Solid state storage technology for ProLiant servers

2nd edition

Technology brief

Introduction

HP is now delivering solid state storage devices based on flash memory in addition to traditional disk drives based on spinning magnetic media. Solid state drives (SSDs) are the most familiar of these new devices, since they are plug-compatible with disk drives and used with traditional SATA/SAS disk controllers. But we are also introducing new types of flash-based storage devices for ProLiant servers that will significantly improve performance for certain types of applications. This technology brief provides an overview of solid state storage and the new high performance products that we at HP are developing using this new technology.

Flash memory technology

Most solid state drives use flash memory technology, a non-volatile computer memory that can be electrically erased and reprogrammed. There are two configurations, NOR flash and NAND flash. NOR and NAND flash both store information in arrays of floating-gate transistors called "cells." But they differ in how the cell arrays are organized and accessed. NOR flash memory cells connect in parallel to the bit lines, letting you read and program the cells individually. NAND flash memory cells connect in a series, and you can only read or program the cells as a group.

NAND's architecture allows you to create memory arrays with almost twice the density of comparable NOR memory and at a lower cost. As a result, most devices use NAND flash memory.

Single-level and multi-level cell NAND flash

There are two primary types of NAND flash technology:

- Single-level cell (SLC) technology works by storing a single level of charge in each cell, representing a single bit of information.
- Multi-level cell (MLC) technology stores one of four different charge states in a cell. This allows each cell to represent 2 bits of information, effectively doubling storage density.

NAND flash memory using multi-level cell technology has quickly become the primary flash technology in consumer products. Compared to SLC, MLC technology has several characteristics that make it less desirable for creating the type of higher performance, high reliability devices required for server storage (Table 1), including the following:

- Higher internal error rates caused by the smaller margins separating the cell states, necessitating larger ECC memories to correct them
- Shorter lifespan in terms of maximum number of program/erase cycles
- Slower read performance and significantly slower write (program) performance

Table 1. Primary characteristics of SLC and MLC flash

	SLC flash	MLC flash
Random access	25 microseconds	60 microseconds
Serial access	50 nanoseconds	30 nanoseconds
Page program (write)	200 microseconds	800 microseconds
Maximum program/erase cycles	100,000 @ 1 bit ECC	5000 – 10,000 @ 4 bit ECC

As Table 1 shows, MLC NAND flash has comparatively poor read and write performance. More important, SLC flash has a program/erase lifecycle—often referred to as endurance—that is 10 to 20

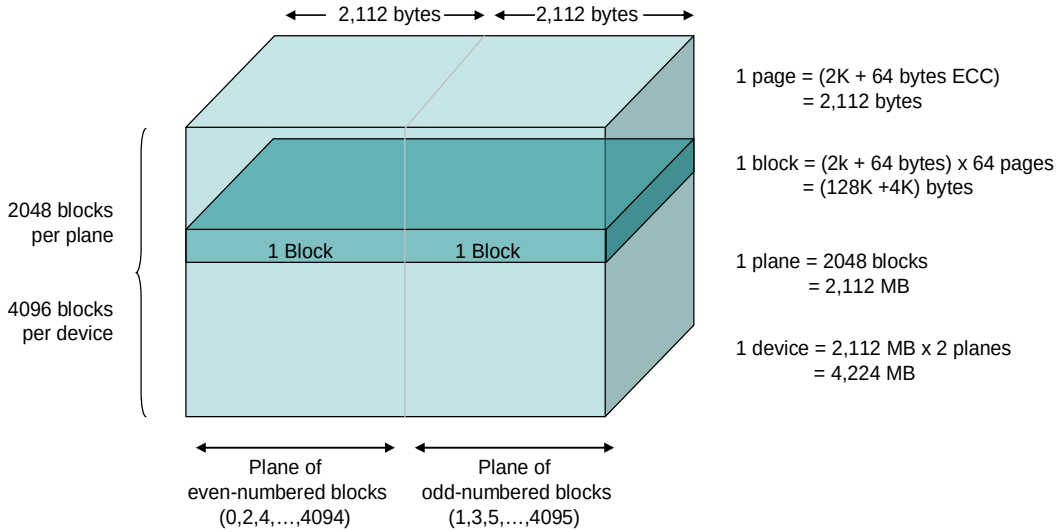
times greater than that of MLC flash. SLC NAND's higher performance and better reliability are preferred for designing solid state drives that meet the requirements of unconstrained workload environments. But there will also be MLC-based SSDs for use in read intensive application environments with limited write requirements.

