Understanding endurance and performance characteristics of HP solid state drives

Technology brief

Introduction

As more organizations seek to harness the power of information, the demand for data intensive and transactional workloads such as data warehousing, real-time analytics and virtualized environments is expanding. For that reason, solid state storage technology is becoming more mainstream because it delivers the performance, energy efficiency and high density ideal for these applications. HP solid state drives (SSDs) for ProLiant servers offer significant performance benefits over traditional disk drives for applications requiring high random I/Os per second (IOPs) performance. In addition, HP ProLiant Gen8 servers and drive controllers have been designed to optimize solid state media performance, delivering up to 6 times the performance with SSDs versus previous generations.

Because they are plug-compatible with traditional SATA and SAS drives, we tend to think of SSDs in the same terms as traditional disk drives. But SSDs have unique functional characteristics that require us to re-think some of our usual assumptions when using them in server-based IT environments. This paper provides an overview of two of the unique aspects of SSDs—SSD endurance and SSD performance characteristics in server applications.

SSD endurance

SSDs are compatible with the SAS and SATA interfaces originally defined for reading and writing data to hard disk drives (HDDs). But behind these interfaces, an SSD is a completely different animal. Instead of storing data as magnetic fields on spinning disks, SSDs store data in NAND memory cells. This essential difference profoundly influences both the endurance and data retention characteristics of SSDs when compared to traditional disk drives.

An introduction to endurance

In terms of data storage, endurance refers to the durability of the medium on which the data is stored. How long will the medium last before it wears out and can no longer effectively store data? With disk drives, endurance is rarely an issue. The effective lifespan of the magnetic medium of the disks is typically longer than the time that most disk drives are in service. With SSDs, this is not true. To understand SSD endurance, we first need to review some basics of SSD architecture.

NAND organization

NAND flash memory arrays consist of pages and blocks. Pages are the smallest units of NAND memory that you can address and write to. Page size can vary between different NAND implementations, but they are typically 4 KB or 8 KB. Once you write to a page, you cannot simply overwrite it the same way you could a disk sector. You must first erase its contents. Pages are organized into NAND blocks, which are typically 256 KB to 1 MB in size and should not be confused with the 512 byte logical block of the SATA or SAS interface.

There are two important things to know about NAND blocks. The first is that you can only erase NAND memory at the block level, not at the page level. This means that the SSD controller must relocate and remap any valid data in a block before the controller can erase and write new data to it. The second point is that the lifespan of NAND blocks is limited. They can only be erased and re-written a certain number of times. This is the basic reason for the limited endurance of SSDs.

SLC versus MLC NAND

There are two primary types of NAND memory—single level cell (SLC) and multi-level cell (MLC). SLC NAND stores a single bit in each memory cell, and MLC NAND stores 2 or more bits per cell. MLC is

far more prevalent than SLC, but it has a significantly shorter lifespan in terms of erase cycles. Typically, you can erase and re-write SLC NAND up to 100,000 times before it wears out. MLC NAND comes in both consumer grade (cMLC) and enterprise grade (eMLC) implementations. You can erase and re-write eMLC NAND about 30,000 times before it wears out, while cMLC has a lifespan of about 5,000 erase/write cycles. Clearly, the use of MLC or SLC NAND significantly affects the endurance of an SSD.

Wear–leveling and Over–provisioning

Wear–leveling and over-provisioning are two design technologies that engineers use to increase the endurance of SSDs.

Wear–leveling works by continuously re-mapping the SSD's logical blocks to different physical pages in the NAND array. This helps achieve the goal of evenly distributing NAND block erasures and writes across the NAND array, preventing the premature wearing out of a NAND block and maximizing the SSDs endurance. Wear–leveling is a background task that uses SSD controller cycles but remains invisible to the application reading/writing to the SSD's SATA or SAS interface.

Over-provisioning the NAND capacity on an SSD also increases SSD endurance. It accomplishes this by supplying the SSD controller with a larger population of NAND blocks to distribute erases and writes over time and by providing a larger spare area so that the controller can operate more efficiently.

