

Top Considerations for Enterprise SSDs

Contents

Introduction1

Form Factors 1

Interface Options 3

Endurance Considerations 4

Error Handling, Power Protection,
and End-to-End Data Protection 5

NAND Types 6

Benchmarking 6

Power of Overprovisioning 8

Monitoring and Management 8


Conclusion 8

Introduction


Evaluating solid state drives (SSDs) for enterprise applications can be daunting. Get the form factor wrong and it may not fit in your preferred server. Select the wrong interface type and you can artificially limit maximum performance, slowing your applications. Choose the wrong endurance rating and you may find yourself replacing the SSD in several months' time. After these basic choices are made a whole array of other specifications can make or break your SSD implementation. Manufacturer datasheets, unfortunately, often impair your ability to make decisions by offering data that was derived from different testing methodologies or burying the user in a vast sea of data points.

This paper enables users to make better-informed choices about the correct SSDs for their applications. We will methodically examine SSDs based upon the physical form factor, the logical interface, the flash technology, various write endurance differentiators, and more. Developing an understanding of each of these factors will make the choice of enterprise SSDs a less daunting and more successful operation.


At A Glance: Top 8 Considerations When Selecting Enterprise SSDs




Form Factor
Main options include 2.5-inch drive, M.2, Add-In-Card, and EDSFF




Interface
Main options include SAS, SATA, and NVMe




Endurance
High endurance, medium endurance, or read-intensive




Error Handling and Data Protection
Embedded technology to improve reliability




NAND
SLC, MLC, TLC, or QLC using either 2D or 3D manufacturing



Performance
IOPS, throughput, latency, and quality of service at macro and micro levels



Power
Configurable power modes to optimize large-scale deployments



Monitoring and Management
Maintain and manage large deployments easily

Form Factors

Why It's Important: The form factor defines where the SSD fits, whether it is possible to replace it without powering down the server, and how many SSDs can be packed into a chassis. A wide array of form factors is available, each with specific strengths and weaknesses that vary depending on your needs.

Because there are no moving parts in SSDs, they can be found in more unique physical form factors than their hard drive counterparts. You may find different versions of the same SSD, with similar performance in different form factors, to work best with your infrastructure.

2.5" Drive

The most common form factor is a 2.5" drive, also known as Small Form Factor (SFF), or U.2. This form factor defines the width and length of the SSD, but be aware that there are multiple heights available. Laptop and consumer SSDs often are 7mm in height. Enterprise SSDs can be 7mm, 9mm or 15mm in height. In the context of a server installation, the height is relatively unimportant, as the vast majority of servers support any of these heights. The deeper vertical dimension or Z-height provides more volume for flash chips and controllers, potentially increasing performance, cooling, and capacity. The 2.5" form factor can support SATA, SAS, and even NVMe™ interface technologies (x4 or 2x2, as described in the following section).

The 2.5" form factor, shown in Table 1, is present in many server and JBOD front panels, allowing for hot-swapping of SSDs without powering down the server. In a 1U server, up to 10 drives may be accessible in a front panel, with 24 slots potentially available on most 2U servers.

Add-In Card (AIC)

Another common form factor, shown in Table 1, is that of the Add-In Card (AIC), also referred to as HH-HL (half height, half length). It consists of a board that connects to a PCIe slot inside the server chassis. Only NVMe SSDs, which communicate natively using PCIe, are available in this form factor. Because they live inside the server chassis, AIC-form-factor SSDs are not hot swappable. They may, however, have a higher bandwidth and power profile than 2.5"-form-factor versions of the same SSD because the communications bus is wider (x8 or x16) and the power delivery capability of the PCIe slots is generally higher.