NAND architecture

NAND flash memory arrays are organized into pages and blocks. A page is the smallest unit. Page size can vary between different NAND implementations, but they are typically 2KB, 4KB, or 8KB. Pages are organized into blocks, with each block typically consisting of 64 pages. We call these units NAND blocks to differentiate them from the 512-byte logical block of the SATA/SAS interface.

SLC NAND also can be implemented in a two-plane architecture that divides the device into two physical planes, consisting of the odd and even blocks. Two-plane flash improves NAND performance by allowing two pages read or programmed concurrently. It also allows concurrent erasing of two blocks. Figure 1 shows a 4GB SLC NAND architecture consisting of 2K pages with 64 pages per block. NAND architecture continues to evolve at a rapid pace, with 8K pages becoming common and four-plane designs on the horizon.

Figure 1. Organization of NAND memory



NAND flash has a specific protocol for writing and retrieving information. The smallest unit that can be read or written is a page. Unlike disk drives, pages that contain existing data cannot be directly overwritten with new data. They are first erased. NAND memory can only be erased in entire NAND blocks, which typically consist of either 64 or 128 pages. One of the more important tasks for any storage device built using NAND flash is effectively managing this asymmetry of the size of writes versus erases. Table 2 provides a list of these basic NAND operations and their execution times.

Table 2. SLC NAND flash operations

Operation	Minimum execution time
Random page read	25 microseconds
Page program (write)	200 microseconds
Block erase	1500 microseconds

As Table 2 shows, writing to NAND flash is a slower operation than reading from it. A page program operation is eight times slower than a random page read. A block erase operation, executed less frequently, is seven times slower than page program operation. Many high- level strategies address this timing disparity; however, it is the primary reason that all NAND-based storage devices, whether USB drives or the more advanced solid state drives, have better read performance than write performance.

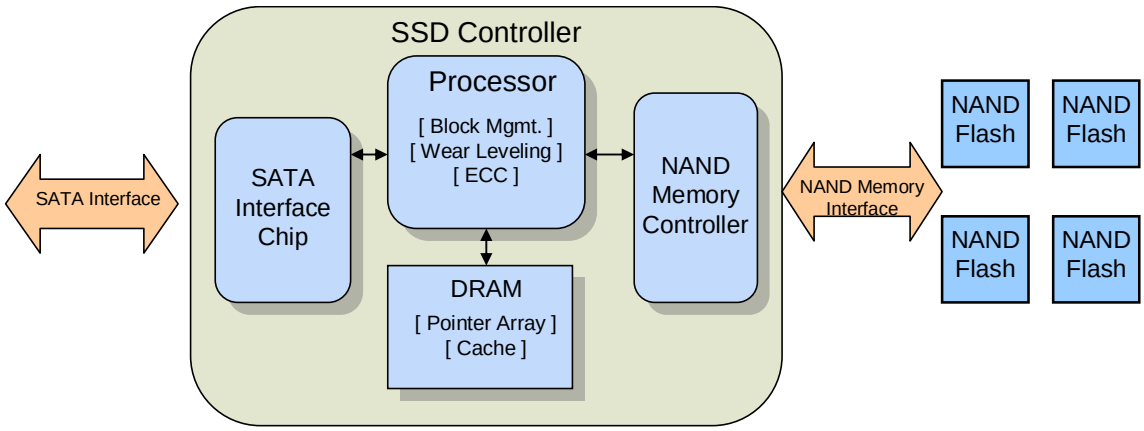
Design of solid state drives with flash memory

A NAND-based solid state drive requires a drive controller subsystem that performs several tasks, including:

- Managing read and write operations to the NAND memory, including error handling and block management
- Enhancing the performance of NAND flash using management algorithms and RAM-based cache
- Maximizing the endurance, or lifespan, of the SSD by employing algorithms to minimize write/erase cycles to the NAND memory
- Providing the translation between the NAND read/write interface and the desired interconnect to the host, typically SAS or SATA

Figure 2 is a functional diagram for a typical SATA SSD. It includes the SSD controller section that contains all of the operational logic necessary to manage NAND flash memory and provide a standard SATA storage interface to the host server.

Figure 2. Functional diagram of a typical SATA solid state drive



Wear leveling for increased SSD endurance

Wear leveling is one of the design techniques engineers use to increase the endurance of NAND-based SSDs. Since NAND-based SLC flash supports only 100,000 lifetime write/erase cycles, the SSD needs to not erase and rewrite NAND blocks any more than is necessary. But application usage may require frequent updates or rewrites of some logical SCSI blocks in a SAS/SATA device. Wear leveling resolves this issue by continuously re-mapping logical SCSI blocks to different physical pages in the NAND array. Wear leveling ensures that erasures and rewrites remain evenly distributed across the medium, maximizing the SSD's endurance. To maximize SSD performance the SSD controller maintains the logical-to-physical map as a pointer array in high speed DRAM. The metadata region in the NAND flash array itself also maintains this information algorithmically. These techniques ensure that the SSD can rebuild the map if you lose power unexpectedly.