Minimizing write amplification

Whenever an SSD executes a write of host data, the SSD controller translates this high-level task into a series of NAND operations. In each operation, the controller writes the host data to NAND pages. The controller also performs additional NAND operations to manage and reorganize NAND blocks as required. Write amplification is a ratio of the total size (in MB) of the NAND data writes executed by the controller to carry out a given size (also MB) of host data writes. Lower write amplification ratios are better, and HP strives to maintain lower ratios for its SSDs by using more sophisticated SSD controller firmware. SSDs with higher write amplification ratios sacrifice performance and endurance because of their less efficient management of NAND.

Overall SSD endurance

HP uses the technologies we have discussed, as well as others, to produce SSDs with the highest endurance possible. To provide an array of solid-state storage solutions, we produce SSDs in three different classes—enterprise value, enterprise mainstream, and enterprise performance. Table 1 summarizes the endurance characteristics for each class.

Table 1: General endurance characteristics for classes of HP SSDs

	Enterprise value	Enterprise mainstream	Enterprise performance
Target Workload	High read/Low write applications	Equal read/write applications	Unrestricted read/write applications
Reliability Endurance	3 year service life @ target workloads	3 year service life @ target workloads	3-5 year service life. Unconstrained workloads
Endurance	3k – 5k NAND program/erase cycles	25k – 30k NAND program/erase cycles	100k – 200k NAND program/erase cycles
Usage environment	-Boot devices. -Read-intensive workloads	- High IOPS applications	- Mission critical - High IOPS applications

Because of the limitations of endurance, SSDs—unlike disk drives—have a limited service life in a server. Once an SSD reaches that service life, you should replace it to avoid a potential data loss from continued operation. To assist with this process, HP has developed a new SSD monitoring feature that tracks and reports SSD endurance. It is the HP SMARTSSD Wear Gauge™.

Endurance versus reliability with SSDs

There is a distinction between endurance and reliability. Reliability deals with how often SSDs or a disk drives fail. We usually measure this as Mean Time between Failure (MTBF). MTBF is the number of aggregate service hours, on average, a population of storage devices operate before a failure occurs on any one device. Fortunately, with modern server drive technology, MTBF is typically in the millions of hours. In general, SSDs that have not reached their endurance limit are just as reliable—if not more reliable—than traditional disk drives. But once they have reached their endurance limit, you need to replace them in order to avoid increasing error rates and possible drive failures.

