IDC
Analyze the Future

WHITE PAPER

Storage Device Reliability and Endurance

Sponsored by: Seagate

Jeff Janukowicz John Rydning

August 2010

IN THIS WHITE PAPER

In this white paper, sponsored by Seagate Technology, IDC discusses storage device reliability and endurance while examining the importance for storage device vendors of defining a common set of metrics to characterize solid state storage. This white paper explores the different reliability aspects of solid state drives (SSDs) and discusses the importance of endurance in the enterprise storage market.

SITUATION OVERVIEW

The hard disk drive (HDD) is a technology over 50 years old, and while many other technologies have come and gone, the HDD remains the predominant storage device for data today. Decades' worth of research and development and billions of dollars have been spent to advance disk drive technology. Through the efforts of many in the HDD industry, HDD storage densities, as measured by the capability to store a given number of bytes within a single drive, have increased to 3 terabytes of capacity within the industry today. The cost of storage has also benefited from the decades of advancements and is measured in cents per gigabyte of capacity.

The dramatic advancements in HDD technology that have resulted in higher-capacity devices at more affordable prices have been a key factor in the success of the HDD. However, the fact that the industry ships over 500 million disk drives annually is a testament to the quality and reliability of the HDD as well. Today, we rely on and trust HDDs to store the data that is critical in our everyday lives — from important PowerPoint presentations on our work laptop, to the family photos on our home computer, to the 24 x 7 operation of HDDs that run in datacenters.

Meanwhile, recent advancements in solid state technology (specifically NAND semiconductor technology) have thrust SSDs into mainstream applications and datacenters. Increasingly, SSDs are becoming a more common storage device in the IT environment. Yet, when compared with the decades of history and experience behind HDDs, the SSD is a relative newcomer in many IT applications. Server and storage system manufacturers are tasked with evaluating various integration strategies for SSD devices within enterprise systems. One integration strategy is simply to leverage NAND-based SSDs in the same way traditional HDDs are leveraged (i.e., replacing an HDD with an SSD). However, a NAND-based SSD does not function in the same manner as an HDD. In fact, the construction of an SSD, with its use of NAND flash components as the storage media, is inherently different from the construction of an HDD.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

The performance aspect of NAND-based SSDs is one of the more obvious distinctions between the two technologies. SSDs have no mechanical armature to move and thus provide random data read access times that are significantly faster than those of HDDs. Yet, one of the other dissimilarities and possibly the most misunderstood difference between the two storage technologies is revealed when looking at device-level reliability and endurance, or the life span of the storage device a user can expect when reading and writing data.

From this perspective, SSD reliability and endurance are impacted by a number of factors that are different from those that impact HDDs. Customers and system designers must understand the parameters that influence device longevity and reliability of the overall SSD, such as NAND component endurance, the efficiency between the data written to the NAND media and the data received from the host (or write multiplication/reduction factors), and even how efficiently the data is spread over all cells in the NAND media (or wear leveling), among other factors.

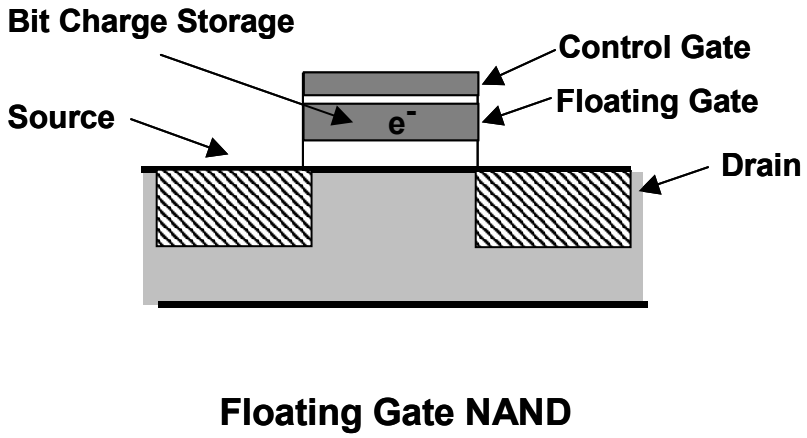
To lay the groundwork for understanding SSD component- and device-level reliability and endurance, we take a closer look at the underlying NAND storage media and its intrinsic properties that impact endurance and reliability. In addition, we examine SSD architectures and a few of the techniques used to manage data written to and read from NAND in a typical SSD. In the end, this paper details the importance for storage device vendors of defining a common set of metrics to characterize solid state storage.

