Hewlett Packard Enterprise

Understanding endurance and performance characteristics of HPE solid-state drives

Contents

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. HPE solid-state drives (SSDs) for HPE servers offer significant performance benefits over traditional disk drives for applications requiring high random I/O operations per second (IOPS) performance.

Because they are plug-compatible with traditional Serial ATA (SATA) and Serial-Attached SCSI (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 rethink 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 SAS and SATA interfaces, as well as the new non-volatile memory express (NVMe) interface. The SAS and SATA interfaces were originally designed for reading and writing data to hard disk drives (HDDs), and over time, evolved into supporting SSDs. If you look past the interfaces and compare HDDs to SSDs, you will quickly notice that an SSD is completely different. 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 write. Page size can vary between different NAND implementations, but they are typically measured in 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 measured in MB 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 rewritten a certain number of times and still be successfully read. This is the basic reason for the limited endurance of SSDs.

NAND types

While NAND memory started storing on bit per cell, the ever-growing need for cheaper and larger capacities of NAND has driven engineers to develop methods to store more bits per cells. Nowadays, the vast majority of NAND is multi-level cell (MLC). There are two primary types of MLC NAND memory: traditional MLC, which supports 2 bits per cell; and tri-level cell (TLC), which supports 3 bits per cell. The increase in bit storage capacity does come with a downside; it decreases the number of times a cell can be written and still successfully read. This specification is called the Program-Erase (PE) specification. The PE specification is only a part of the properties that determine the endurance of an SSD. Along with these general-type categories of NAND, the industry also produces different quality of NAND within these groups. For example, NAND fabs produce enterprise versions of TLC and consumer-grade TLC with different PE and endurance characteristics.

Wear-leveling and overprovisioning

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

Wear-leveling works by continuously remapping the SSDs 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 and can impact the performance of the SSD, but it is mandatory to have a functioning device.

Overprovisioning 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 unseen by the user 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 Hewlett Packard Enterprise 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. In general, the larger and more sequential a host workload is, the lower the write amplification.

Overall SSD endurance

HPE 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: Read Intensive (RI), Mixed-Use (MU), and Write Intensive (WI). HPE uses a metric called Drive Writes Per Day (DWPD) to help classify the different endurance characteristics of each device. DWPD is the number of times the capacity of the device can be written with a particular workload every day for a certain amount of time. At HPE, our DWPD number uses a random 4 KiB transfer size for the workload over five years. For example, consider a 400 GB drive with a DWPD of 1, a host can write 400 GB of random 4 KiB data every day for five years and not wear out the drive.

Table 1. General endurance characteristics for classes of HPE SSDs

	Read Intensive	**Mixed-Use**	**Write Intensive**
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 DWPD	<= 1	> 1 < 10	=> 10
Usage environment	• Boot devices • Read-intensive workloads	• High IOPS applications	• Mission-critical workloads • High IOPS applications

Because of the limitations of endurance, SSDs—unlike hard 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, HPE has an SSD monitoring feature that tracks and reports SSD endurance, called the HPE SmartSSD Wear Gauge. The HPE SmartSSD Wear Gauge uses HPE specific data generated by the SSD controller to calculate and report SSD endurance continuously. Various HPE storage tools access and report this data, allowing you to monitor SSD endurance in real time.

Endurance versus reliability with SSDs

There is a distinction between endurance and reliability. Reliability deals with how often SSDs or 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. However, once they have reached their endurance limit, you need to replace them in order to avoid increasing error rates and possible drive failures.

In order to get the most endurance and reliability out of our SSDs, HPE runs an industry-leading qualification program, which rigorously tests drives in servers under a variety of real-world conditions for over 3.35 million hours,[1] ensuring that customers receive the reliability they've come to expect from HPE SSDs.

[1] h20195.www2.hpe.com/v2/GetPDF.aspx/4AA4-7291ENW.pdf

HPE SmartSSD Wear Gauge Report

The HPE SmartSSD Wear Gauge Report uses the power-on hours and the percentage used of an SSD's to calculate the expected life remaining of SAS and SATA SSDs (NVMe SSDs currently not supported). Various HPE storage tools access and report this data, allowing you to monitor SSD endurance in real time:

- HPE Smart Storage Administrator (SSA) GUI (only supports SR controllers)
- HPE SSA CLI
- HPE MR Storage Administrator (MRSA) (only supports MR controllers)
- HPE Integrated Lights Out (iLO) RESTful interface

Figure 1 shows the HPE SmartSSD Wear Gauge information screen for an array of SSDs attached to an HPE Smart Array P408i-a SR Gen10 controller.

