Western Digital.

White Paper

SSD Endurance and HDD Workloads

A comparison of data transfer ratings for HDDs and SSDs

January 2023

Introduction

Solid-state drives (SSDs) and hard disk drives (HDDs) are complementary storage devices in certain applications and are competitive storage devices in other applications. In the latter, it is common to look at SSD and HDD specifications and to try to compare them side-to-side.

SSDs are subject to a very well-known limitation, based on the write endurance specification of NAND flash, and are specified for the total quantity of data written from the host to the drive. HDDs are subject to less well-known workload ratings, usually specified as the total quantity of data transferred between the host and the drive, whether read or write. Both are specified in total data transfer over a period of time.

Given the similarity between these two metrics, it is easy to see how they may be compared directly. However, an SSD endurance specification and an HDD workload rating serve two very different purposes and are not directly comparable between technologies.

SSD Endurance

Physical Structure

NAND flash memory at its core is comprised of many individual cells as shown in Figure 1. The cell is effectively a transistor with a storage element built in. A transistor works as a relay, where the charge applied to the gate can turn on or turn off current between the source and the drain. A flash cell is similar, but contains an additional storage element, such that either the gate voltage or the stored charge—or a combination of the two—can turn on or off current between the source and the drain. There are two common types of flash cells in the market, the floating gate (FG) or the charge trap cell (CTC). In addition, there are many physical differences between the construction of 2D and 3D NAND. While there are many very important differences between these, from the standpoint of operation and endurance, they can be treated as similar.

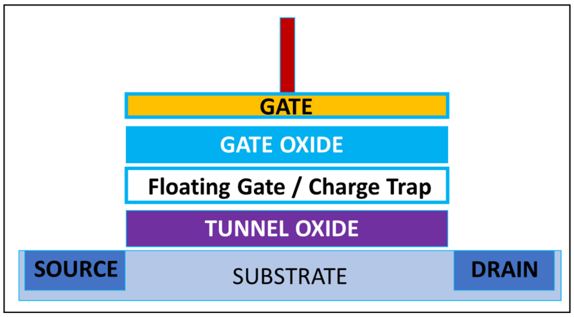

Figure 1: NAND cell structure

A flash cell is built to retain charge. As such, there needs to be isolation between the storage element and the rest of the electronics. This isolation, the insulating oxide layers shown in Figure 1, are what allows the cell to remain charged in the absence of power making this a non-volatile memory (NVM). Read operations are only used to detect the charge in the cell, not add or remove charge, and do not meaningfully affect the isolation elements. The isolation elements, however, present a significant problem when it comes to programming or erasing the cell. To program or erase the cell requires using a large enough voltage to overcome the isolation and force charge into (program) or to force charge out of (erase) the cell.

This process of repeated program-erase (PE) cycles degrades the oxide layers over time. As the oxide layers degrade they can no longer effectively work as insulators to retain charge within the cell. As charge dissipation increases, particularly apparent at higher temperatures, the cell becomes unable to store charge in a state that can be reliably read over time. This behavior is inherent to NAND flash technology and the reason NAND-based storage devices specify a write endurance rating.

Endurance and Data Retention

As described above, it is the PE cycles that degrade the oxide layers and cause the NAND to become unable to store charge over long periods. Endurance (the number of PE cycles that a NAND cell can undergo) and data retention (the amount of time that NAND must be able to store data reliably) are two sides of the same coin.

There is an inverse relationship between PE cycles and data retention - as PE cycles increase, data retention diminishes. This endurance / data retention dichotomy is integral to the way that NAND flash is specified for use. The relationship is defined by JEDEC specification JESD218 for client and enterprise SSDs, shown in Table 1 below.

The data retention is only specified for “power off” data retention. Most modern SSDs will have power-on background tasks which monitor how long it has been since a specific NAND block has been written and will actively refresh blocks which are showing higher bit error rates or are near the end of their designed retention period.