Overview of NAND

NAND flash memory is a nonvolatile form of memory, which means that a source of power is not needed to maintain the data stored within the device. NAND flash is a solid state semiconductor device that can be electrically programmed and erased, which makes NAND an effective device to store data. The basic NAND memory structure consists of a substrate with a source and drain and a floating-gate transistor, as shown in Figure 1. The floating-gate region in a NAND cell is designed to store a charge (electrons) that is used to "program" the cell for data storage.

FIGURE 1

NAND Flash Cell



Source: IDC, 2010

NAND Read/Write Process

In writing (or programming) a NAND cell, current carriers are generated in the substrate between the source and drain, and a voltage is applied to the control gate that causes some electrons to tunnel through the gate oxide insulator into the floating gate, where they are effectively trapped. The presence of electrons trapped within this floating gate subsequently changes the threshold voltage of the transistor, resulting in the transistor being turned on during the read (i.e., a "1" is written). In erasing, electrons are expelled from the floating gate by reversing the respective polarity of the control gate. This process is one of the fundamental concepts of NAND flash and a vital aspect in understanding the reliability and endurance of the technology as NAND-based SSDs must perform this process to read or write data that is stored and ultimately ensure data integrity.

NAND Page-Based Program/Erase Cycles

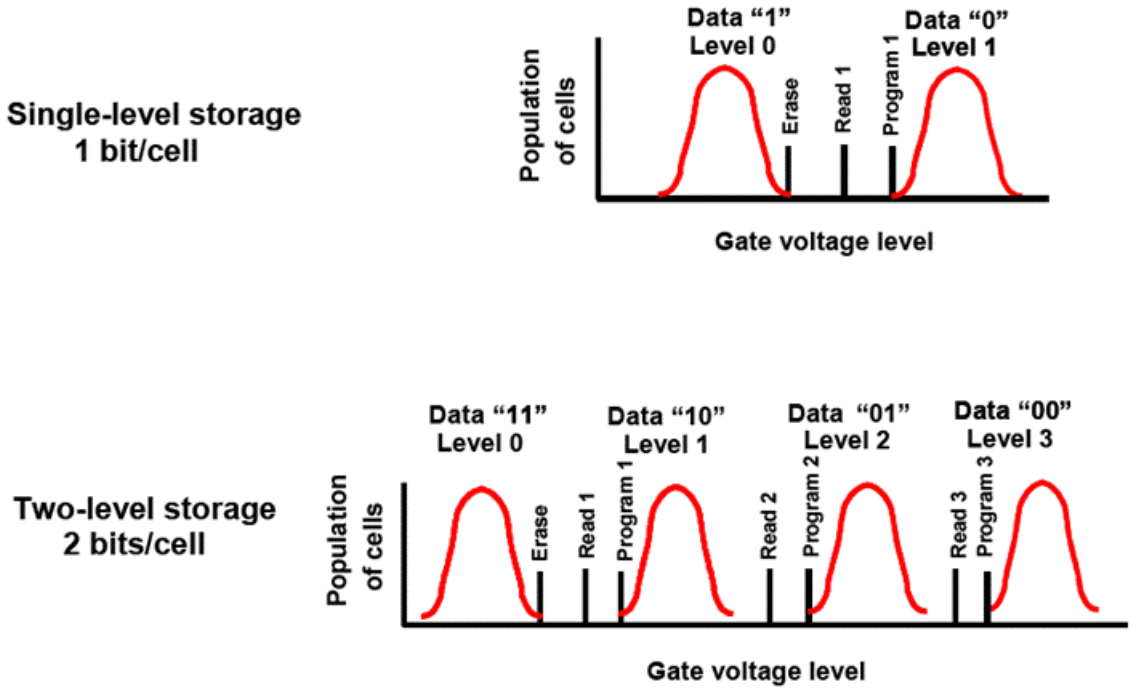
Another important factor in NAND flash memory architecture is that individual bit cells are organized in pages and blocks. As such, individual bits are not accessible. Instead, entire blocks or pages are addressed within a NAND chip. Today's flash memory is read or programmed on a page basis, typically in sizes of 2K–8K bytes. Yet, it must be erased at a block level (typically 16K–512K bytes). In this organization, to write data, an entire block must be erased prior to writing (assuming it has already been written to, or programmed). This is called a program/erase (P/E) cycle, which is another aspect that impacts the endurance of the technology as it introduces significantly more wear and tear on the NAND media because data must first be erased before data is written and a fundamental mismatch exists between erase block size, page size, and the data.

