



Solid-State Drive Technology Brief for Workstation Markets

Contents

- Technology Overview 1
- NAND Characteristics That Differ From HDDs 2
- NAND Cell Technology 3
- SSD read performance. 4
- Mixed SSDs and HDDs 6
- Mapping an SSD to your needs 7



HP 128 GB Solid State Drive, SSD Hard Drive, (AY490AA)

HP recommends Windows® 7.

Technology overview

Flash Memory

Flash memory is a non-volatile computer storage technology that can be electrically erased and reprogrammed. There are two types of flash memory NAND and NOR. NAND technology provides higher density solutions than NOR and enables products such as memory cards, USB flash drives, solid-state drives (SSDs) and PCIe storage solutions. Furthermore, NAND memory reads and writes blocks of data while NOR memory provides random access and direct byte execution. This paper will focus on NAND memory and its application in solid-state devices.

Solid-State Drives versus Hard Disk Drives (HDD)

Solid state drives based on NAND memory arrays have several advantages over traditional hard disk drives based on rotating magnetic media.

FIGURE 1

Power	SSDs consume significantly less power than HDDs. While comparison of specific drives will vary the power savings can be significant. Note that the power savings varies due in part to the HDDs operating state.
Random access	The random access time is much quicker for SSDs. HDDs have to move the head and rotate the disks in order to access randomly placed data. NAND based SSDs has some overhead for random accesses but it is much shorter than rotating media.
Read Performance	An SSD's read performance is better than the HDDs. While the actual improvement over HDDs vary based on data access patterns and a host of other variables highlighted below it is not uncommon to see a significant performance improvements.
Spin-up	SSDs having no moving parts have an almost instantaneous spin-up time while HDDs can take several seconds. Note that SSDs may need to power up from a sleep state which will take a few hundred milliseconds.

NAND Characteristics that Differ from HDDs

An understanding of NAND technology and architecture is required to get at the root causes for the characteristics discussed below. This paper will focus on the characteristics that affect user's perception and use of NAND technology and not an in-depth discussion of NAND architecture.



Block Erasure

NAND memory is arranged in blocks for quicker access and erasure. Blocks can vary in size but are larger than OS pages and standard HDD 512 byte sectors. For any flash cell to be programmed (written), the cell must first be erased. Flash memory can only be erased a block at a time. Erasure consists of setting all bits in the block to 1. Once erased any bit can be programmed, but once the bit is set to zero it can only be erased by setting the entire block to 1.

Garbage Collection and TRIM

Data is typically written to SSDs in pages but must be erased in larger units called blocks. Garbage collection in an SSD is the act of moving and coalescing the valid pages into new blocks so that partially filled blocks can be erased and made available for new data. By its nature garbage collection requires a read and write of data. TRIM is a command supported in Windows7 and in select vendor's tools that will can tell SSDs that all pages within a block are no longer in use and can be erased. TRIM provides the sector and page information to the SSD and reduces the garbage collection and most importantly reduces the number of writes required.

Memory Wear

Flash memory has a finite number of program-erase cycles or P/E cycles. A SSDs firmware and hardware is designed to dynamically remap blocks in order to spread write operations out, prolonging the life of the device. This technique is called wear leveling.

Read disturb

If a NAND memory cell is read over and over again hundreds of thousands of times without a rewrite of surrounding cells it is possible to change the value of the surrounding cells. Controllers track total reads across the NAND device and rewrite the surrounding cells to mitigate read disturb errors.

Overprovisioning

SSD manufacturers reserve a certain amount of NAND memory, keeping it hidden from the OS so that it has space to temporarily store data during garbage collection and wear leveling activities. This allows even reasonably full SSD drives to maintain a reasonable performance level and in practice has the benefit of improving reliability and extending the usable life of the SSD. The amount of over provisioning varies by device. Mainstream/client devices are in the range of 7% to 10% while devices focused on the enterprise/server markets are greater than 20%.

Write Endurance

Flash devices support a limited number of writes and will eventually fail if the cell experiences too many writes. Wear leveling algorithms help to extend the life of a flash device by spreading writes uniformly across the device.

One complication as NAND geometries shrink is that fewer electrons per cell are available to reliably control the NAND cell state, which leads to reduced raw write endurance at the NAND cell level. This requires more advanced management and Error Correcting Codes to offset the increased bit error rates.

To simplify matters, manufacturers will typically quote write endurance in Terabytes Written (TBW) based on the underlying technology of the NAND device taking into account the wear leveling, ECC, and management algorithm based on typical access patterns.