M.2

Another SSD form factor becoming more prevalent in data center environments is M.2. This is a long, thin bare-card form factor that attaches directly to the motherboard and typically uses NVMe or SATA to communicate. Its length is variable, and it may also have components on one or both sides. The "XX" portion of the size identifier M2-22XX identifies the length in millimeters, with common sizes being 42, 60, 80 or 110mm. Table 1 also features an M2-2280 M.2 SSD form factor. Attachment is often difficult, and these SSDs definitely are not hot-swappable. NVMe versions use either x2 or x4 PCIe lanes, while SATA-based M.2 SSDs use standard SATA-III signaling. Due to their smaller size and often awkward locations, thermal management may be a significant issue for sustained performance of M.2 drives. These small drives often have insufficient surface area to dissipate heat, resulting in thermal throttling and poor long-term performance stability. A common use case for M.2 drives is as a boot device where the host is mainly reading from the device.

EDSFF

A new addition to the SSD form factor list is the Enterprise & Data Center SSD Form Factor (EDSFF). There are 3 main variants of EDSFF: E1.L, E1.S, and E3. The E1.S form factor is a little longer and wider than an M.2 and targeted toward the 1U compute-optimized server design. The E1.L is optimized for 1U storage servers to enable larger density in a small server footprint, and it helps increase petabytes-per-server rack storage in cloud infrastructures. The EDSFF 3" version, the E3, comes in 4 sub variants: E3 Short-Thick, E3 Short-Thin, E3 Long-Thick, E3 Long-Thin. The E3 FF is designed for 2U-server and storage designs. The EDSFF uses the same connector specification (SFF-TA-1002) across all form factors and uses the PCIe NVMe protocol.

Key benefits of the EDSFF include:

- Higher capacity and density per square foot in the data center than the U.2 or M.2
- Common-connector standardization for future scalability in generic systems
- Compelling airflow advantages and heat dissipation capability without thermal throttling
- Support for various power levels to optimize in system thermal and performance balance

For more information, see <http://www.snia.org/sff/specifications>.



| Form Factor: | 2.5" Drive (U.2) | Add-In-Card (HH-HL) | M.2 | EDSFF (E1.L) |
|---------------------|---------------------------|--------------------------------------|--------------------------------------|---|
| Dimensions | 70x100mm 7-15mm height | 65x170mm | 22x30-110mm | 38.4mm x 318.75mm x 9.5mm (25W) / 38.4mm x 318.75mm x 18mm (40W) |
| Typical Power | 11-20W | Up to 25W | < 8W | < 8W |
| Hot-Swappable | Yes | No | No | Yes |
| Front Serviceable | Yes | Maybe | Maybe | Yes |
| Typical Drive Loads | Up to 24 | 4-6 (depends on PCIe lanes from CPU) | 1-2 (depends on PCIe lanes from CPU) | Up to 32 SSDs |

Table 1: Enterprise SSDs are available in a variety of form factors.

Interface Options

Why It's Important: The interface is the electrical and logical signaling between the SSD and the CPU. It defines the maximum bandwidth, minimum latency, expandability, and hot-swap capability of the SSD.

The interface is the logical protocol that the SSD uses to communicate with the host. There are three basic interface options for SSDs today:

SATA (Serial ATA), SAS (Serial Attached SCSI), and NVMe (PCIe). Unlike SATA and SAS interfaces on enterprise SSDs, which are generally only available in the 2.5" form factor, NVMe is available in 2.5", add-in-card, M.2, and EDSFF form factors. Each interface was developed with a specific audience in mind: SATA for the cost-conscious home user, SAS for the enterprise user who required capabilities like multipath to support high-availability access, and NVMe for performance applications due to its focus on the lowest latency and highest bandwidth. SATA and SAS can support both SSDs and HDDs, while NVMe is generally an SSD-only protocol.

SATA Interface

SATA is generally the least expensive, least expandable, least highly available, and lowest-performance interface for SSDs. The latest generation of SATA, SATA-III, provides a maximum of around 600 MB/s transfer rate and is hobbled in its latency due to the legacy protocol

optimized for rotating media. No high availability is possible on the interface, so users who need to survive link or controller failure have to resort to very low performance application-level replication or other strategies. SATA also does not generally support high levels of expansion, with most servers having the ability to support fewer than 6 SATA devices. However, given its lower cost and complete support by motherboards and chipsets, SATA is very useful for things such as boot devices or scale-out NoSQL databases that already implement the necessary application logic to ensure data availability through replication. Data protection in the form of RAID can be implemented by most major operating systems at the software level.