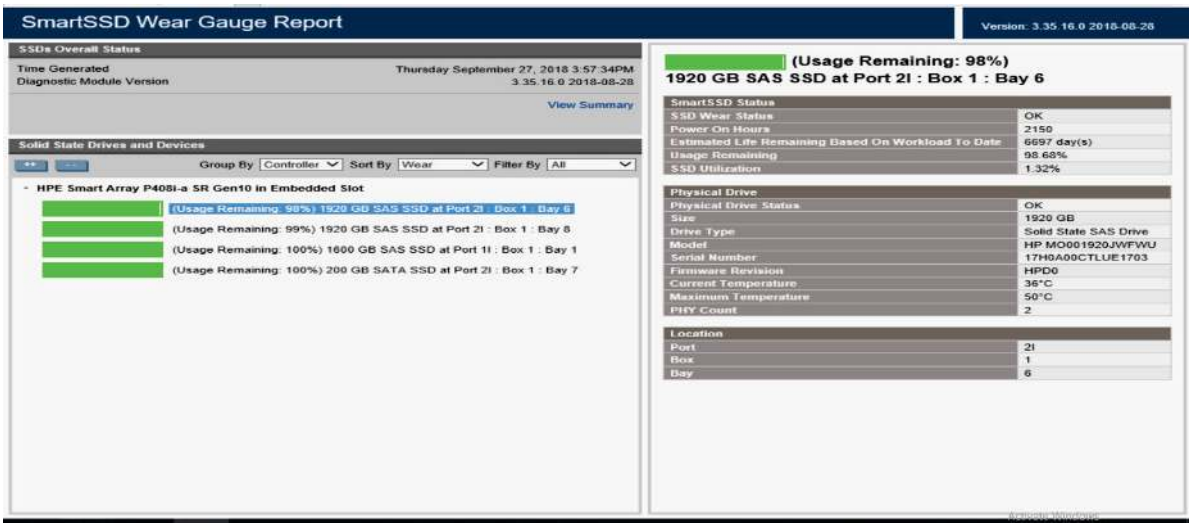

Figure 1. Example of the HPE SmartSSD Wear Gauge Report

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

- Power On Hours
- Estimated Life Remaining (in days)
- Usage Remaining (in percentage)
- SSD Utilization (in percentage)

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 toward 30 days. Over time, the Estimated Life reading will start to reflect the workload change.

HPE SmartSSD Wear Gauge alerts and indicators

The HPE SmartSSD Wear Gauge Report 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. HPE SmartSSD Wear Gauge indicators

HPE SmartSSD Wear Gauge indicator	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 SSD 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.

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 SSDs remaining endurance (lifespan) when you remove it from service.
- The SSDs 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. In general, higher storage temperature negatively affects the nonpower on storage retention. The industry, through the JEDEC standards body, generally specifies end of life retention to be three months at 40°C for enterprise SSDs.

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 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 SSDs are compatible with the NVMe, 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 reexamine 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 IOPS. You can measure SSD performance by using a benchmarking tool such as IOmeter to compare it with that of a HDD. What you'll discover is 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.