Write Amplification

This is the ratio of writes to the NAND device divided by the total number of writes required to the NAND array. This ratio provides a metric for quantifying the overhead of the management algorithms mentioned above. The workload including the data access pattern and occupancy rate of data on the drive will affect the write amplification.



IOPS and Bandwidth

A common performance measure for HDDs and SSDs are Input/Output operations Per Second (IOPS). IOPS are typically measured for four types of accesses: random reads, random writes, sequential reads, and sequential writes. For NAND devices, IOPS will start high for a new or erased device and drop to a steady state value that accounts for wear leveling and garbage collection. SSDs even in steady state have higher IOPS than HDDs. Bandwidth numbers are related to IOPS and can be derived if the transfer size of each I/O operation is known.

Data Retention

SSD vendors try to meet the requirements of many JEDEC Standards. JESD218A specifies that SSDs must meet the following retention specification when the device has reached 100% of its write endurance limit or wear limit:

FIGURE 2

Application Class /Workload (see JESD219)	Active Use (power on)	Retention Use (power off)
Client	40°C, 8 hrs/day	30°C, 1 year
Enterprise	55°C, 24 hrs/day	40°C, 3 months

The definition of the application classes above is an attempt to recognize that there are different levels of demand for SSDs based on different application workloads. JESD219 clearly specifies an enterprise workload that consists of random data distributed in a predefined manner across the entire SSD. JESD219 has yet to specify a client workload. It does assume that client workloads are not as intensive as enterprise workloads and that the SSD will not have the same level of data distributed across the entire device.

Temperature and the amount of wear on the device also affects the data retention time. The higher the active use temperature the longer the data will be retained when the device is powered off. Unfortunately, the higher the temperatures during power off the shorter the retention time.

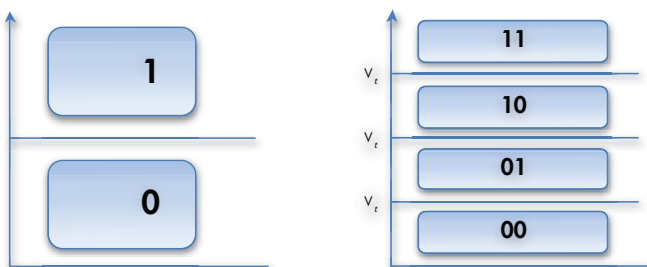
As mentioned previously the one year retention time in the table above assumes that the NAND device has reached 100% of its endurance rating, essentially in a state where it should be replaced. JESD47H.01 states that a device at 10% of its endurance limit should provide a data retention time that is 10 times longer than when it is worn out. This last specification implies that a device at 10% of its endurance rating will have a data retention time of 10 years, assuming the power on temperatures of 40°C and power off temperatures of 30°C.

To summarize JEDEC has provided guidelines for client workloads that specify a data retention time of 10 years for new devices. This data retention time decreases as the device wears until it reaches one year for devices at their endurance limit. The data retention times will vary based on active temperature, storage (power off) temperature, and on the workload used as the device is consumed.

NAND Cell Technology

NAND memory cells are based on either a single-layer (SLC) or multi-layer (MLC) design. An SLC cell can hold only one data bit. An MLC cell can hold more than one bit of data per cell.

FIGURE 3



Because MLC designs can hold more than one bit of data per cell the density of MLC devices is higher. Likewise the cost per bit of MLC devices is lower. Due to the higher complexity of MLC cells, the endurance and write/erase speeds are often lower while data integrity management costs are higher.



The table below summarizes these differences:

FIGURE 4

	Single-Layer Cell	Multi-Layer Cell
Bit density	Lower	Higher
Capacity	Lower	Higher
Cost / bit	Higher	Lower
Write/erase programming	Fastest	Slower
Write Endurance	Higher	Lower
Reliability	Higher	Lower
Data integrity management	Lower complexity	Higher complexity

Workstation Categorization of SSDs

The Industry categorizes SSDs in the following manner:

FIGURE 5



The categorizations above are rough approximations. To complicate matters even SSDs within the same product family can vary greatly. For example, Intel® Solid-State Drive 320 Series a MLC device from the enthusiast class has specifications that change over the product family. The specification differences are often due to differences in the number of NAND chips that can be written in parallel as the capacity of NAND memory changes across the product family.