SAS Interface

SAS provides a significantly more robust enterprise feature set, including dual ports, expander capability, and higher data rates. A state-of-the-art SAS12G interface can support over 1 gigabyte/s over each of its two links. Those links can be connected to different cabling, controllers, and even servers, allowing for a single hard drive to fail over to another server should the primary server fail. Enterprise SAN and NAS arrays often require this feature, which is why SAS drives are found in most vendors' offerings. SAS drives often require a special host bus adaptor (HBA) or RAID card to support them. HBAs simply provide protocol support, while RAID cards often implement read and battery-backed write caches, as well as hardware RAID and RAID recovery offload. Another useful feature of the SAS protocol is its support for large-scale expansion. JBODs with 24 or more drive sleds are commonly available, allowing for massive quantities of SSDs to be connected to a single server.

NVMe Interface

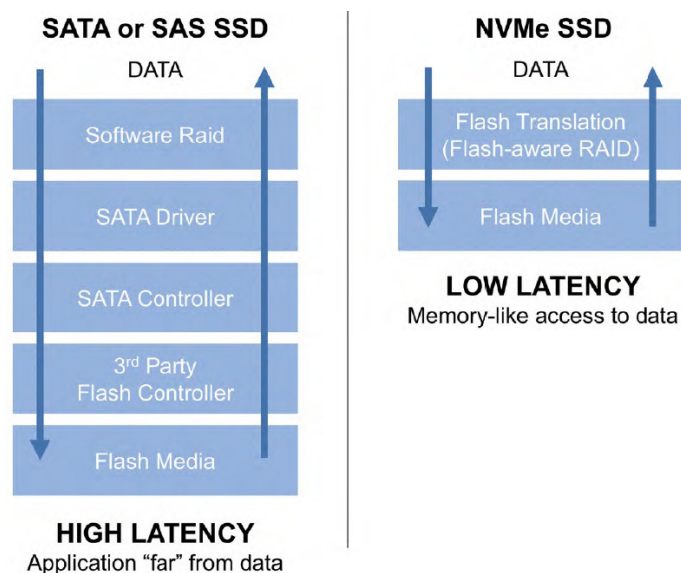


Figure 1: NVMe vs SATA or SAS protocol layers

NVMe is based on PCI Express, which is present in all modern server systems. It is a serial, point-to-point bus with a variable number of data "lanes" in the interface, identified as "x1", "x2", "x4", "x8", or "x16." These lanes are connected either directly to the processor, with no intervening logic, or through a chipset or PCI switch (which can add latency and reduce maximum bandwidth if not properly selected). NVMe, as it is PCI Express-based, is generally only found within a

server chassis, but there are efforts underway to allow NVMe devices outside the server, such as NVMeoF™ (NVMe over Fabric). NVMe was designed from the beginning as an ultra-high-speed connection interface for memory-speed devices, not hard drives. Much of the complexity and layers present in the SAS and SATA protocol stacks are completely eliminated. Even things such as the number of uncached memory accesses (which effectively stalls the processor) are minimized. This technology allows for unprecedented low latencies for interfaces—on the order of 1–2 microseconds—several times faster than storage based protocols, as shown in Figure 1.

Because NVMe is based on PCI Express, it is important to understand the concept of "generation" for PCIe. All PCIe slots are identified as "Gen X," where X is the generation. Most modern servers today provide "Gen 3" interfaces, which provide up to 1GB/s of bandwidth per lane (an "x4 Gen 3" slot can potentially achieve 4 GB/s into an NVMe SSD). The major difference between generations is the bandwidth: it effectively doubles with each generation. This means that an "x4 Gen 3 slot" can provide the same bandwidth as an "x8 Gen 2 slot." This increase has allowed for the proliferation of front-loading U.2 NVMe based SSDs.

