SNIA CMSI | COMPUTE, MEMORY, AND STORAGE

Endurance of NVMe®, SAS, and SATA SSDs

Jonmichael Hands
Co-Chair, SNIA Solid State Drive Special Interest Group
Strategic Planner, Intel Corporation

Table of Contents

List of Tables

Introduction – An Overview of Solid State Drive (SSD) Endurance

SSDs have finite endurance, or the amount of data you can write to the SSD before the device wears out and can no longer store data safely. The SSD industry uses the term endurance, but it is also referred to as SSD life or SSD wear out. SSD vendors generally specify this in two ways, TBW (terabytes written) or DWPD (drive writes per day) which is supposed to be an easy metric of how much you can write to the device every day of the warranty period.

SSD endurance will vary greatly from what the SSD vendor specifies due to the dependency on the workload (random write vs sequential write, large block size vs small) and as a function of free space or "overprovisioning" on the SSD. Mainstream SSD firmware use unwritten LBAs as spare area for garbage collection until written to, and can mark used LBAs free again with a “TRIM” command. The most important thing is that endurance can be accurately measured and estimated with a few simple equations.

Terminology and Math Related to Endurance

- NAND P/E Cycles: amount of program / erase cycles NAND can do before wearing out.
  - NAND programs (writes) in pages and erases in blocks (contains multiple pages)
- Wearing out: SSD no longer meeting UBER (uncorrectable bit error rate), retention (keeping data safe while powered off), functional failure rate, or user capacity
- UBER = number of data errors / number of bits read
- $UBER = \frac{number\ of\ data\ errors}{number\ of\ bits\ read}$
- WAF (Write Amplification Factor) = NAND writes / host writes
- $WAF = \frac{NAND\ writes}{host\ writes}$
- TBW or PBW – amount of host writes to SSD before wearing out
- $TBW = \text{drive capacity} \times cycles \div WAF$
- DWPD (drive writes per day): amount of data you can write to device each day of the warranty (typically 5 years) without wearing out
- DWPD = TBW/365/warranty/drive capacity
- $DWPD = \frac{TBW}{365 \times 5 \times drive\ capacity}$
- Overprovisioning is the amount of spare NAND capacity that is used for garbage collection, wear leveling, and background operations. SSDs also need some amount of spares or reserve blocks for failures.
- WAF is inversely proportional to the amount of overprovisioning, the closer OP gets to 50% the closer the WAF will get to 1 (exactly the amount of spare area vs user data should yield almost perfect write amplification)
- TRIM: deallocate in NVMe marks LBAs as not in use so the SSD can claim the space back to use. TRIM is important for keeping WAF down because the host has to tell the SSD which data is not in use. TRIM gets sent during a format or sanitize operation, during filesystem creation (quick format in Windows, discard in Linux), and during filesystem deletes. TRIM is the communication between the host software and SSD to show which data is needed, and tell the SSD when it is not needed anymore so that it can reclaim the space.

- Note: The operation of the Deallocate function is similar to the ATA DATA SET MANAGEMENT with Trim feature described in ACS-4 and SCSI UNMAP command described in SBC-3.
- Most common filesystems in Linux disable *discard* on mount in favor of doing a scheduled *fstrim* task, which sends TRIM commands to all unused space in the file system on a daily or weekly basis. Enabling discard will improve endurance and performance by sending TRIM immediately when files are deleted, but may decrease performance, latency and quality of service due to blocking IO commands. This will vary greatly between drive model, interface, and firmware handling, as newer drives generally handle this type of workload much better.

Differences in Endurance Between Client/Consumer and Enterprise/Data Center SSDs

SSD vendors use JEDEC spec for endurance (JESD219) to demonstrate TBW. This is with a fixed workload trace specific to segment (client or enterprise).

Table 1 - SSD Classes and Requirements JESD218B.01

Application Class	Workload (JESD219)	Active Use (power on)	Retention Use (power off)	UBER
Client	Client	40° C 8 hrs/day	30° C, 1 year	$\leq 10^{-15}$
Enterprise	Enterprise (10% 512B-4k, 67% 4k, 23% 8k-64k)	55° C 24 hrs/day	40° C, 3 months	$\leq 10^{-16}$

Different classes of drives have very different endurance characteristics. Consumer drives often employ a cache (like dynamic or static use of SLC NAND) to absorb the writes. This means that small bursty workloads that don't spill out of the cache will have great endurance and performance. This is done to improve performance in common scenarios, make the spec sheet look better, and give a boost to small capacity SSDs. Most consumer workloads are write once, read many (like installing a game and then playing it). Only heavy content creators and power users are regularly moving around tens to hundreds of GB of data.