FIGURE 6

	Sustained Sequential Reads (up to)	Sustained Sequential Writes (up to)	Random 4KB Reads (up to)	Random 4KB Writes (up to)	Write Endurance
40 GB	200 MB/S	45 MB/S	30,000 IOPS	3,700 IOPS	5 TB
80 GB	270 MB/S	90 MB/S	38,000 IOPS	10,000 IOPS	10 TB
120 GB	270 MB/S	130 MB/S	38,000 IOPS	14,000 IOPS	15 TB
160 GB	270 MB/S	165 MB/S	39,500 IOPS	21,000 IOPS	15 TB
300 GB	270 MB/S	205 MB/S	39,500 IOPS	23,000 IOPS	30 TB
600 GB	270 MB/S	220 MB/S	39,500 IOPS	23,000 IOPS	60 TB



In order to focus customers on the appropriate device for their specific environment HP workstations has partitioned SSDs according to workloads and defines the following classes:

FIGURE 7

Mainstream	Extreme	Enterprise
Average write load up to 25 GBytes/day over 3 years	Average write load up to 110 GBytes/day over 3 years	Average write load > 110 GBytes/day over 3 years
Mainstream Use Cases	Extream Use Cases	Enterprise Use Cases
Power Office User	Streaming data acquisition	Heavy demand imaging
Administrator	Heavy digital media	Large data capture
Entry CAD	Medium demand imaging	Heavy SSD caching
Boot Device (Fast Boot)	SSD caching	
Other light write load application environments		

Note that the use cases are examples only and customers should monitor their data usage in order to determine the correct categorization.

SSD Read Performance

SSDs fundamental value to improved performance is due to the ability of the SSD to quickly retrieve data from the NAND memory. There are many reviews comparing HDD and SSD performance. These reviews show that performance varies based on the architectures of the HDD and SSD, size of the blocks, the randomness of the accesses, and the queue depth. All reviews invariably show that SSDs provide some read performance improvements. The graph below, which highlights the windows experience index and its components, is typical of the read advantage seen with SSDs.

FIGURE 8

Windows Experience Index

HP Z210 Workstation, Genuine Windows® 7 Professional 64-bit, E3-1290, 4G DRAM





This read performance improvement provides system and application performance improvements in several different ways. Some examples numbers using a HP Z800 with Genuine Windows® 7 Professional 64-bit, two X5560 2.8GHz processors and 12 GB of DRAM are shown in the table below.

FIGURE 9

	SSD— 160GB MLC SATA (seconds)	HDD— 160GB 7.2K SATA (seconds)	Improvement
Boot	13	22	1.6X
Application installation from drive	10	24	2.4X
Data load	6	13	2.2X
Application startup	2.5	6	2.4X
Shutdown	4	6.5	1.6X

All of the numbers above are shown as an example. Performance improvement will vary dependent on system configuration, load on the system during boot, application and data characteristics and a user’s workflow.

Mixed SSDs and HDDs

There are several ways that SSDs and HDDs can be mixed to provide a performance advantage. A few of these options will be discussed in this section.

A small SSD can be used to hold the OS, recovery partition, and application image to provide quick boot and application start. An HDD is added to hold all the user specific data. This model dramatically increases the lifetime of the SSD since the OS and application start is mostly a read operation. Page files and other data which may be cached on the OS/application partition could be stored on the HDD to improve the lifetime even further.