High Availability NVMe "2x2" Mode

Some U.2 NVMe SSDs can also support a "dual-port" mode, also known as "2x2." The x4 physical link is broken into two separate logical channels, each at half the width. While 2x2 mode does limit the maximum bandwidth available on the interface to 2GB/s, it provides the same capability that SAS drives provide with their dual ports, namely a redundant link to the storage through a separate controller. This feature is used in highly available NVMe JBOD devices with dual controllers to allow for seamless failover should an NVMe link or controller go down.

Endurance Considerations

Why It's Important: Each SSD warranty allows for a limited amount of written data over its useful lifetime because the underlying flash supports only a finite number of erase and write cycles. Choosing too high of an endurance SSD for a read-mostly application will unnecessarily increase costs, while choosing too low of an endurance SSD for a high-write workload could result in premature application failure.

Hard drives use a magnetic domain to record ones and zeros to effectively provide an unlimited number of writes. By contrast, flash cells (the spot where data is recorded) actually move charges in and out of an insulator to store bits, and they can be written upon only a finite number of times. The physical process of erasing and writing bits into flash, which can be performed from 100 to over 10,000 times, effectively destroys the device. This is why flash drives have different endurance ratings and why flash error correction technologies matter.

SSD Lifetime Calculations

SSD lifetime, commonly known as endurance, is generally specified either in Drive Writes per Day (DW/D or DWPD) or Terabytes Written (TBW). These numbers represent the amount of user data the device is guaranteed to be able to write over the device's lifetime.

Drive Writes per Day is the most common measurement used to express an SSD's endurance specification. This number, which can vary from under 0.1 to over 10, indicates the amount of data that can be written each day for the warranty period.

For example, a 1TB SSD with 3 DW/D and a warranty of 5 years should allow for writes of $1\text{TB} * 3 \text{ DW/D} * (365 * 5) = \sim 5.5\text{PB}$.

$$\text{TBW} = \text{DW/D} * \text{Warranty} * 365 * \text{Capacity}$$

Note: Capacity should be converted to TB if necessary.

Comparing SSDs with different DW/D specifications can be complicated. Two SSDs with the same DW/D specification can have vastly different total amounts of write lifetime, depending on the drive capacity. Sometimes SSDs with lower DW/Ds can actually support writing vastly more data than SSDs with higher DW/D ratings. For example, a 256GB SSD with a relatively high 10 DW/D and a 4-year warranty can write $256\text{GB} * 10 * 365 * 4 = 3.65\text{PB}$, while a 6.4TB SSD with a much lower 1 DW/D and the same 4-year warranty can write nearly three times the amount: $6.4\text{TB} * 1 * 365 * 4 = 9.3\text{PB}$.

SSDs with lifetimes that are specified as "Terabytes Written" have the math completed already, and their lifetimes can generally be compared directly. Simply enough, a drive with a 1000TB Written specification can write two times the amount of data as one specified as 500TB Written.

SSD Endurance Levels

There is a broad mix of SSD endurance levels, sometimes even within a single product line. Many industry leaders refer to the different levels as High Endurance (HE), Medium Endurance (ME), Read-Intensive (RI), or Very Read-Intensive (VRI).

The selection of an SSD's endurance rating will depend on its intended use. If you have a measurement infrastructure, you can monitor your own application to get exact values (but be sure to account for any expected growth). If not, then Table 2 provides general rules of thumb for selecting the right rating:

Error Handling, Power Protection, and End-to-End Data Protection

Why It's Important: A true differentiator between consumer and enterprise SSDs is error case handling. Unexpected power failure, random bit flips in the controller or data path, and other flash errors can all cause data corruption—and there is a wide variance in how effectively, if at all, these conditions are covered.

SSDs are fast, but without using proper enterprise-level data protection you can put your data at risk. The data protection guarantees of enterprise SSDs cover three main areas: NAND error handling, power-failure protection, and end-to-end data path protection. These areas are intended to prevent data loss or erroneous data retrieval at different stages of the SSD processing pipeline.