Data center SSDs at a similar capacity may look like they have worse performance than a high end consumer SSD. Data center SSD have prioritized performance consistency, quality of service, and worst case workload for measurement. Consumer SSD performance is often tested when the drive is

empty or “fresh out of box” in the best conditions (which makes sense, as most users workloads live frequently in the cache). Generally speaking, data center NVMe SSDs are higher power, high performance, and higher endurance than consumer drives and are rated for continuous workloads. Consumer NVMe SSDs also employ various low power states to save battery life, where most often in data center these are turned off.

NVMe Features Related to SSD Endurance

Overprovisioning SSDs

Overprovisioning is the amount of spare NAND capacity that is used for garbage collection, wear leveling, and background operations. SSDs come with a factory amount of overprovisioning (or may be called “spare” area) with capacity that is not accessible to the host. SSDs need some reserve or spare blocks in the event of defects and failures inherent over time in NAND flash. Some amount of overprovisioning is paramount to SSD firmware to function and do garbage collection. In NVMe the easiest way to overprovision a drive is to delete the namespace, and create a new one that is smaller. Alternatively, a workload can specify an LBA range to write that is smaller than the total size of the drive. This can also be achieved through the use of partitions.

Streams

Streams is a feature in NVMe called “Directives” that was added in NVMe 1.3 which allows the host to tag and classify data with a stream ID, which the SSD can use to do intelligent data placement. The purpose of this feature is to put data with different velocity (e.g. hot, warm, cold) into different physical locations to improve garbage collection efficiency and reduce WAF. The SSD firmware can then decide where on NAND to physically place data that is tagged with the same stream ID in the same set of erase blocks on NAND so that when garbage collection happens, efficiency is improved and write amplification decreases.

Sets

NVM Sets was added in NVMe 1.4 to be able to both logically and physically isolate data.

An NVM Set is a collection of NVM that is separate (logically and potentially physically) from NVM in other NVM Sets. One or more namespaces may be created within an NVM Set and those namespaces inherit the attributes of the NVM Set. If data from different workloads or hosts is placed on their own NVM Set, write amplification should be improved due to not mixing velocity of data, and quality of service will improve by avoiding the noisy neighbor problem.

Endurance Groups

This feature allows for multiple NVM Sets to be part of an endurance group, that all share endurance and will be wear leveled together (keeping endurance of the NAND similar across die).

ZNS

The new NVMe Zoned Namespaces has great potential for further improving endurance by breaking up an NVMe SSD into Zones. ZNS is a brand new command set, and the scope of ZNS is far more complicated than can be covered here. The drive, host, file system, and software all have to be ZNS aware for this to work properly. Zones are sequentially written, and can be the same size or larger than a NAND erase block. ZNS aims to solve many challenges for storage workloads by eliminating the need for overprovisioning, doing garbage collection at the zone level with zone resets, and forcing write amplification to be close to 1. The tradeoff for this amazing improvement in SSD cost and endurance is more complex software and management at the host, and the ability to stage data before writing to durable storage.

Examples of Drive Models in Each Class

SNIA recently published a list of different types of NVMe SSDs for data center and enterprise use.

https://www.snia.org/technology-focus-areas/physical-storage/nvme-ssd-classification

Endurance is a large differentiator in the cost of SSDs due to media type (MLC, TLC, QLC) having different program erase cycle capability with various ECC engines, as well as overprovisioning (more overprovisioning = more NAND = more cost). It is typical now to see 1 DWPD for mainstream read intensive use in enterprise server, as well as cloud applications that tune the endurance by overprovisioning to specific workloads. The enterprise segment also uses 3 DWPD for “mixed use” which is more suitable for caching, database, and higher write performance workloads. New storage class memory SSDs, such as Intel Optane, don’t use NAND and have a very different endurance capability, with drives in the market ranging from 30, 60, and 100 DWPD.