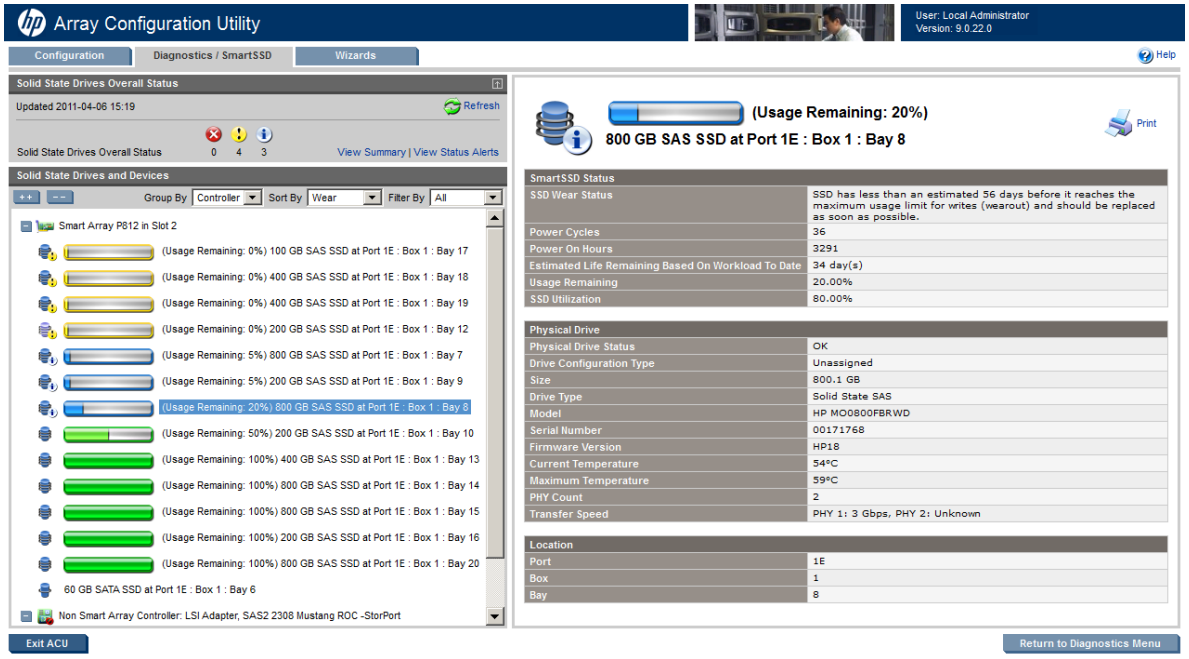
HP SMARTSSD Wear Gauge

In 2011, we implemented the SMARTSSD Wear Gauge. It uses HP-specific data generated by the SSD controller to calculate and report SSD endurance continuously. Various HP storage tools access and report this data, allowing you to monitor SSD endurance in real time.

Integration of the SMARTSSD Wear Gauge into storage tools and utilities

Several storage tools and utilities monitor and report the SMARTSSD Wear Gauge information. The most prominent is the Array Configuration Utility (ACU). Figure 1 shows the Wear Gauge information screen for an array of SSDs attached to an HP Smart Array P812 controller.

Figure 1: This is an example of the SMARTSSD Wear Gauge information displayed by the ACU.



The bar gauges on the left show the Usage Remaining in percentages for each SSD. The SMARTSSD status table at the top right displays detailed information for the highlighted SSD. This information includes the following:

- Power cycles
- Power-on hours
- Estimated life remaining (in days)

The Estimated Life Remaining for the SSD represents the number of days of use remaining for the SSD based on its time in service and the percentage of lifecycle consumed to date. The Estimated Life Remaining number assumes that the server will continue to use the SSD in the same manner going forward. If its workload profile changes significantly, then the Estimated Life reading won't be accurate. As an example, let's consider an SSD that has a stated Estimated Life Remaining of 60 days. If its workload changes abruptly so that its write activity has doubled, then the Estimated Life Remaining would actually trend towards 30 days. Over time, the Estimated Life reading will start to reflect the workload change.

A number of tools and utilities report the SMARTSSD Wear Gauge information, not just the ACU. You can use the following tools:

- Array Configuration Utility—Graphical and command line versions
- HP Array Diagnostic and SMARTSSD Wear Utility—Graphical and command line versions
- System Management Homepage—SNMP Storage Agents

SMARTSSD Wear Gauge alerts and indicators

The Wear Gauge defines certain indicators and status level metrics for an SSD's condition. All the tools that report SMARTSSD status use these indicators consistently. Table 2 shows the indicators and the status levels they represent.

Table 2: SMARTSSD Wear Gauge indicators.

SMARTSSD Wear Gauge	SSD Status
	Drive has sufficient endurance remaining
	Drive has reached one or more of the status metrics indicating its remaining endurance is low. ▪ 56 days of usage remaining at current workload ▪ 5% of usage remaining ▪ 2% of usage remaining
	Drive has reached 0% usage remaining and has been marked with a Predictive Failure.

Reaching 0% usage remaining does not mean that the Solid State Drive will stop operating immediately. It indicates that the SSD has reached the end of its calculated lifespan and that you should replace it to avoid potential data loss.