Over-provisioning NAND

Design engineers can increase the endurance and performance of an SSD by over-provisioning NAND capacity on the device. Over-provisioning increases the endurance of an SSD by distributing the number of writes and erases across a larger population of NAND blocks Over-provisioning also increases SSD performance by giving the SSD controller additional buffer space for managing page writes and NAND block erases. On higher-end SSDs, NAND memory may be over-provisioned by as much as 25 percent above the stated storage capacity.

HP solid state drives for ProLiant servers

HP introduced the first SSDs for servers in 2008. The SSDs were not hot-pluggable and intended for specific BladeServer environments. In 2009, we introduced the first hot-pluggable SSDs in traditional drive carriers. These 3 Gb/s SATA SSDs are usable across the ProLiant server line, wherever you would use a traditional midline SATA disk drive. Unlike PC-based solid state drives, SSDs for servers meet the higher standards for server storage devices. At the same time, they provide the performance and reliability characteristics associated with SSDs.

Performance of HP server SSDs

Disk access time, or latency, which is the total time required to retrieve data from the drive, influences the performance of a traditional disk drive. Disk drive latency is the sum of the seek time, rotational delay, and transfer time.

With SSDs, there is no seek time or rotational delay. Latency is a function of the memory access and transfer times combined with controller overhead. Given these facts and the knowledge of how NAND flash operates, we can make the following suppositions:

- Read operations should be faster on SSDs than write operations, because of the relative slowness of NAND program (write) operations.
- Random reads on SSDs should be faster than to random reads on disk drives, because the SSDs has no seek time and rotational delay for each read operation.

Table 3 is a side-by-side comparison of the performance of a 32-GB small form factor HP server SSD with that of a 15K Midline SAS hard disk drive (HDD). While performance on sequential operations is comparable between the two drive types, performance on random operations is significantly better for SSDs. In random read performance, the SSD achieved more than 50 times the performance of the HDD.

Table 3. Comparison of typical SSD and HDD performance (actual performance numbers may vary)

	HP 3 Gb/s SATA SSD	HP SFF 15K SAS HDD
Random reads (4 KB)	18,000 IOPs	340 IOPs
Random writes (4 KB)	3000 IOPs	285 IOPs
Sequential read throughput (64 KB)	220 MB/s	105 MB/s
Sequential write throughput (64 KB)	120 MB/s	150 MB/s

HP server SSD reliability

Reliability is an important criterion when selecting a storage device for use in a server. But not all SSDs are more reliable than hard disk drives. Higher reliability SSDs must include controller designs that manage NAND flash arrays while correcting problems caused by the following NAND error modes:

- Read disturbs
- Program (write) disturbs
- Hot charge injection (bad cells can flip bits)

HP solid state drives for servers employ a variety of mechanisms that deliver a high level of reliability required in server environments. They are:

- Longer-lasting SLC NAND technology
- Over-provisioning NAND memory to provide a longer lifecycle
- Wear leveling and block management
- Read and write algorithms that significantly reduce the frequency of NAND error modes
- End-to-end data path error detection
- Surprise power loss protection

With these technologies, HP solid state drives for servers achieve a level of reliability equivalent to or slightly greater than current HP Enterprise disk drives for servers.

Perhaps more important for particular applications, HP server SSDs deliver this level of reliability under conditions that are unsuitable for traditional disk drives, including environments of high-temperatures and those of greater shocks and vibrations. Table 4 compares the operating envelope of an HP server SSD with that of an HP small form factor SAS enterprise drive.

Table 4. Comparison of SSD and HD operating envelopes

	HP 3 Gb/s SATA SSDs	HP SFF 15K SAS HDD
Operating temperature	0° – 55° C	10°– 35° C
Operating shock	1500 g (.5 ms half sine wave)	30 g (2 ms half sine wave)
Vibration	20 g peak 10 – 2000 Hz	1.5 g (RMS) 10 – 500 Hz
Power consumption (active)*	Under 2 watts	8 – 9 watts

*Note: Power consumption will increase as performance levels improve on future devices

Server SSDs for different needs – value, mainstream, and performance

Our current 3 Gb/s SATA solid state drives are just the first building blocks for an HP family of SSDs designed to meet a variety of workload, capacity, and performance requirements. All HP SSDs are Enterprise-class devices because they deliver I/O performance – particularly read performance – that is as good as or better than Enterprise-class disk drives. SSDs differ in the read/write workload levels that they support and their endurance, or expected service life. Today's HP SATA SSDs are considered Enterprise mainstream storage devices. They address workload-constrained environments and have a 3-year service life. The first Enterprise performance SSDs will be available in early 2011. They will address unconstrained workload environments. Enterprise value SSDs will provide relatively large storage capacities at lower costs, but they will not have the endurance of the mainstream or performance SSDs. Table 5 compares the endurance and workload characteristics of the SSD classes that we expect to offer.