Single-Level Cell and Multilevel Cell Technology

Another aspect of NAND is its ability to store a varied number of electrical charges in the floating gate such that the cell can be programmed by varying the threshold voltage to have one of multiple possible voltage levels in each cell. Each programmed threshold voltage can be assigned to a specific logic state (see Figure 2).

FIGURE 2

NAND Technology: SLC Versus MLC



Source: IDC, 2010

As a result of this technique, current NAND flash that is used to construct SSDs is based on two types of technologies: single-level cell (SLC) and multilevel cell (MLC). Today's MLC NAND effectively stores 2 bits at four different voltage levels in a given cell, as opposed to SLC, which stores only 1 bit at two distinct voltage levels. The natural result is twice the capacity (today) for the devices when using MLC technology compared with SLC technology.

However, because both SLC and MLC NAND use a voltage threshold window with a similar size, the tolerance between adjacent voltage levels in MLC is much smaller than in SLC NAND. As a result, achieving this precise distribution of charge on the floating gate of the MLC flash device requires a more sophisticated and time-consuming programming algorithm. This impacts data reliability in terms of a higher probability of error that can affect data integrity and endurance by impacting total

number of program cycles of the NAND media and making the MLC NAND wear out faster. Nevertheless, this is an evolving market, and new techniques and technologies are under development to both optimize performance and increase the bit density.

NAND Flash "Wear-Out": The Impact on Endurance and Reliability

Fundamental to the design of NAND flash is the potential for irreparable damage to the floating gate due to multiple program/erase cycles. Simply put, the endurance (which means the number of cycles in which a block can be erased and programmed) is limited. The relatively strong electric fields used during the program/erase cycle are capable of damaging the floating gate, which, if damaged, permanently alters the NAND cell characteristics. The potential for this problem is exacerbated when the SSD has a limited number of NAND blocks or a fixed amount of capacity available to use. Thus, multiple program/erase cycles based on the amount of data written to the device (or workload), the efficiency with which the program cycles are spread over all cells in a flash device evenly (or wear leveling), or the efficiency between the data written to the NAND media and the data received from the host (or write multiplication) can cause the NAND cells to "wear out" prematurely and negatively impact the endurance of the overall SSD device and reliability/accessibility of the data contained therein.

Because additional program cycles are required to operate MLC NAND and its tighter voltage threshold window, an MLC NAND cell inherently will wear out faster than an SLC NAND cell because the signal to noise of the NAND media degrades over time. It is important to recognize the difference between these attributes of SLC and MLC flash because it affects the endurance specified for a given block:

- SLC NAND generally is specified at 100,000 write/erase cycles per block.
- MLC NAND typically is specified at 10,000 write/erase cycles per block.

Additionally, data retention (or the integrity of the stored data on a flash cell over time) is impacted by the state of the floating gate in a NAND cell where voltage levels are critical. Leakage to or from the floating gate, which tends to slowly change the cell's voltage level from its initial level to a different level after cell programming or erasing, may change the voltage level. This altered level may be interpreted incorrectly as a different logical value by the system. Thus, due to the tighter voltage tolerances between MLC levels than SLC levels, MLC flash cells are more likely to be affected by leakage effects. Consequently, care must be taken to ensure the long-term data retention capabilities of both SLC and MLC NAND when used in enterprise storage.