Integration of SMARTSSD Wear Gauge with Storage Agents

The Storage Agents for various Operating Systems support SMARTSSD data. The agents deliver information on SSD wear status, power–on hours; percentage of endurance used and estimated endurance remaining in days. The HP Systems Management Homepage for the server displays this information. SMARTSSD data is currently supported by the Storage Agents for the following operating environments:

- SNMP Storage Agents for Windows, Linux and VMWare® ESX
- WBEM Storage Agents for Windows and VMWare® ESXi

The storage agents also deliver SMARTSSD status and updates to the OS event logs and via SNMP Traps. The agents send traps whenever the SSD endurance passes one of the four pre-defined milestones.

- 56 days estimated endurance remaining
- 5% endurance remaining
- 2% endurance remaining
- SSD Wear Out (0% endurance remaining)

SSDs and data retention

Data retention is the ability of a storage device to retain data after you remove it from service. SSD data retention characteristics are different from those of traditional disk drives. Three factors influence an SSD's data retention:

- The percentage of the SSD's remaining endurance (lifespan) when you remove it from service.

- The SSD's operating temperature when it was in service
- The temperature you store the SSD at after removing it from service.

The data retention period of an SSD is actually greater when you operate the SSD at higher operating temperatures while it is in service and store it at lower temperatures once you remove it from service. As an example, an SSD operated at 50°C and stored at 30°C should retain its data for 28 weeks if you remove it from service at the end of its rated endurance.

The important thing to remember is that an SSD has a limited data retention window once you remove it from service. This is different from disk drives, which typically retain data for years. If an SSD has used all of its rated endurance, the only truly safe assumption that you should make when removing it from service is that it will not retain its data for any significant period.

Understanding SSD performance characteristics

Because Solid State Drives are compatible with the SAS and SATA interfaces, you can measure their read and write performance using the same tools measuring disk drive performance. But their underlying storage technology is different from that of disk drives. As a result, their performance characteristics are also distinctly different. With SSDs, we need to re-examine our assumptions about storage performance and understand how SSD performance changes in different environments and under different workloads.

Measuring SSD performance

SSDs are capable of delivering exceptional performance, particularly for random I/Os per second (IOPS). You can measure SSD performance by using Iometer or other tools to compare it with that of a disk drive. But you'll discover that an SSD's performance can vary significantly each time you run the same test unless you use the proper methodology. We can attribute these differences to the varying overhead of the background management tasks associated with the NAND memory architecture.

SSD NAND organization and performance

In addition to fulfilling read/write requests, an SSD controller is executing background tasks to manage the NAND memory. They include NAND block management to maintain a pool of free blocks, and data re-mapping tasks associated with wear-leveling. The level of background activity can vary significantly, due in part to the organization of the NAND data and the type of read/write activity going on. The changing level of background activity influences SSD performance.

When running benchmarks on SSDs, all of the following are true:

- **Performance drops by as much 50% when written data starts to fill a new SSD's storage capacity.** As the SSD fills with data, the level of background NAND management activity rises dramatically, increasing overhead and lowering performance.
- **Performance drops within the first few minutes of starting a benchmark.** The level of background activity gradually increases because of the read and write activity of the benchmark.
- **Performance can increase after running a sequential write test.** The test leaves the SSD NAND in a more organized, less randomized state. This lowers NAND management overhead for the next test.

Pre-conditioning an SSD for accurate performance measurements

To obtain accurate, repeatable performance measurements for SSDs, you should first pre-condition them to a steady state that reflects how the SSD will operate in most production environments. At HP, we pre-condition SSDs for benchmark tests using the following procedure:

- We execute 100% sequential write tests at 256 KB request size, which writes the entire capacity of the SSD at least twice. For a 200 GB drive with a throughput of 100 MB/s, we run two 33-minute test passes. As a result, the SSD is fully written and properly conditioned.
- We run several hours of 100% random writes tests at 8 KB request size and 4 KBKB aligned. We continue to run the tests until the IOPS performance drops and then stabilizes.

This procedure ensures that an HP SSD has reached a steady state in terms of its operational overhead. It also ensures that performance test results after the pre-conditioning reflect SSD performance in a real world application environment.