Error Correction Codes and Signal Processing

The most basic NAND error protection involves the error correcting code (ECC) that is used, and the number of flipped bits it can detect and repair in a certain read area. The strength of this ECC, measured in terms of the number of bits it can correct, directly influences the device's reliability and lifetime. The ECC allows older, "noisier" NAND to continue to provide valid data over the device lifetime. The actual ECC implemented inside the SSD will vary depending on the NAND generation and geometry, but generally the more levels per cell, the higher the ECC requirements. So, TLC usually requires a significantly higher ECC correction capability than MLC.

SSD lifetime and reliability can also be increased by advanced signal processing techniques. These dynamically manage the NAND operation over the SSD lifetime. Some best-in-class controllers modify the NAND cell programming and read algorithms dynamically as the device ages, significantly reducing the raw read-and-write error rate that the error correcting code needs to fix. Western Digital's enterprise-class Ultrastar SSDs implement advanced error correction technology.

Power-Fail Protection

Power-fail protection is crucial for devices storing transactional databases and for other applications that need to ensure that no written data is lost. Due to the block-based nature of flash programming, all SSDs have small RAM caches where data is stored before being written into a previously erased flash block. In typical operation, the SSD returns a signal to the application that it has "completed" the write, when in fact the written data is generally still present only in the RAM cache, and the actual flash write is still underway. Should power to the server be lost before this flash update has completed, the write may never make it into flash. On power restore, the application will attempt to recover any lost transactions from the log but won't be able to find the incomplete flash update, leading to data loss or corruption.

Enterprise SSDs protect against this by including sufficient power storage in the SSD, normally with multiple capacitors on the PCB. These capacitors are charged from the main server. In an unexpected power failure, they guarantee enough power to run the SSD and complete any uncompleted writes left in RAM before they discharge. True enterprise-level SSDs also verify the performance and lifetime of these capacitors when the server powers up. This is similar to how traditional RAID battery backup units periodically perform lifetime tests on their battery backup units (BBUs).

End-to-End Data Path Protection

Finally, end-to-end data path protection helps guarantee that all user data transferred into the SSD is protected against transient errors (random "bit flips"). In the same way that main memory in a server uses ECC memory, and internal processor data paths include parity information, the SSD adds additional check bits for all user data and verifies their state before performing operations. Without this protection, "silent" bit flips can propagate through the controller, eventually resulting in corrupt data being written into flash.

NAND Types

Why It's Important: SSDs today are built on flash cells, with a wide variety of implementations. These range from single-layer, single-bit-per-cell configurations to three-dimensionally stacked groups where each flash cell stores up to 16 different charge levels. Understanding each NAND type's strength and weaknesses helps you choose the appropriate lifetime and reliability SSD for a particular application.

The NAND cell is the most fundamental storage component in an SSD. At a high level, the important metric is the number of bits stored per cell. This measurement dramatically influences endurance and the NAND cell array layout, which can significantly impact density and costs.

NAND Flash Types

A major difference between NAND storage and other storage types is that each NAND element, or cell, can store more than one bit of data by very carefully adjusting the writing and reading algorithm. This has dramatically increased the usable flash bits per SSD while keeping costs reasonable. In single-level cell (SLC) technology, a NAND cell can store a value of 0 or 1 only. This method was used in early SSDs, but due to cost and a narrow performance-and-reliability gap it is not commonly used today. Multi-level cell (MLC) technology uses one of four different charge levels to represent two bits (00, 01, 10, or 11).

This technology essentially doubles the capacity of a single NAND chip as compared to SLC and has been responsible for some of the dramatic reductions in costs seen in SSDs today. A large percentage of enterprise-class SSDs use MLC flash and provide great performance and endurance.

The latest commercial technology is triple-level cell (TLC) NAND, which stores three bits per cell in eight different charge levels (000, 001, 010, 011, 100, 101, 110, or 111). Read performance is often similar, but due to the underlying nature of the technology, the write performance of TLC is generally reduced compared to MLC.