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

Table 1: JESD218 data retention specification

An SSD, during the entirety of its rated endurance, must be able to meet these power-off data retention requirements. Once the NAND inside the SSD has exhausted its rated endurance there is no longer any expectation that the data will be retained for a specific length of time. It may seem counterintuitive that a client SSD has a longer data retention period than an enterprise SSD. However, typical use cases for enterprise SSDs are 24/7 power-on operation from the time a drive is deployed until it is decommissioned with only limited power interruption. Client SSDs, on the other hand, may be used in consumer devices that will have limited power-on operation each day, and may have long periods of non-operation. Thus, the expected duration for power-off data retention is longer.

Client and Enterprise Workloads

NAND is specified to meet the client or enterprise data retention requirement at a certain number of PE cycles. SSDs are specified for a certain amount of data written by the host. In between, the SSD controller will have a flash translation layer (FTL) and firmware which will handle writing data, erasing blocks, performing garbage collection, handling background tasks. The FTL will typically require more PE cycles than the amount of host data written, with the ratio between the two known as write amplification (WA). Many factors influence WA, such as NAND page and erase block sizes, SSD overprovisioning, quantity of DRAM, and firmware. One of the most impactful, however, is workload. Large-block sequential write workloads typically have low WA, while small-block random write workloads typically have high WA.

Because SSDs are general block storage devices it is unknown what sort of application and workload the SSD will be deployed into when sold. In order to standardize endurance ratings across the industry we rely on JEDEC to define the workload under which endurance is characterized, this time the JESD219 specification. The specification defines the workloads, the preconditioning requirements for the SSD prior to test, and the test procedure, to determine a drive's endurance rating.

Similar to differences in SSD Class and Requirements between client and enterprise SSDs, the workloads used for client and enterprise SSDs are very different. The client workload consists of a reference trace of I/O commands that are played back to the SSD. This trace is based on recording commands sent during an extended period of time in a user PC with an operating system installed. The enterprise workload defines a mix of random write operations to the drive of varying block lengths, heavily biased to 4KB and 8KB random write operations, but including both shorter and longer transfer lengths.

An SSD vendor can determine how much WA is occuring in the drive by testing with these workloads. Based upon the raw NAND PE cycle specification and the capacity, an SSD vendor can then calculate the amount of host writes that can be written using that workload to determine its endurance rating. For client devices, this endurance rating is typically provided in terabytes written (TBW), whereas for enterprise devices, this endurance rating is typically provided in drive writes per day (DW/D). In either case, the rating tells a user how many writes the drive is expected to withstand before reaching the NAND's raw PE cycle specification.

Of critical importance, however, is that an endurance rating is based on a specific workload. Real-world workloads will invariably differ from the workloads used for rating. Users selecting an SSD may use this endurance rating as an estimate of how long a drive will last in their application but their application may result in higher or lower WA than the JEDEC workloads used for rating. Thus, the most important metric for understanding a drive's endurance will be the actual percentage life remaining provided by the drive firmware. Modern SSDs, whether SATA, SAS, or NVMe™, will report their percentage life remaining in SMART or log data. This percentage life remaining will be based on the actual PE cycles of the NAND utilized in the drive, not based on host writes.

Endurance vs Reliability

Beyond endurance, SSDs may also have ratings for mean time between failure (MTBF) and/or annualized failure rate (AFR). This rating is mostly unrelated to endurance.

MTBF/AFR is the likelihood of failure of the SSD. It is typically based upon the failure rates of the various individual electrical components within the SSD, using a statistical model that determines the expected failure rate for the drive as a whole. Drive failure is a random process that will be expected to affect a certain portion of a drive population over time. As with any electronic system, this failure rate is influenced by total activity of the system. Drives that are more heavily used, and more heavily written, will have higher incidence of these failure modes than drives which are mostly unused. In addition, drives operated at higher temperatures will be expected to fail at higher rates, and failure rate expectations should be derated if operated at high temperatures.

The endurance rating, on the other hand, predicts an individual drive's ability to read back the data which has been stored on the drive. A drive which has reached its endurance rating may be able to operate without electrical failure for a significant lifetime beyond its endurance rating but read error rates will significantly increase and the drive should no longer be trusted to store data for any period of time.

Takeaway

SSD endurance is based upon a well-known phenomenon, the predictable rate of degradation of NAND flash cells as they undergo repeated PE cycles. The endurance rating relates to the ability of the drive to retain stored data over a certain period of time, and should not be interpreated as having a meaningful relationship to the likelihood of drive failure.