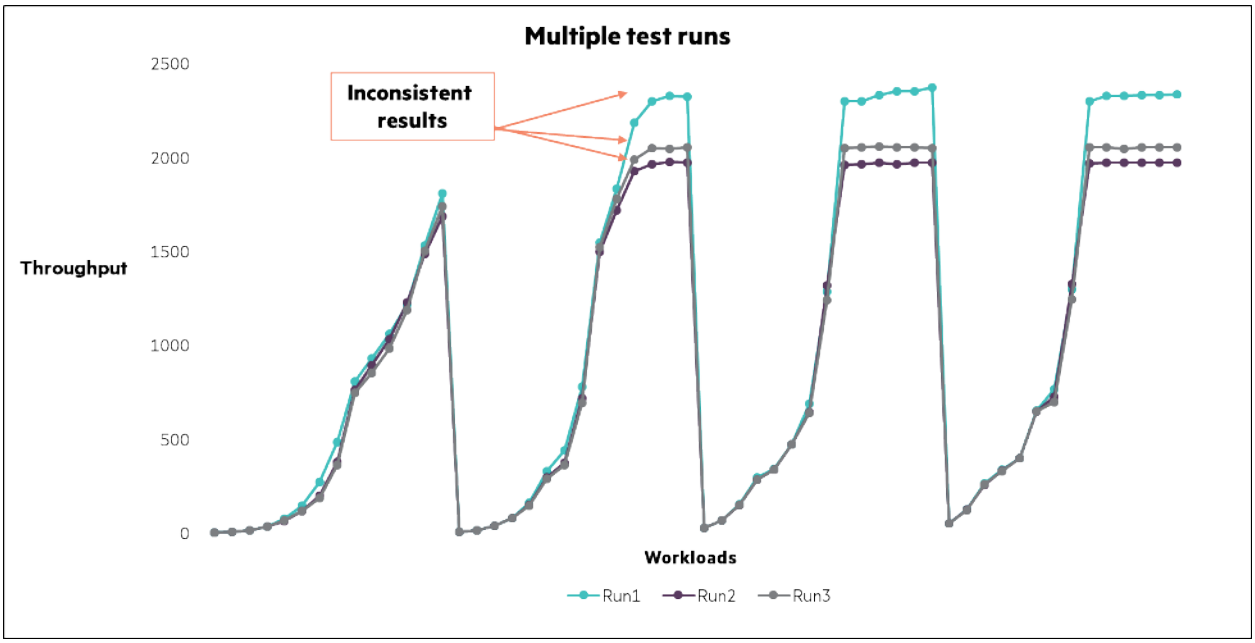

Figure 2. Sequential performance without steady state

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 can drop by 50% or more when the data written starts to fill a new SSDs storage capacity. As the SSD fills with data, the level of background NAND management activity rises, increasing overhead and impacting performance.
- Performance drops once there are no more free or empty pages to store data on the media at which time background NAND management activity begins to run more frequently. The level of background activity gradually increases because of the read and write activity of the benchmark. The worst-case is typically when the drive has been written to with random write workloads using small transfers (~4 KiB or smaller) causing significant write amplification.

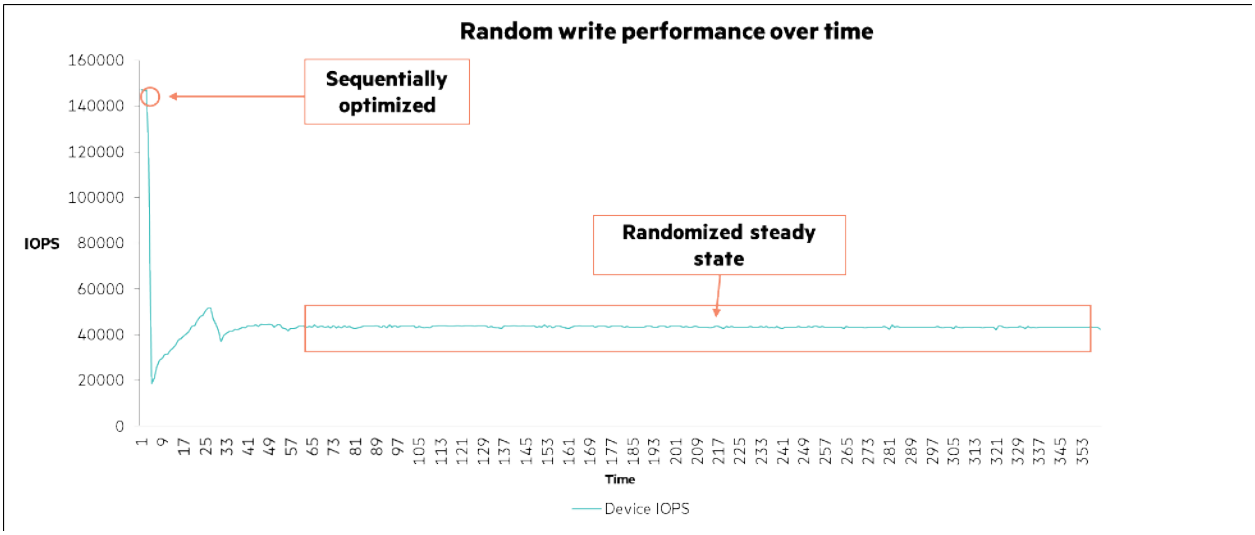

Figure 3. Randomized worst-case steady state—Optimized to randomized

- Performance can increase after running a sequential write test. The test leaves the SSD NAND in a more organized, sequentially optimized state. Sequential write workloads have very low write amplification and as a result provide much less negative impact to performance and NAND management efficiency.