In response to these issues, NAND flash OEMs recently have announced technology (called Enterprise MLC, or eMLC) that dramatically extends the life span of flash-based storage for enterprise applications. The new technology delivers a higher program/erase cycling capability than any other NAND technology, and demonstrations of devices capable of achieving 1 million write/erase cycles have been shown.

Enterprise-Class SSDs

As mentioned earlier, SSDs are fundamentally different from HDDs in that they are built using solid state semiconductor technology and thus use no mechanical moving parts to store the data. For example, the Seagate Pulsar SSD — the first product in the new Pulsar family — leverages nonvolatile flash memory rather than spinning magnetic media to store data (see Figure 3).

FIGURE 3

Enterprise SSD

Source: Seagate, 2010

It is important to understand not only the characteristics of an individual NAND cell but also the underlying NAND flash components that are used to construct an SSD. SSDs work by storing data on semiconductor memory, typically NAND flash memory as the storage media. To accomplish data storage in a manner similar to an HDD, SSDs utilize an architecture that contains a number of NAND flash integrated circuits (ICs) arranged in a parallel manner on a printed circuit board (PCB) to store the data. Another critical component — the controller IC — is used to store this data efficiently. The controller acts as the brain of the SSD and manages all the interactions between the different SSD components. The SSD controller typically contains a microprocessor, buffer memory, and an interface to the host device.

Due to the nature of NAND's operation discussed previously, data cannot be directly overwritten as it can in an HDD. When data is written to an SSD initially, the cells all start in an erased state so that data can be written directly using pages at a time, with no erase cycle necessary — an important dynamic to understand. The SSD controller has a Flash translation layer (FTL) that manages the interactions to the NAND memory and uses intelligent algorithms to maintain a logical to physical mapping system known as logical block addressing (LBA). When new data is sent to the SSD

to update previously written data, the SSD controller will write the new data in a new location and update the logical mapping to point to the new physical location. The old location is no longer holding valid data, but it will eventually need to be erased prior to being written again. Consequently, care must be taken in managing the data flow intelligently to ensure long-term reliability when used in enterprise storage.

Techniques Used to Improve Enterprise NAND-Based SSD Reliability and Endurance

On the surface, many of the issues associated with NAND as a storage media may appear too overwhelming or challenging for the technology to be used in the enterprise environment. However, enterprise SSDs integrate a number of advanced techniques and intelligence to help overcome the endurance and reliability limitations at the NAND flash media level, such as the following:

- **Error correction code (ECC).** ECC is used to detect and correct errors by adding additional bits to the data. ECC algorithms, such as Reed-Solomon codes, Hamming codes, Bose-Chaudhuri-Hocquenghem (BCH) codes, and others, typically are used in storage applications. In general, the more bits of ECC that are used, the higher the level of error correction. Hence, an SSD with effective ECC will be able to correct more errors, which ultimately improves the time to wear-out.

- **Wear-leveling techniques.** Wear leveling is a process that SSDs utilize to minimize the impact of the NAND endurance limitation by spreading the program cycles over all cells in a flash device evenly. Two primary techniques, static and dynamic, commonly are used in SSDs to manage access to the NAND media. Static wear leveling prevents infrequently accessed data from remaining stored on any given block for a long period of time. Static wear leveling is designed to distribute data evenly over an entire system by searching for the least used physical blocks and then writing the data to those locations. Dynamic wear leveling distributes the data over free or unused blocks. In the end, the combination of these wear-leveling techniques increases the life span of an SSD by spreading the data across all the cells in the device evenly to avoid individual cell wear-out.

