SSD ENDURANCE

Application Note

Document #AN0032 – Viking SSD Endurance | Rev. A

viking TECHNOLOGY

A RF, Optical, Microelectronics and Memory Company

Table of Contents

Table of Tables

1 Introduction

Endurance is the total amount of data that can be written to the SSD over its full lifespan. SSD manufacturers measure it in two different ways:

TBW – terabytes written, which is the total data that can be written to the drive.

DWPD – Drive Writes Per Day is the total number time the drive can be filled to full capacity per day

The relationship between TBW and DWPD is:
TBW = DWPD * CAPACITY * 365 (days) * years of warranty

2 Factors Affecting Endurance

It is this estimation of SSD life available and is used to determine the relationship between host writes and NAND cycles. The latter being the number of program/erase cycles applied to any NAND block, and use this relationship to estimate the SSD endurance rating.

The relationship may be different for different types of NAND components used in the SSD. Consider an SSD containing only one type of NAND and no features of the drive design that would make the WAF (Write Amplification Factor, determined by manufacturer) change over the lifetime of the drive. Suppose further that the design of the wear leveling method is expected to result in the most heavily-cycled erase block receiving twice the average number of cycles.

The WAF may be obtained from SSD data using the specified workload for endurance testing. Measurement of WAF requires access to information about NAND program/erase (P/E) cycles. Under the assumption in this example where WAF is constant, WAF may be measured after operating the SSD long enough to reach a steady state, without needing to operate the drive to its full endurance rating. This defines workloads for the endurance rating and endurance verification of SSD application classes.

Qualification of a solid state drive involves many factors beyond endurance and retention, so such qualification is beyond the scope of this standard, but this standard is sufficient for the endurance and retention part of a drive qualification. Endurance applies to solid state drives based on solid-state non-volatile memory (NVM).

NAND Flash memory is the most common form on memory used in solid state drives.

Erase blocks used by the SSD during read, write, or erase operations at a specific point in time. NOTE The SSD may have additional erase blocks besides those in the current cycling pool that may be used as spares or for other purposes. The cycling pool is typically larger than the user-accessible LBA count.

The relationship between TBW, write amplification factor and the wear-leveling efficiency are highly dependent on the workload applied for the characterization of endurance. TBW rating may also be expressed in an alternative form called drive-writes-per-day (DWPD), with the assumed lifetime also stated. The value of DWPD is TBW/(C*Y*365), where TBW is the endurance rating in terabytes written, C is the capacity in terabytes, and Y is the lifetime in years that is stated along with the DWPD rating.

Erase blocks are the smallest addressable unit for erase operations, typically consisting of multiple pages. Gigabyte (GB) is approximately equal to 109 bytes when used in reference to SSD capacity.

TBW specification for example, a 100 GB drive with a 100 TBW specification could be in the same family as a 50 GB drive with a 50 TBW specification. Terabyte (TB) For the purpose of this standard, a terabyte is equal to 1·1012 bytes.

Wear leveling methods employed by the drive to spread the p/e cycles across the NVM physical locations even when the workload may be unevenly distributed across the logical drive capacity. The detailed sequence of host writes and reads (including data content and timing) applied to the drive during endurance testing.
Write amplification factor (WAF) The data written to the NVM divided by data written by the host to the SSD. For the purpose of calculating WAF, data written by the host is considered to be in multiples of drive capacity. For example, if 150GB of data is written to an SSD with capacity of 100GB, the data written by the host is considered to be 1.5. Also, data written to the NVM is considered to be the average number of p/e cycles experienced by NVM blocks in use in the SSD. For example, if the average number of p/e cycles is 3, then the data written to the NVM is considered to be 3. In this example, the WAF would be 2. Write amplification factor will depend on the workload and may vary over the lifetime of the device.

Since the endurance of an SSD is dependent up the workload applied to it and the conditions (temperature, duty cycle, etc.) in which that workload is applied, it is necessary to define standard application classes under which endurance for a particular type of device is to be rated and verified. This allows devices within a particular class to be compared to one another as regards the standard endurance rating.

These classes are not all-inclusive and it is understood that variations such as the operating system and application architecture make a significant impact to the workload of the SSD. These classes provide a means to provide parameters for standardized endurance ratings so that the end user may use the endurance rating as a factor in

determining if an SSD is suitable for a particular application. This standard defines two application classes: Client and Enterprise.

3 SSD Application Class Definitions

SSD application classes are defined by usage model and the workload associated with an application. These definitions provide a common set of guidelines around which to specify SSDs. Not all SSD suppliers follow these guidelines however, and it is not mandatory. At the moment, the JEDEC JC-64.8 SSD committee defines application classes only for client and enterprise SSDs in document JESD218. The workloads associated with these application classes are explained in JESD219. Viking recommends a review these document to get a better understanding on benchmarking and comparing SSD performance numbers from different vendors. JEDEC definitions are helpful in specifying client and enterprise SSDs, but they don't cover all of the considerations for datacenter, embedded or military SSDs. Therefore, it is important for designers to always look at SSD datasheets to fully understand the assumptions and conditions under which the product performance and endurance numbers were specified. For example,

Endurance workload testing might assume the following:

- Active use (power on) time and temperature
- Retention use (power off) time and temperature
- Functional failure and uncorrectable bit error rate requirements

and

Performance workload testing might assume the following:

- Preconditioning
- Testing with min and max entropy (compressed and uncompressed data)
- Varying que depth, various block sizes, LBA boundary (4G, 8G LBA or Full LBA)
- Use of different workload models (i.e. workstation, server, database)

The quantitative requirements for the application classes are defined by how the SSDs are actively used for a period of time, during which the SSDs are written to their endurance ratings (Active Use), followed by a power-down time period in which data must be retained (Retention Use).