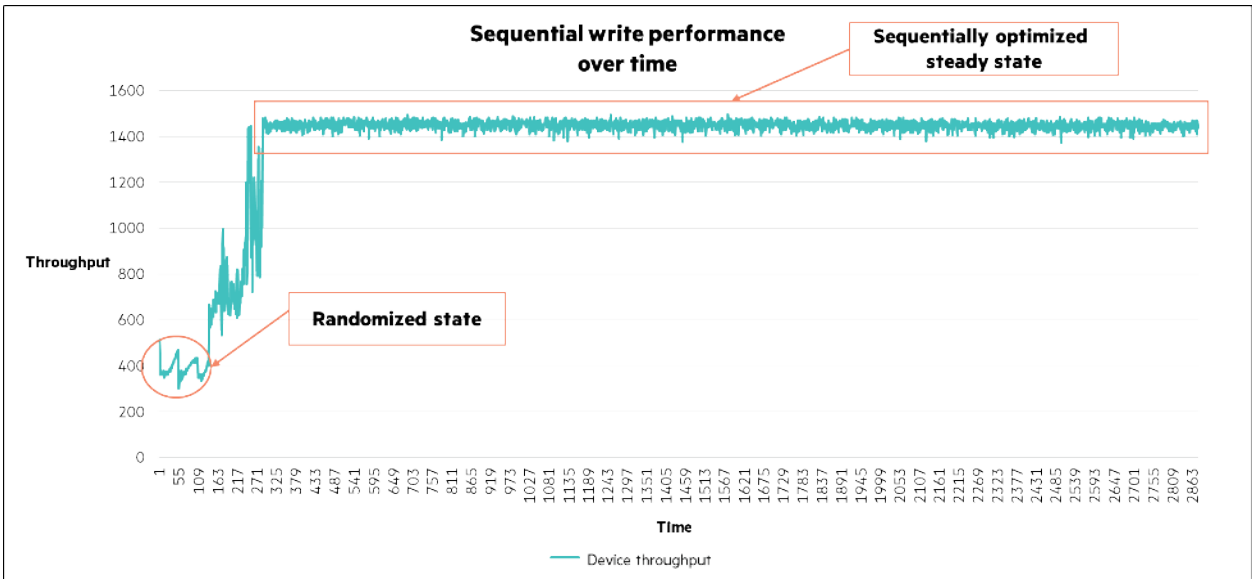

Figure 4. Sequentially optimized steady state—Randomized to optimized

Preconditioning an SSD for accurate performance measurements

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

- After a low-level format of the device, we execute 100% sequential write tests at 512 KiB request size over the entire advertised user capacity of the SSD twice.
- We run several hours of 100% random writes tests at a 4 KiB request size and 4 KiB aligned. We continue to run the tests until the IOPS performance drops and then stabilizes to a steady state.

This procedure ensures that an HPE SSD has reached a steady state in terms of its operational overhead. It also ensures that performance test results after the preconditioning are consistent and reproducible.

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 and max. Latency over a given period while executing a predetermined profile of read commands, write commands, or both. Latency can differ depending on a variety of variables. They consist of, but are not limited to, things such as request size, mixture of read and write I/O, intensity of workload or queue depth, and other background I/O.

Origins of SSD latency

With traditional disk drives, the head seek time and the rotational latency of the disk drives are the primary contributors to overall latency. SSDs do not function mechanically like HDDs, but they do have latency. The primary difference is the media used and how it is managed.

With SSDs, latency comes primarily from the processing overhead associated with managing and executing individual NAND operations. These operations are required to fulfill high-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
- 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 15K HDD
4 KiB random	0.55 ms	36 ms

Using SSDs with HPE Smart Array controllers

HPE SSDs are suited to enterprise environments with highly random data under a variety of write-workload applications. SSDs provide significantly better random read and write IOPS compared to HDDs. While sequential read and write throughput is also improved over HDDs, the greatest benefit is recognized in random data applications. As a result, these high-performance, low-latency, and low-power SSDs provide significant system benefits for applications that previously overprovisioned HDD capacity to achieve better performance. This advantage extends to the use of SSDs when in a RAID volume when connected to an HPE Smart Array controller.