For data that is important, drives which have reached their endurance rating should no longer be used, as there is no specification for how quickly the data retention falls off after the endurance rating has been reached. In that case, the drive should be decommissioned and replaced.

HDD Workload Rating

Physical Structure

An HDD is a complex electromechanical device, consisting of electrical components and mechanical components that are designed and built to operate within extremely tight tolerances. Drives today can have as many as 10 platters, the disks upon which data is stored, with up to 20 read/write heads that aerodynamically fly at nanometer scale above the platters. The design and manufacture of an HDD is truly interdisciplinary and the task of building, selling, and shipping millions of high-quality and reliable HDDs every quarter is nothing short of herculean.

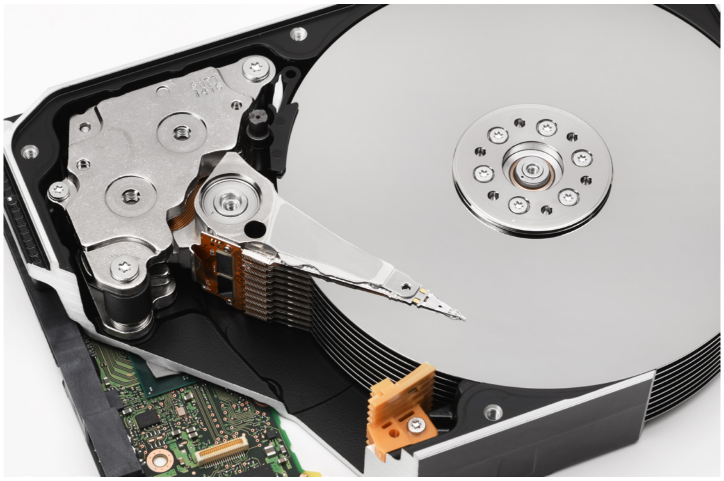

Figure 2: Cutaway image of 10-platter HDD

As complex as these devices are, there are many potential failure modes. Obviously, they are delicate and fragile devices, and external shock can cause failures. The fly height of the heads above the platter is lower than the size of most particles, and particles between the head and platter can scratch and damage the platter or accumulate on the head leading to failure. To counteract this, HDDs are built in cleanroom environments to avoid contamination that will cause failures. As with any mechanical device, operating temperature and time in use as measured in power on hours (POH) correlate to failure rates.

Advances in recording technologies that have reduced the functional flying height of the heads only during read and write activity, and additional energy-assisted recording technologies have led to the introduction of a new specification for HDDs: workload rating.

Impact of Read/Write Workload

As HDD areal densities increased, the need to move the read-write head closer to the platter to improve the capability to read and write to the platter became apparent. However, keeping the heads flying at read-write height at all times with these areal densities increased the likelihood of damage due to head-disk interaction (HDI), the highest pareto failure mode of an HDD. Something was needed to adjust the fly height in order to read and write, but only during read and write operations.

In 2007, the company HGST, acquired by Western Digital in 2012, introduced a technology called thermal fly-height control (TFC). This technology uses heat to force the pole tip of the slider, containing the read-write head, to protrude closer to the platter during read and write operations. The pole tip can then retract and fly at a higher and safer height when not actively transferring data. This allowed the best of both worlds; reduced spacing for read and write operations and safer fly height at all other times. Similar technologies are now used in all commercial hard drives.

The effect was that the likelihood of HDI only became relevant during reads and writes, and not during other idle modes where there was no active data transfer. HDD reliability no longer correlated strongly with POH; it began to correlate with the amount of data transferred to and from the platters. Across the industry, HDD workload ratings, measured as amount of data transfer from the host as expressed in terabytes/year (TB/yr), were born.

In modern drives, additional recording technologies have been introduced and will be introduced that use additional energy within the head during write operations. These technologies are referred to as energy-assisted magnetic recording (EAMR). Technologies such as ePMR, MAMR, and HAMR will make the total write workload a factor in failure rate due to this additional energy flowing through the head.