SSD Latency

With any data storage device, latency is the time it takes to execute a read or a write command. In benchmarks and in real life, we measure latency as the average latency over a given period while executing a predetermined profile of read commands, write commands, or both. Naturally, average latency varies depending on the size (4 KB vs. 256 KB) and type (random vs. sequential) of commands executed. Latency also varies depending on the mixture of read versus write commands in the workload.

Origins of SSD latency

At first, you may be slightly surprised to see a discussion of SSD latency. After all, with traditional disk drives the head seek time and the rotational latency of the disk drives are the primary contributors to overall latency. SSDs have no rotational latency, but they do have latency. It is simply from different sources than for disk drives.

With SSDs, latency comes primarily from the processing overhead associated with managing and executing individual NAND operations. These operations are required to fulfill the higher-level host read or write. This includes any or all of the following:

- Managing the contention for the limited number of channels between the NAND controller and the NAND flash
- Translating host logical addresses into physical NAND memory addresses
- Executing the individual NAND reads or writes needed to complete a command
- For writes, erasing NAND blocks before a write can be completed
- Executing general NAND background management activity, including the NAND block management associated with wear-leveling

SSD writes tend to incur a greater overhead than reads. That's because writes tend to generate NAND block management activity in the SSD controller, where simple reads do not. As a result, SSD performance in standardized benchmarks such as Iometer tests will tend to decrease as the percentage of writes in the test increases. As Table 3 illustrates, average SSD latency remains significantly lower than that of HDDs.

Table 3: Typical average latencies for SSDs versus HDDs

Iometer benchmark (70%/30% read/write, Queue=16)	**Typical Average Latency Enterprise Mainstream SSD**	**Typical Average Latency SAS HDD**
8 KB Random	.55 ms	3.0 ms

Using SSDs with Array Controllers

SSDs perform much better than HDDs in applications requiring large, random IOPS performance. This advantage extends to the use of SSDs RAIDed behind Array Controllers. But there are some subtleties to keep in mind, including how various RAID levels affect SSD performance differently than HDDs.

Array performance scaling with SSDs

Just as IOPS performance scales when using multiple HDDs behind an HP Smart Array controller, the same is true for multiple SSDs. An individual SSD, of course, is capable of delivering thousands of random IOPS compared to hundreds for the highest performing HDD. But IOPS will still scale linearly when you add SSDs behind a Smart Array RAID controller. With typical workloads, random IOPS performance for an array controller such as the HP Smart Array P410 scales linearly up to six SSDs. Past this number, the performance starts to become constrained by the throughput capability of the array controller. It can process a maximum of 50,000 to 60,000 I/O operations per second. Because of this, we recommend a maximum of eight SSDs behind an array controller when you are using the current generation of SmartArray controllers. This number increases significantly with the Gen8 SmartArray controllers. They are capable of processing up to 200,000 I/O operations per second.

SSDs and RAID levels

You can use SSDs with Smart Array controllers in redundant RAID configurations. In RAID 5 and RAID 6 configurations, SSDs have several unique advantages. Both RAID 5 and RAID 6 require the Smart Array controller to perform two read-modify-write operations to the back-end array drives for every host-level random write it executes. With HDDs, the performance penalty for using RAID 5 and RAID 6 is severe. Each read-modify-write operation requires an extra revolution of the disk media, causing the system to incur the worst possible latency. This is why you often use RAID 1 or RAID 10 with HDD arrays. They achieve redundancy without suffering as great a performance penalty. With SSDs, however, there is no rotational latency. As a result, you can create RAID 5 or RAID 6 arrays using SSDs that have about the same performance as RAID 1 or RAID 10 arrays.

With any redundant RAID level, a host-level write operation results in multiple low-level writes to the drives. SSD write operations typically have a higher overhead and generate more background NAND management activity than SSD read operations. In redundant RAID configurations using SSDs, this performance difference is magnified. A redundant array using SSDs may deliver better performance than an HDD-based array. But the relative fall off in performance as the write load increases is greater for the SSD-based configuration.