Table 5. Comparison of HP solid state drive classes

	Enterprise value	**Enterprise mainstream**	**Enterprise performance**
Interface(s)	3 Gb/s SATA	3 Gb/s SATA 6 Gb/s SAS (Early 2011)	6 Gb/s SAS
General description	SFF and LFF Hot Plug	SFF and LFF Hot Plug	SFF and LFF Hot Plug
Availability	Early 2011	Currently Shipping	Early 2011
Capacities	200 – 800 GB	60 GB and 120 GB 200+ GB in 2011	200 GB +
NAND technology	MLC	SLC MLC in 2011	SLC
Workload	High read/Low write applications	Equal read/write applications	Unrestricted read/write applications
Reliability **Endurance**	1year service life @ constrained write workloads	3 year service life @ constrained write workloads	3-5 year service life Unconstrained workloads
Data Retention	< 1 year	< 1 year	< 1 year
Usage environment	Boot devices Applications high in reads but few or no writes or data is transient	High IO/s applications	Mission critical High IO/s applications

In addition to better workload and endurance characteristics, Enterprise performance SSDs will provide improved throughput and IOPs over the other classes of SSDs. Performance class SSDs are expected to deliver random read performance of 60,000 IOPs compared to 18,000 for today's mainstream SSDs, and should be able to support 450 MB/s of sustained throughput for sequential reads.

I/O accelerators – the new kid on the storage "block"

I/O accelerators are a new class of storage product based around flash memory technologies. As with all modern storage, I/O accelerators present themselves as standard block level storage devices, allowing applications to access them as they would any other storage volume. Unlike SSDs, however, they are not plug-compatible storage components and not accessed through a standard drive

controller and its SAS or SATA channels. An I/O Accelerator is its own controller and storage device, requiring its own specialized driver and delivering block I/O directly across the PCIe bus. As we will see, this gives it some distinct performance advantages for use in particular application environments.

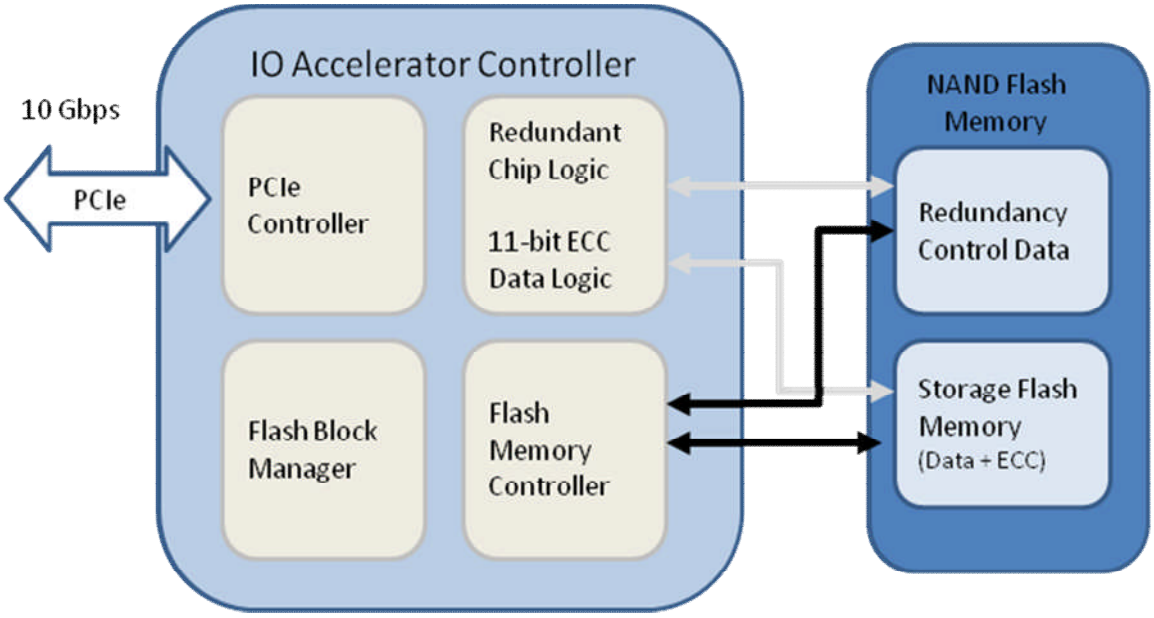
I/O accelerator architecture

I/O accelerators use the same basic NAND memory and NAND memory controller technology as SSDs to perform the low-level storage and retrieval of data to and from flash memory. After that, the similarities end.