Quad-level cell, or QLC NAND, stores 4 bits of data in a single NAND cell and has also been announced by several vendors, including Western Digital. Because it requires the flash to accurately store 16 distinct charge levels (0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, or 1111) in any cell, this technology will have a very small write lifetime and is expected to be present in flash archival storage only, where updates will be infrequent.

| NAND Type | Bits per Cell | Charge Levels | Write Lifetime | General Use Cases |
|-----------|---------------|---------------|----------------|--|
| SLC | 1 | 2 | Very High | Not commonly found, but in early SSD days it was used in the highest read- and-write performance applications. |
| MLC | 2 | 4 | High-to-Med | General usage with a wide variety of write lifetimes. |
| TLC | 3 | 8 | Med-to-Low | Large consumer adoption with enterprise adoption increasing. Often optimal price-to-endurance ratio for medium performance applications. |
| QLC | 4 | 16 | Very Low | Not generally available yet, but envisioned for write-once-read-many (WORM) type archive application. |

Table 2: NAND Flash Types.

Industry Transition to 3D NAND

There are different ways to build NAND cells, but from a user perspective the biggest difference is 2D vs. 3D NAND. 2D NAND has a single layer of cells built on top of a silicon wafer, similar to the manner in which DRAM or processors are built, and is the original method for building NAND devices. To increase 2D NAND capacity, manufacturers decreased the NAND device size. However, there is a physical limit to how small NAND cells can shrink and still reliably store the required number of electrons needed to read properly. So the industry is evolving from 2-dimensional NAND to 3-dimensional NAND to enable higher capacities.

Manufacturers are transitioning to a 3D process, where individual NAND cells are actually larger than in the latest generations of 2D NAND, to ease semiconductor lithography constraints and ensure enough volume to store charges safely. To make up for this increase in size, NAND charge storage is built vertically, like a skyscraper. As layers are added, the total amount of data a single die can store increases, along with the complexity of stacking and aligning all of these NAND cells.

Benchmarking

Why It's Important: The desire for increased performance is often the first reason architects look at SSDs. Benchmarking real performance, however, is neither simple nor easy. The proper figures of merit must be obtained under a preconditioned, representative workload. This is a necessary step to understanding SSD performance in a specific application.

Performance measurement is often a "black art." There is no one-size-fits-all approach to determining the true application performance of an SSD without actually running that application. But even when the application can be tested on a specific SSD, there must be a method of testing the entire breadth of input possibilities (e.g., a web store's peak load during Black Friday or Singles' Day, or an accounting database's year-end reconciliation). Otherwise, the results might not be completely reliable.

SSD Performance Metrics

Because of this difficulty in testing with real-world application workloads, most SSDs are measured using synthetic benchmarks, under specific conditions. Typical figures of merit for an SSD are Input/Output Operations per Second (IOPS), throughput, loaded and unloaded latency, and quality of service (QoS) and outliers.

IOPS is the number of I/O operations the SSD can perform per second. These operations are generally all the same block size (4KB and 128KB are common sizes, but the size should be specified on a datasheet). The mix of reads and writes should also be specified, as very few workloads are pure read or write tasks, and mixed workloads are often more challenging for SSDs.

Throughput is the amount of data that can be transferred to or from the SSD in a unit of time. It's normally measured in megabytes per second (MBPS) or gigabytes per second (GBPS). The direction of data transfer—whether pure writes, pure reads, or a mixed write-and-read workload—also needs to be specified here.

Latency is the amount of time it takes for an operation to travel from the loaded application to the SSD and return, either with an

acknowledgement or the requested read data. Effectively, it is the round-trip time for a data transfer through the SSD. Latency is normally measured in milliseconds or microseconds, but upcoming persistent memory devices may reduce this to nanoseconds. When specifying latency, it is important to include the number of outstanding I/Os. "Unloaded latency" is the latency of an I/O operation with no other work being performed in the system, while "loaded latency with a queue depth of X" is the latency of an I/O when it has to share SSD resources with X total of I/Os in parallel. The loaded latency will normally increase as the number of outstanding I/Os increase, so it is important to measure this at an expected workload level.