SSD and SmartArray array accelerator cache

For HP SmartArray controllers, the Array Accelerator cache module typically delivers significant performance improvements for logical volumes consisting of traditional disk drives. The dramatic decreases in latency for both writes and read hits far exceed the processing overhead for determining cache hits and misses on reads and for aggregating cached writes. When the logical volume consists of SSDs, this may not be the case.

For logical volumes consisting of SSDs, we generally recommend disabling the array accelerator cache. The overhead of the cache lookup typically outweighs the more marginal latency gains that the accelerator cache provides over the SSDs' natively low latencies. This guideline is not absolute. Some applications may actually benefit from using the array accelerator, and in those cases, the most advantageous cache ratio is usually 0% read/100% write. But nearly all synthetic benchmarks will show better performance with the array accelerator disabled.

With SSDs and SmartArray controllers, you should still install a cache memory module even if you don't activate the array accelerator function. Without the memory module installed, the array controller operates in Zero Memory RAID mode (ZMR). In ZMR, the controller code runs from non-volatile memory

on the card rather than the much faster SDRAM on the memory module. This significantly affects command latency and overall storage performance.

SATA versus SAS SSD performance

HP offers both SATA and SAS Solid State Drives. In general, the SAS interface is more robust and features better error detection and error processing. All of our highest performing SSDs, the enterprise performance class drives, are SAS-based. But SATA SSDs may meet your performance requirements applications.

Some enterprise mainstream SATA SSDs deliver surprisingly good performance in lower workload environments. But SAS SSDs consistently outperform SATA SSDs for IOPS in higher workload applications. This difference is more a function of the speed and robustness of the SAS processing core on the SSD than of the SSD back-end architecture.

You should almost never use SATA SSDs behind SAS expanders. When you use an SAS expander, a relatively large number of storage devices on the expander's backside share the limited number of dedicated SAS channels on the expander's front side. SAS devices will only take control of a shared channel when they have data to transfer. On the other hand, SATA drives, including SATA SSDs, take control of a channel for the entire request/transfer cycle. They relinquish it only when they have completed the entire operation. This major difference between the protocols significantly affects performance when many SATA devices vie for access to the SAS channels.

HP IO Accelerators for ProLiant servers

HP IO Accelerators deliver the fastest solid state storage solution for HP ProLiant servers. Unlike SSDs, which operate through the SATA or SAS interface of HP SmartArray controllers and Host Bus Adapters (HBAs), IO accelerators provide solid state storage that is accessed directly across the PCI bus. IO accelerators are ideally suited for environments requiring very high random read performance (greater than 100,000 IOP/s) as well as significantly higher read and write throughput (up to 600 – 700 MB/s).

Conclusion

HP Solid State Drives are a storage solution ideally suited for certain types of server applications—particularly those requiring superior random IOPS performance. You can use SSDs wherever you would use a disk drive. But SSDs have some distinctly different performance characteristics. You should consider them before deploying SSDs in a particular application environment. Additionally, SSDs have a shorter lifespan, or endurance, than enterprise-level disk drives. Treat them as consumables to replace when they reach the end of their usable lifespan. To help maximize an SSD's lifespan, we have developed HP SMARTSSD Wear Gauge technology for monitoring SSD usage and wear, and then integrating this information into all HP management tools.

For more information

Visit the URL listed below if you need additional information.

Resource description	Web address
HP Solid State Storage Technology Technology brief, 2nd edition	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c01580706/c01580706.pdf
Innovative Technologies in HP ProLiant Gen8 servers Technology brief	http://h20000.www2.hp.com/bc/docs/support/SupportManual/c03227849/c03227849.pdf
HP Solid State Storage Technology Homepage	www.hp.com/go/solidstate

Send comments about this paper to TechCom@HP.com


Follow us on Twitter: http://twitter.com/ISSGeekatHP

TC1109885, April 2012