Estimating Endurance, Measuring WAF, and Monitoring Endurance Through Software

WAF can be estimated:

- WAF will be close to 1 for sequential workload, close to 5 for 1 DWPD class drive, 2-3 for 3 DWPD class drive.
- WAF is a function of garbage collection efficiency. If drive is <60% full and TRIM commands are getting sent to SSD during file deletes, then WAF should be close to 1. WAF will get worse the more random the data pattern for writes, and the more full the SSD is. Enterprise SSDs are speced at JEDEC, which is close to full LBA span (100%) random write where drive is full and preconditioned (worst case)

WAF can be measured:

- Read SMART data from drive to get host writes, read vendor specific log page to get NAND writes
- Run workload with known amount of data writing (or take ave MB/s * time ran)

- Read SMART again
- Calculate with NAND writes / host writes

With an estimated or measured WAF, total endurance in TBW or DWPD can be easily calculated! In the case where a vendor specifies a given TBW at worst case, one can estimate or measure the worst case WAF and find the true program erase cycles.

Reading endurance with NVMe-CLI - this is the gas gauge that shows total endurance used:

```
sudo nvme smart-log /dev/nvme0 | grep percentage_used
```

Reading amount of writes that the drive have actually done

```
sudo nvme smart-log /dev/nvme0 | grep data_units_written
```

smart-log data units written can be confusing…need to reference the NVMe spec to decode the output

Bytes written = output * 1000 * 512B

TBW = Bytes written * 1000 * 512B / (1000^4)

TiBW (binary) = Bytes written * 1000 * 512B / (1024^4)

Source, NVM Express 1.4 section 5.14.1.2 SMART / Health Information (Log Identifier 02h)

Data Units Written: Contains the number of 512 byte data units the host has written to the controller; this value does not include metadata. This value is reported in thousands (i.e., a value of 1 corresponds to 1,000 units of 512 bytes written) and is rounded up (e.g., one indicates that the number of 512 byte data units written is from 1 to 1,000, three indicates that the number of 512 byte data units written is from 2,001 to 3,000). When the LBA size is a value other than 512 bytes, the controller shall convert the amount of data written to 512 byte units.

For the NVM command set, logical blocks written as part of Write operations shall be included in this value. Write Uncorrectable commands and Write Zeroes commands shall not impact this value.

A value of 0h in this field indicates that the number of Data Units Written is not reported.

For NAND writes and calculating WAF, you need vendor plugins for NVMe-CLI (built in). Showing the command for an Intel NVMe SSD

```
sudo nvme intel smart-log-add /dev/nvme0
```

To find out NAND writes, you will have use the vendor plugins for NVMe-CLI.

```
sudo nvme <vendor name> help
```

Example with an Intel SSD

```
sudo nvme intel smart-log-add /dev/nvme0
```

In SATA you can use the following commands

```
sudo apt install smartmontools
sudo smartctl -x /dev/sda | grep Logical
sudo smartctl -a /dev/sda
```

looking for `Media_Wearout_Indicator`

note this does also work for NVMe for basic SMART health info

```
sudo smartctl -a /dev/nvme0
```

SAS

```
sg_logs /dev/sg1 --page=0x11
"Percentage used endurance indicator: 0%"
```

Additional Resources

- https://www.jedec.org/sites/default/files/Alvin_Cox%20%5BCompatibility%20Mode%5D_0.pdf
- https://www.jedec.org/standards-documents
  https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/over-provisioning-nand-based-ssds-better-endurance-whitepaper.pdf
- http://intel.com/endurance
- https://wintelguy.com/endurance-calc.pl

About the SNIA

The Storage Networking Industry Association (SNIA) is a non-profit organization made up of member companies spanning information technology. A globally recognized and trusted authority, SNIA's mission is to lead the storage industry in developing and promoting vendor-neutral architectures, standards and educational services that facilitate the efficient management, movement, and security of information. More information is at www.snia.org.

About the Compute, Memory, and Storage Initiative

The Compute, Memory, and Storage Initiative supports the acceptance and growth of computational storage (CS) solid state storage (SSS), and persistent memory (PM) in the marketplace. Our member companies support SNIA work in their technology focus areas of computational storage, physical storage, and persistent memory; promote their results in the marketplace, influence standards activities, and educate vendor and user communities. Find out more at www.snia.org/forums/cmsi.

SNIA CMSI | COMPUTE, MEMORY, AND STORAGE

Storage Networking Industry Association
4360 Arrows West Drive • Colorado Springs, CO 80907 • Phone: 719-694-1380 • Fax: 719-694-1389 • www.snia.org