Quality of Service (QoS) measures the consistency of performance over a specified time interval, with a fixed confidence level. There is great variation in this kind of reporting, but generally "macro QoS" reports on the consistency of average IOPS over time, while "micro QoS" plots the latencies of individual I/Os and determines measurements such as exceedance.

Choosing the Right Test Parameters

Simulated workloads are characterized by their block sizes, their access patterns, the queue or I/O depth, and the read-to-write ratio. The block size is simply the natural I/O size for the task at hand. For databases this may be 4KB, 8KB, or 16KB, while for streaming media a larger size of 128K may be used. The access pattern is defined as either sequential (contiguous ranges of the SSD are accessed in sequence), or random (the position of each I/O operation is independent of the prior I/Os). The queue depth, or I/O depth, is an indication of the parallelism of I/O operations, reflecting how many are in flight at any given time.

Most applications can have multiple threads, each reading or writing a different I/O stream, so queue depths of 16 up to 256 are often employed to mimic this configuration. Even single-threaded applications, when run in a virtualized or containerized environment, exhibit high queue depths. This is done by aggregating multiple application streams that use a single queue depth.

The read-to-write ratio (R:W or R/W) indicates the percentage of I/O operations that read pre-existing data vs. writing new or updated data. While many SSD datasheets show 100% read or 100% write performance, in the real world such pure read or write workloads are very rare. Because SSDs can be more easily optimized for these pure workloads, the reported results may be above the level that a real application can achieve, so it is very important to include some mixed workloads in any benchmarking. A more realistic read-to-write ratio of 60:40 or 70:30 can be useful for testing OLTP and OLAP databases, while a 90:10 ratio may make sense for infrequently accessed databases or logs.

SSD Preconditioning Importance

The pre-conditioning state of the drive also needs to be accounted for in any testing. Most SSDs have a very different performance profile "fresh out of the box" (FOB) versus "steady state" (having been completely written by a prolonged series of I/O operations). See Figure 2 for a view of SSD performance, starting with a completely fresh SSD that was subjected to a random 4KB write workload over the course of several hours. While the FOB performance can provide a good indication of the initial performance of the drive, after being deployed for days or months the SSD will generally be in the "steady state" mode with a lower performance level. Because many enterprise

SSDs are expected to be in service for multiple years at very high drive utilizations, steady state is more representative of the application performance than FOB and must be accounted for in any testing.

To place SSDs into steady state performance, they must be pre-conditioned. The best testing methodologies completely fill the drive multiple times (blocks overwritten more than once). Random writes of the selected block size are performed for hours or even days, until the observed performance shows the drop-off to steady state. So for a workload testing 16KB writes, the drive would be repeatedly filled using a high-queue-depth, 16KB, 100%-write pattern. This practice becomes onerous, however, as most testing involves multiple block sizes, and each pre-conditioning stage can take a long time. The best compromise between accuracy and testing time is to simply perform a high-queue-depth, random 4KB workload at the start of any test series.