- **Use of spare blocks (or overhead).** Providing spare blocks of additional NAND capacity is another way to improve endurance. For example, an SSD marketed as a 25GB SSD may show 25GB of capacity available to the user to store data. Yet, the SSD may be constructed with 32GB of actual NAND capacity. The 7GB of overhead (or spare blocks) in this example can be used to improve wear-leveling efficiency and other program/erase operations to increase the endurance and performance at the device level. This is commonly referred to as overprovisioning.

- **Buffering the data.** In an SSD, and also with an HDD, buffering the data with a small amount of DRAM memory can improve performance. Yet, in an SSD, buffering the data also improves device-level endurance by optimizing writes, limiting the program/erase cycles, and eliminating any mismatch between erase block size and the data size.

The Need for Device-Level Specifications

At the beginning of this decade, enterprise storage system OEMs began using desktop-class HDDs (designed primarily for use in a desktop PC) in storage systems as a substitute for enterprise-class 10,000rpm and 15,000rpm HDDs. The idea was to use a high-capacity HDD that carried with it a lower cost per GB compared with an enterprise-class HDD. However, in the early years of implementation, as system OEMs attempted to leverage desktop-class HDDs in large storage arrays and 2.5in. mobile-class HDDs on server blades, IDC uncovered drive reliability issues based on numerous discussions with system OEMs. We believe the demands of multidrive environments, high workloads, and 24 x 7 power on hours (POH) were out of the design envelope of this class of drives and necessitated a significant round of customer education and device redesigns.

The HDD industry responded by integrating capacity-optimized HDDs with advanced technology that allowed the drives to be used in enterprise system applications and to withstand the more demanding workloads and power on hours of operation. HDD vendors, such as Seagate, proceeded to communicate clearly the workloads that capacity-optimized HDDs were designed to support and the endurance or operating life that could be expected without failure when integrating this class of HDD in an enterprise application.

This experience provides a good lesson for SSD OEMs that are new to the enterprise market and that seek to deliver SSDs for enterprise applications: The SSD device must be designed for enterprise workloads, and there must be a standard metric that indicates the expected endurance of the device such that system OEMs can architect the entire system sufficiently.

Terminology and metrics that clearly convey the reliability, endurance, and expected life of an SSD device in enterprise applications are vital in order to avoid confusion in the marketplace and advance adoption by users. Some vendors within the industry have suggested that the number of SSD cycles is equal to the expected total data written to the SSD over its life divided by the total capacity of the device. However, this definition/methodology is too simplistic. The reality is that many factors come into play to determine device-level endurance. Use of overly simplistic measures for SSD endurance can manifest later as problems in "real-world" environments.

Currently, the storage industry does not have a good, consistent way to describe SSD device-level endurance and reliability. JEDEC (JESD22-A117A, JESD47) specifies the relationship between endurance and data retention at the NAND component level. JEDEC's JC64.8 subcommittee, of which Seagate is a founding member and active leader, focuses solely on issues relevant to SSD device-level metrics. The JEDEC 64.8 set of standards clearly communicates device longevity in a variety of usage scenarios (or workloads) that allow customers and system designers to compare SSDs and HDDs for specific applications so that the appropriate storage medium can be integrated. This strategy should provide storage device vendors and system OEMs with an effective tool to communicate device endurance and accurately make decisions around device usage, warranties, and qualifications for the most proficient use of the device over its life cycle.

FUTURE OUTLOOK

Customers expect that data stored on an SSD or an HDD will always be there and that it will be accurate regardless of the conditions: loss of power, temperature fluctuations, vibrations, and shock. They also expect storage to be relatively low cost. The use of NAND-based SSDs is a relatively new phenomenon in the critical data-sensitive, enterprise storage market. As end users seek higher-performing systems, SSD adoption will grow, and as the installed base grows and the market matures, the reliability of SSDs will be exposed — for good or for bad. Hence, it is most beneficial for device and system OEMs (for the sake of end users) to define a common set of metrics that characterize solid state storage reliability consistently and appropriately and to establish these definitions sooner rather than later. The development of a set of standards around reliability will allow customers and system designers to evaluate SSDs for their use in their given application, to set expectations appropriately, and to increase customer satisfaction.