As we showed in Figure 2, each SSD uses an onboard processor to perform the translation between the NAND read/write interface and the storage block interface that it presents to a Smart Array host controller. Block data moves across a 3 Gb/s or 6 Gb/s link to the controller before being delivered across the PCIe bus to the calling application.

An I/O accelerator, on the other hand, is its own controller and storage device. It is a PCIe card requiring a device driver. With I/O accelerators, all of the logic that translates standard block level I/O into NAND reads and writes is contained in the accelerator's device driver. The same is true for the NAND management functions, including wear leveling, error correction and bad block management. This architecture allows the I/O accelerator to leverage the server CPUs much greater bandwidth and multi-core processing capabilities to achieve significant improvements in block storage throughput and lower latencies than are possible with onboard processors.

Figure 3. HP IO Accelerator architecture



I/O accelerators also use a wider and flatter array of NAND cells than SSDs. The HP PCIe IO Accelerator uses 8 to 12 NAND channels in its NAND array compared to 4 NAND channels in a typical SSD. This architecture allows our IO Accelerator to perform more NAND accesses in parallel, leading to further performance improvements over typical SSDs.

Finally, the architecture of the HP PCIe IO Accelerator ensures maximum performance by removing intermediaries that could cause bottlenecks between the solid state memory and the PCIe bus. With no

SAS channels and no Smart Array controller front end, the only physical restraint to throughput is the bandwidth of the PCIe bus.

HP PCIe IO Accelerator models

HP delivers the PCIe IO Accelerator in two different product configurations – the ioDrive and the ioDrive Duo. The ioDrive is a low-profile PCIe x4 device available with either 160 GB of SLC NAND memory or 320 GB of MLC NAND.

The ioDrive Duo, at the hardware level, consists of two separate NAND storage devices. The device driver presents a single storage volume to the operating system and applications. This architecture, combined with the wider PCIe x8 interconnect, allow the ioDrive Duo to deliver still better I/O performance that the ioDrive. The 320 GB ioDrive Duo delivers the best performance of all models since it uses SLC NAND combined with the wider bandwidth of the ioDrive Duo architecture.

HP PCIe IO Accelerator performance

The HP PCIe IO Accelerator products provide extremely fast block storage performance for those application environments that can benefit from extremely high IOPs and throughput performance combined with very low latencies. These include, but are not limited to, the following:

- Database and Database acceleration
- Web servers
- Video, rendering, animation

Table 6 provides a look at the key performance characteristics of these devices.

Table 6. IoDrive performance

	HP 160 GB ioDrive	HP 320 GB ioDrive Duo	HP 640 GB ioDrive Duo
NAND Type	SLC	SLC	MLC
Sequential write throughput (32k blocks)	670 MB/s	1.4 GB/s	1.0 GB/s
Sequential read throughput (32k blocks)	750 MB/s	1.5 GB/s	1.4 GB/s
IOPs (4k random reads)	116,000	185,000	122,000
Read access latency	26 microseconds	50 microseconds	80 microseconds

Summary

Solid state storage products that use NAND flash memory are a new, rapidly evolving class of products for ProLiant servers. Today, the primary use for solid state storage products is in application environments with significantly greater random IOPs performance requirements than traditional spinning media can deliver. HP solid state drives are capable of delivering significantly improved I/O performance while seamlessly integrating with the Smart Array storage environment and all of its features, including hardware-based RAID and the ability to hot-plug drives. HP IO Accelerators are a new class of block storage devices designed to provide the maximum possible storage I/O throughput by operating directly across the PCIe bus.

For more information

For additional information, refer to the resources listed below.

Resource description	Web address
HP Solid State Storage web page	www.hp.com/go/solidstate
HP ProLiant drives (including solid state drives)	www.hp.com/products/harddiskdrives
HP PCIe IO Accelerator web page	http://h18004.www1.hp.com/products/servers/proliantstorage/solid-state/index.html?jumpid=reg_R1002_USEN
Comparison of SSD, ioDrives, and SAS rotational drives using TPC-H Benchmark	http://h20195.www2.hp.com/v2/GetPDF.aspx/4AA0-0248ENW.pdf

Call to action

Send comments about this paper to TechCom@HP.com

TC100909TB, October 2010