise SATA HDD technology in compelling ways—accelerating access of the file system namespace (metadata) and access of small files via SSDs and using HDDs for large file performance.

Now, a consumer grade SSD can be easily obtained in retail channels at less than a dollar per GB. This, of course, gives storage vendors expanded component options in order to create faster solutions at costs suitable for almost all pricing tiers.

File I/O Optimization

An important piece of research that Panasas undertook was to determine how much SSD capacity customers would need and whether it would make a big enough performance difference to be worth the incremental cost of including SSD storage in the system. To do this, we extracted key data from a number of production file systems in the field to learn more about file size distributions and overall number of files stored.

Perhaps surprisingly, data set size for even predominantly large-file throughput-oriented workloads involve a very large number of small files under 64KB in size. The chart above illustrates that most of these HPC file systems are comprised of approximately 70 percent of small files by count. Without any additional data points, you might think that an all-SSD storage solution might be the right approach even though it would be extremely costly on a per-TB basis.

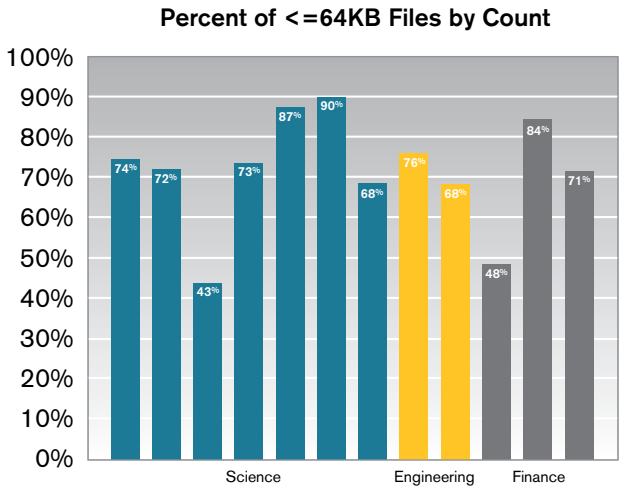
Fortunately, the opposite is true. Even though most files are small files by count, large files dominate by capacity and these are the files mostly accessed in streaming workloads where SATA HDDs excel. In the graph above you can see that small files less than 64KB typically consume well less than 1 percent of file system capacity (not including file system RAID overhead or metadata consumption).

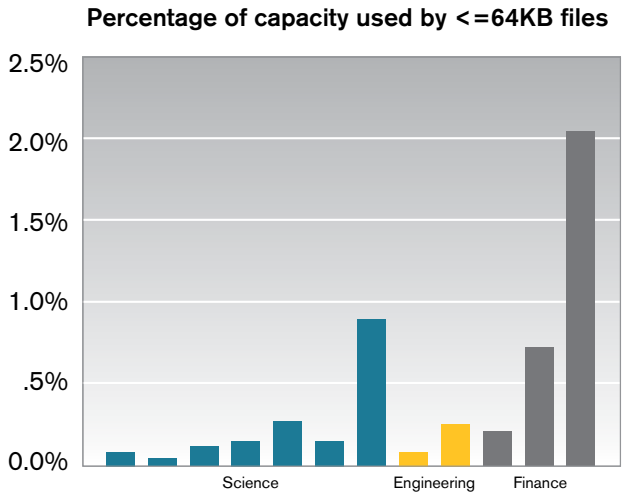
ActiveStor 14 is fundamentally designed for both large file throughput and small file IOPS workloads alike. It is the first truly general purpose parallel file system in the HPC/big data space capable of handling mixed workloads with high performance.

SSD TIER CAPACITY OPTIMIZATION

Next, we looked at how much SSD capacity would actually be needed to store small files and all file system metadata for these workloads, assuming a Panasas storage blade with 8TB of enterprise SATA disk.

The result was three ActiveStor 14 models, each with a larger amount of SSD capacity to address a wide variety of big data workloads. The 81TB configuration uses ten storage blades, each with two 4TB SATA HDDs and one 120GB 1.8" SSD. Even with SSD representing only 1.5 percent of storage capacity, this model meets the needs of most workloads and maximizes $/TB of the overall system. The 83TB configuration steps up to a 300GB SSD per storage blade for workloads that are more metadata and/or small file oriented. Finally, the ActiveStor 14T 45TB model uses storage blades with two 2TB SATA HDDs and a 480GB SSD—a full 10.7 percent of total capacity on the SSD





for workloads in Finance and other markets that are particularly skewed towards small file performance.

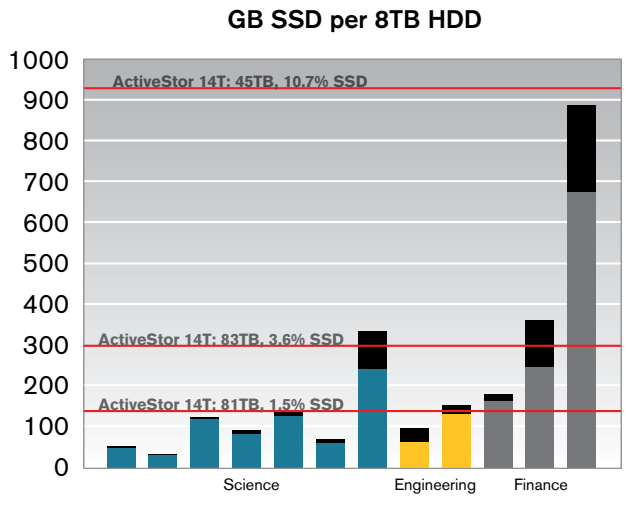
ACTIVESTOR 14: FILE SYSTEM RESPONSIVENESS DELIVERED

ActiveStor 14 is fundamentally designed for both large file throughput and small file IOPS workloads alike. It is the first truly general purpose parallel file system in the HPC/big data space capable of handling mixed workloads with high performance. Unlike caching or tiering approaches where having the namespace and/or small files that you need on the SSD tier is not guaranteed, the Panasas approach makes accessing and managing your data fast and easy—a single, unified SATA HDD/SSD global namespace based on a unified, high performance SATA HDD/SSD tier.

When it comes to performance, the results are clear. Panasas has measured random small file reads at 11x faster than ActiveStor 12, its previous fastest product and multiple-times faster performance over a wide range of other small file and namespace access metrics—especially in directory listing speed, file deletes, NFS v3 performance, and most other namespace-focused benchmarks. The end result is higher file system responsiveness across the board with the HDD and SSD each doing what they do best, all at a price point that represents solid value.

LOOKING FORWARD

Panasas expects many exciting developments from the SSD industry that companies like Panasas will be able to take advantage of in future storage solutions. For example, the emergence of standards-based PCIe SSDs with Non-Volatile Memory Express (NVMe) and SCSI Express specifications promise to bring more performance and better management and storage system integration. And as flash memory pricing continues to drop, the use of SSD technology will become more and more prevalent for HPC applications, even in capacity-focused environments.



04152013 1102