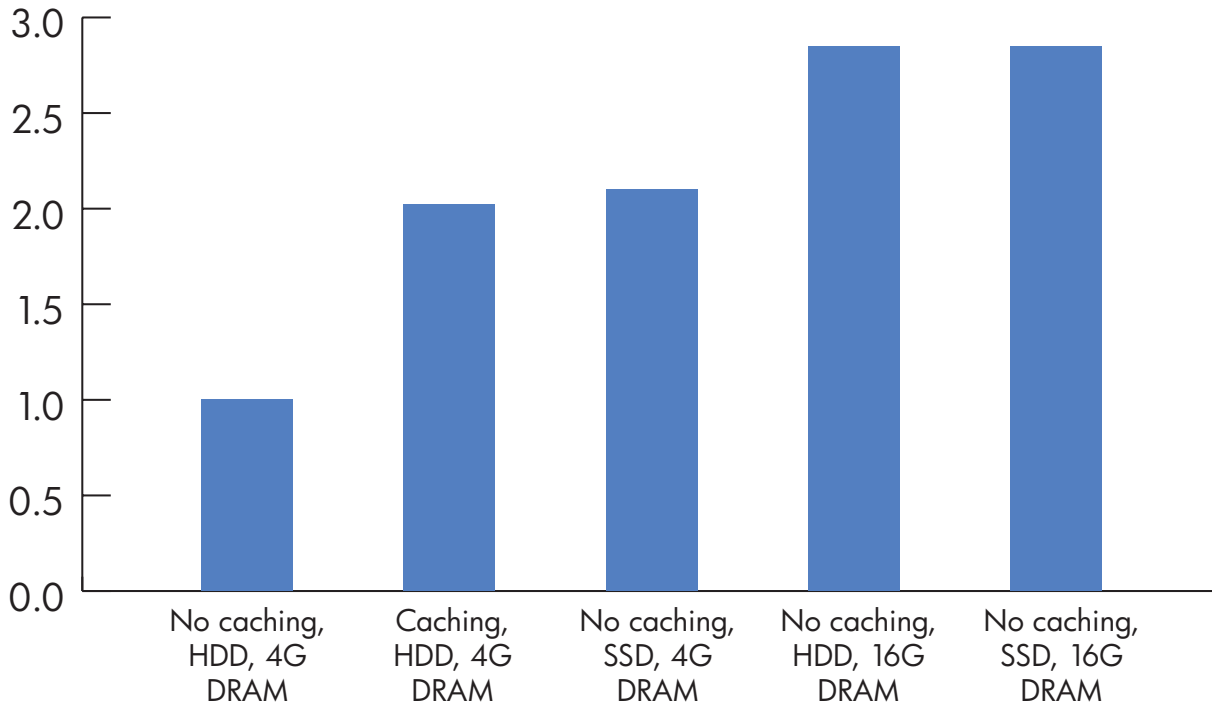
Many workstation customers do not reboot or restart their workstation platforms on a regular basis. Their concerns are not with OS loading or application start, but with data access. SSDs can be used for long-term data storage to allow for quick data access or it can be used for temporary data storage. Many applications require scratch files or data caches in order to operate efficiently. Using a NAND based solution such as an SSD can greatly improve performance for these applications.

A technique called flash caching is also available. Flash caching is a software technique of monitoring HDD access and storing blocks that are used most frequently onto an SSD. These techniques offered by vendors such as Intel®, NVELO®, Diskeeper®, and others have the potential of providing performance improvements, but are dependent on vendor specific algorithms. Installing one of these applications can improve OS boot, application startup, and even the data load. Application runtime improvements vary and are often absent depending on system configuration and workloads. Specifically, if you have enough DRAMS to hold the entire dataset during an application’s runtime, performance improvements will be minimal since the dataset can reside in memory. An example of this can be seen using SPECcapc for 3ds Max™ 2011.



FIGURE 10

SPECapc 3ds Max 2011 Large Model CPU



Mapping an SSD to Your Needs

Since SSDs are more costly than hard drives of similar capacity, the usage of SSDs should be carefully considered. Below are a few questions that should be answered before purchasing a SSD.

Is managing a mixed SSD/HDD environment acceptable?

- Multi-partition systems require more thought and care than a single partition system.
- OS, application placement, and data placement has to be considered to provide the proper users experience.

Is the SSD to be used to improve OS load time or boot time?

- Size of the OS image when deployed
- Size of the recovery partition, if any

Is application startup important?

- Size of application footprints when installed.

Is data load time important?

- How much data is required for the application loads that are of concern?
- How easily can I train users to store data on the SSD assuming a mixed SSD / HDD environment?



How much write endurance do I need?

- How much data is written per day on average?
- How many days will the SSD be in use before replacement?
- Is the workload significantly different than what the vendor has used to qualify the device?
 - Accurate calculations require an understanding of:
 - How data is aligned when written to the NAND array
 - The profile of sequential versus random data
 - Underlying understanding of the management algorithms
 - The information above is hard to determine and using the simpler GB/day measure along with assumption that the workload is similar to the client workload will provide a reasonable starting point for determining which SSD class to choose from.
 - SSDs should be monitored to determine if the wear is meeting expectations
- Based on the gigabytes written per day determine which workstation SSD class is appropriate (Mainstream, Extreme, or Enterprise).

The last question above is often difficult to answer and most SSD vendors recommend trying a SSD drive, determining how it is wearing (with vendor specific tools), and then choosing an appropriate SSD solution. This approach should be considered.

With the questions above answered and knowledge of the HP workstation platform to be used a mapping to an appropriate workstation qualified SSD is possible. Refer to the HP website or your local HP representative for the SSD that fits your need.