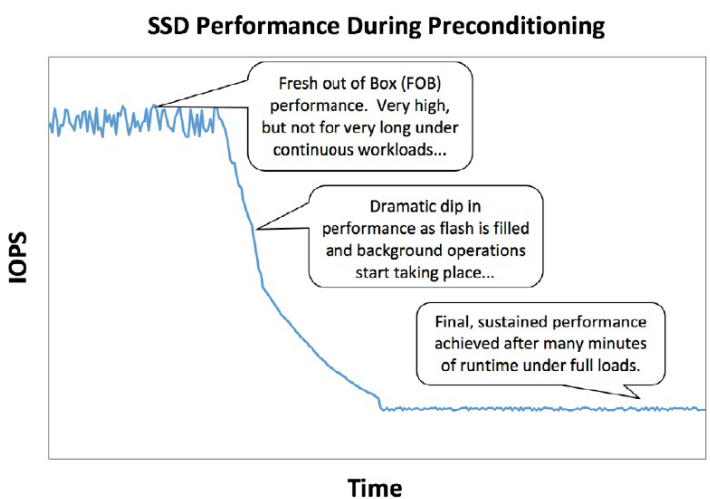


Figure 2: Performance changes from FOB to steady-state

Outliers and Their Effect on Application Responsiveness

One often overlooked SSD performance characteristic is Quality of Service (QoS), or how greatly individual I/O operations may vary in latency. The concern is that "long-tail I/Os"—those that take significantly longer than average—could end up causing your application to break its SLA. For example, you might have an SSD whose average latency for I/O operation is 100 microseconds, but for 5% of the I/Os this latency spikes up to 1,000 microseconds (1 millisecond). Now, suppose you have a NoSQL database that needs to guarantee a 99% lookup time of, say, 500 microseconds per operation. Even though the SSD had a very good average access latency, it could not meet your upper-level application SLA. An SSD with a slightly higher average latency of 150 microseconds, but with only 1% of I/O latencies over 1 millisecond, would be a better choice and would meet your SLA. It is also important to look at the outliers of your specific read-to-write ratio and block size, as the outlier performance of SSDs can vary widely.

Power and Overprovisioning

Why It's Important: SSDs often can be tuned in place to optimize power or performance envelopes. By intelligently utilizing these options, you can realize a significant data center-wide power savings or performance gain.

All enterprise SSDs are specified with a default power envelope, usually within the interface requirements (around 25 watts maximum for NVMe SSDs, around 10–14 watts for SAS, and around 5–10 watts for SATA). For special workloads with hard limits for power consumption, such as when deployed in massive racks constrained by power or cooling, some enterprise-class SSDs can be configured to set a lower power usage limit. In this case, the SSDs power-throttle to reduce maximum energy consumption, often at the cost of performance. If you use a power-throttling configuration in your workload, you must verify the real performance impact by testing the SSDs with the power throttling setting enabled.

Another configuration option that some enterprise SSDs expose for data center deployments is variable over-provisioning. Enterprise SSDs generally ship with a user-accessible space from 1% to 20% less than the total amount of flash on the card, with this extra space used to increase performance or device lifetime. In some SSDs this percentage can be changed; generally, increasing the over-provisioning (reducing usable space) will also increase usable write performance (but not read performance). As this over-provisioning change affects the usable size, it is data-destructive and needs to be completed before the filesystem or application is used. If you use this configuration option, be sure to complete the preconditioning steps and to re-run the entire suite of performance tests to verify the true impact.

Monitoring and Management

Why It's Important: Deploying SSDs is relatively easy; however, as more SSDs are installed, using tools that can monitor health, performance, and utilization from a centralized platform will save time and reduce stress.

Different interface technologies support different monitoring technologies. The basic monitoring technology, available in SATA, SAS, and NVMe interfaces, is called SMART (Self-Monitoring, Analysis, and Reporting Technology). It provides a monitoring tool for checking basic health and performance of a single drive. The SAS interface builds upon this by adding dozens of additional error logging pages, allowing for a finer grained view of drive health. Vendors today are defining a similarly detailed monitoring and logging infrastructure for NVMe drive health. More advanced SSD manufacturers can provide monitoring tools that can integrate and manage a whole data center's SSD portfolio. They may also provide Active Directory/LDAP integration, automated email alerts, and endurance reporting, as well as the ability to perform device-specific and/or enterprise-wide functions like formatting, sanitizing, resizing and updating firmware.

Conclusion

Enterprise SSDs have completely revolutionized the data center. While hard drives still have a bright future in data centers for storing massive amounts of archival and bulk data, the unparalleled latency and bandwidth advantage of SSDs for high-speed databases and other applications is undeniable.

Choosing the right SSD for a specific deployment can be difficult, as there is a spectrum of enterprise-class SSDs that span the gamut of price, performance, form factor, endurance, and capacity. When evaluating enterprise SSDs, look beyond the simple IOPS or bandwidth numbers to match performance to your needs. Consider quality of service to ensure your application SLAs can be met, mixed workload performance to better match real-life workloads, and form factor to assist in hot-swap or fail-in-place architectures.