Workload/Temperature is part of reliability calculation

There is no known and predictable degradation model for how an individual head will degrade with continued read-write activity. This differs from NAND, where the breakdown of cells as a result of PE cycles is known and the NAND components are rated per industry standard (JEDEC) specifications. Thus, there is no belief that an HDD will become increasingly “worn out” based on its workload.

Instead, HDD reliability values are based upon substantial large population testing and statistical models that state that if you have a large enough population of drives it is reasonable to predict failure rates across the entire population given known workload and temperature. Western Digital reliability testing of each drive model encompasses the wide majority of field workloads. This ensures that potential failure modes of the most common workloads are well understood and modeled, and within our testing envelope.

Typically, HDDs will be given maximum workload and temperature values. In the case of Western Digital Ultrastar HDDs, these are typically 550 TB/yr of data transfer and 60°C maximum drive-reported temperature. These values are based on a maximum upon which reliability calculations can be predicted for workload, and upon a maximum that an HDD can be safely operated for temperature.

MTBF/AFR for HDDs should be considered on a continuum of workload and temperature within those maximum ratings. The datasheet MTBF/AFR values will be based upon a typical workload and temperature rather than maximum. These must be derated positively or negatively based upon the actual workload and temperature that a specific application will present to the drive. For drives which incorporate EAMR technologies, there may be additional derating, positive or negative, based on the balance between writes and reads in the workload.

Why Doesn't Workload Increase with Capacity?

SSD endurance specifications typically increase linearly with capacity. Due to wear-leveling, the PE cycles will be distributed roughly equally across all NAND components in the SSD. The more NAND is available at a specified PE cycle rating, the more write endurance is available to the host.

For HDDs, there is no predictive degradation model that defines the failure mechanism of a specific component as a function of data transfer. There is also no ability to wear-level the amount of data that is transferred through any specific head. While populating a greater number of heads and platters in a higher-capacity drive will mean that on average each head will transfer less data, the higher number of components provides more potential points of failure.

The failure of a single head is typically deemed a failure condition for the entire drive. For example, if a large drive containing 20 heads experiences a failure of only 1 head the drive must be taken out of service and replaced. In large datacenter deployments this can be mitigated by utilizing a repurposing depopulation (RDP) feature, which allows a drive with a single failed head to be reformatted at a lower capacity with that head unused. In smaller deployments, however, RDP is unlikely to be implemented.

Western Digital data over many generations of products shows that the factors of workload and component count roughly balance each other. Across a spectrum where a drive has 6 heads and 3 platters up through a drive which has 20 heads and 10 platters, the failure rates for a given host workload are close to identical.

Takeaway

HDD reliability, many years ago, was dominated by simple factors such as the number of hours a drive was powered on, and the temperature of the drive when powered on. In order to advance technology to reach the capacities we see today, the fly heights during data transfer—and thus the workload—became relevant, and the total POH of the drive became less meaningful.

In a single-drive scenario, an HDD that is operated beyond its workload rating does not mean that there is any specific concern about its ability to perform its duty. A drive that is beyond its workload rating that is still operating properly—and for which SMART or SAS log data do not show any red flags—is not expected to individually fail simply by being beyond the rating. HDD workload ratings are best understood as a population-level prediction of failure rates.

Conclusion

SSD endurance ratings are based on known, predictable, attributes of the underlying storage media. When an SSD reaches the end of its endurance, that individual SSD should no longer be trusted for long-term storage of data, even if the drive is not in a "failed" state.

HDD workload ratings are not predictive expectations of a known wear-out mechanism at the individual HDD level. HDD workload ratings are based upon population-level reliability testing. They are best used as one factor, albeit a very important one, in the expected statistical failure rates of a large population of drives in a specific workload.

All Western Digital storage devices, whether HDD or SSD, are designed and produced to exacting reliability standards as measured by MTBF/AFR. SSDs have endurance specifications, and HDDs have workload ratings, and it is easy to conflate the two. However, these measure very different phenomena and should not be considered the same thing.

Learn More

Data Center Solutions

SSD vs HDD - Choosing a Drive that's Right for You

Technology Brief - SSD Endurance Explained

Whitepaper - Therman Fly-height Control (TFC) Technology

Western Digital®

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com

January 2023