Array performance scaling with SSDs

IOPS will still scale linearly when you add SSDs behind an HPE Smart Array RAID controller. The scaling is dependent on the type of workload. With random read workloads, the IOPS performance can scale linearly up to eight SSDs. Past this number, the performance starts to become constrained by the throughput capability of the array controller as it reaches up to the maximum throughput of 1.6 million IOPS. Sequential read workloads scale linearly up to six SSDs. While sequential write operations scale linearly up to six to eight SSDs dependent on the RAID level, both sequential reads and writes are constrained by the throughput of the array controller up to a maximum throughput of 6900 MB/s.

RAID levels

You can use SSDs with HPE Smart Array controllers in redundant RAID configurations, including RAID 0, 1, 5, 6, 10, 50, 60; 1 Advanced Data Guarding (ADM); and 10 ADM (MR controllers do not support ADM.). During random read and sequential read workloads, the I/O performance (1.6 million IOPS) and MB/s performance (6900 MB/s) is comparable in all RAID levels respectively. During random write and sequential write workloads, performance is maximized with RAID 0 and performance declines with advanced RAID levels as data is written on a disk for mirroring or creating a parity.

HPE SSD Smart Path and HPE Smart Array MR Fast Path

SSDs require special tactics to capture the full advantage of their low-latency capabilities. HPE SSD Smart Path (for SR controllers) and HPE Smart Array MR Fast Path (for MR controllers) enable the high performance of SSD-based logical volumes by allowing certain types of I/O requests to take a more direct path to the physical disks, bypassing most of the firmware layers of the RAID controller. HPE SSD Smart Path and HPE Smart Array MR Fast Path are enabled by default when you create an array. This process accelerates reads for all RAID levels and writes for RAID 0. HPE SSD Smart Path and HPE Smart Array MR Fast Path will not run in conjunction with the flash-backed write cache enabled on the controller.

HPE Smart Array SR SmartCache and HPE Smart Array MR CacheCade

HPE Smart Array SR SmartCache and HPE Smart Array MR CacheCade are controller-based caching solutions that cache the most frequently accessed data (hot data) onto lower latency SSDs. This helps dynamically accelerate application workloads and increases storage performance overall.

NVMe, SAS, and SATA SSD performance

HPE offers NVMe SAS and SATA SSD options. The highest performing SSD products are available in our NVMe offerings. They typically have the highest performance in comparison to SAS and SATA offerings of the same endurance class with the exception of some cost-optimized mainstream NVMe offerings.

Our SAS offerings provide performance that is typically lower than our high performance NVMe offerings but higher than SATA SSDs when comparing products in the same endurance class. The SAS protocol also enables our products to perform more efficiently than SATA products when configured behind expanders.

SATA provides a more mainstream cost-optimized group of options that do not typically have the performance of SAS and NVMe products but are typically more affordable. They still provide significantly higher random I/O performance and lower latency than HDDs can typically produce.

You should try to avoid using 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.

Conclusion

HPE SSDs are a server storage solution ideally suited for certain types of server applications—particularly those requiring superior random IOPS performance and lower latency than HDDs can produce. You can use SSDs wherever you would use a disk drive, but it is important to remember that SSDs have some distinctly different performance characteristics, which should be considered before deploying them in a particular application environment. Additionally, SSDs have a shorter lifespan or endurance, than enterprise-level disk drives. To help maximize an SSD's lifespan, HPE SmartSSD Wear Gauge helps monitor SSD usage and wear in tandem with HPE SSA, providing drive status alerts before a drive fails to ensure data integrity. Remember, HPE's industry-leading qualification program rigorously tests drives in servers under a variety of real-world conditions for over 3.35 million hours, assuring customers of the performance, endurance, and reliability they've come to expect.

Resources

Visit the URLs listed in the following for additional resources and information.

Resource description	Web address
HPE Solid-State Drive Selector Tool	ssd.hpe.com
HPE Server Storage home page	hpe.com/info/serverstorage
Driving HDD and SSD value with HPE Firmware (white paper)	h20195.www2.hpe.com/v2/GetPDF.aspx/4AA4-7291ENW.pdf
Refresh servers, consolidate workloads, and cut data center costs (brochure)	hpe.seismic.com/Link/Content/DCPO7Zg2ZkkE2Golhb16u6lw

Learn more at
hpe.com/info/serverstorage

a00076380ENW, July 2019