The table below outlines these elements for the two JEDEC application classes.

Table 3-1: JEDEC Application Classes

Class	Workload	Active Use (Power ON)	Data Retention[1] (Power OFF)	Functional Failure Requirement	UBER Requirement
JEDEC Client	JE5D219 client	40°C (8 hours per day)	30°C, for 1 year	≤3%	< 1 sector in 10^{15} bits read
JEDEC Enterprise	JE5D219 Enterprise	55°C (24 hours per day)	40°C, for 3 months	≤3%	< 1 sector in 10^{16} bits read

NOTE:
1: After endurance requirement has been met

All of these metrics are interrelated when it comes to endurance and changes in assumptions for one parameter can lead to changes in another.

- Workload - consists of the types of data, file sizes, whether that data is sequential or random, and the read and write requirements of the application.
- Active Use - defines the assumed case temperature inside the host system, generally on the SSD case, at which the SSD is written and read. It also defines how often the SSD is used.
- Retention - defines the storage temperature and the length of time the SSD can be powered off while still keeping the data intact after the SSD has reached its endurance specification.
- Data Retention Time - is an important metric point for industrial SSDs. If the SSD has barely been written, the retention time is significantly longer than an SSD that has been in use for a long time.
- Functional Failure Requirement - outlines the number of "acceptable" failures for a given sample size subject to specifically defined conditions.
- UBER - measures the number of sectors that return an Uncorrectable Bit Error Rate based on the number of bits that have been read.

Table 3-2: Comparison of SSD Application Classes

Definitions	Client/Consumer	Enterprise/Datacenter
Lifetime	500 TBW	1-5 DWPD
Endurance(UBER)	10*E-15	10*E-16
Use Case	Mostly Read (80/20), 8hr Duty cycle, 0 to 70.C 1 –3Yr Service Life	Read & Write Intensive,1-5x DWPD, 24/7 Duty Cycle, 0 to 70C, 5Yr Service Life

4 Enterprise SSD Endurance workloads

The enterprise endurance workload consists of random data distributed across an SSD in a manner similar to some enterprise workload traces that are publicly available for review.

Prior to running the workload, the SSD under test shall have the user-addressable LBA space filled with valid data (e.g., the drive does not return a data read error because of the content of the LBA prior to being written during the test routine). Initialization may not be necessary if the formatted SSD satisfies this requirement.

a) The enterprise endurance workload shall be comprised of random data with the following payload size distribution:

512 bytes (0.5K) 4%
1024 bytes (1K) 1%
1536 bytes (1.5K) 1%
2048 bytes (2K) 1%
2560 bytes (2.5K) 1%
3072 bytes (3K) 1%
3584 bytes (3.5K) 1%
4096 bytes (4K) 67%
8192 bytes (8K) 10%
16,384 bytes (16K) 7%
32,768 bytes (32K) 3%
65,536 bytes (64K) 3%

b) The data payloads greater than or equal to 4096 bytes data payload sizes shall be arranged such that the data payloads less than 4096 bytes are pseudo randomized among the data payloads greater than or equal to 4096 bytes. Data payloads greater than or equal to 4096 bytes are aligned on 4K boundaries.

c) The workload shall be distributed across the SSD such that the following is achieved:

1) 50% of accesses to first 5% of user LBA space (LBA group a)
2) 30% of accesses to next 15% of user LBA space (LBA group b)

3) 20% of accesses to remainder of user LBA space (LBA group c)

d) To avoid testing only a particular area of the SSD, the distribution described in c) is offset through the user LBA space on different units under test such that all of the SSD LBAs are subjected to the highest number of accesses (e.g., SSD 1 has LBA group a applied to the first 5% of LBAs, SSD 2 has LBA group a applied to the next 5% of LBAs, etc).

e) The write data payload size distribution shall be applied to each of the three LBA groups concurrently. The write sequence across the LBA groups may be applied in either a deterministic fashion or randomly, depending on the capabilities of the test tools, so long as the percentage of accesses to each LBA group conforms to those specified in 4c and the LBA access shall be random (not sequential) within each LBA group. An example of a deterministic method is given in Annex A.1 and an example of a random method is given in Annex A.2.

f) Random data is considered to have 100% entropy. The randomization of the data shall be such that if data compression/reduction is done by the SSD under test, the compression/reduction has the same effect as it would on encrypted data. It is acceptable to substitute a few bytes in each sector with metadata (such as the LBA number) or other information in place of the random data in those bytes if the addition of such information provides enhanced test interpretation capability. The random data for the payload may be generated by various means. An informative example script for generating the random data and the read/write distribution across LBA groups is provided in Annex A.2. This script uses open-source software that may be found at vdbench.org.

5 Client SSD Endurance Workloads

The client endurance workload consists of a standard reference trace of standard ATA I/O commands, played back on the target device. The standard TBW rating for a client drive shall be derived for and verified under the following workload conditions:

a) PreCond%full = 100%;
b) trim commands enabled; and
c) random data pattern.

Additional TBW ratings may optionally be derived and verified for different values of these parameters; however, such ratings shall be reported only in addition to the standard TBW rating.

The application of the client workload occurs in two phases:

1. Preconditioning Phase

The preconditioning phase establishes an initial pattern of free LBA distribution within the LBA Address Space of the drive under test. This phase creates the logical fullness of the drive at the start of the test phase, as defined by the PreCond%full parameter specified for the test.

The preconditioning phase comprises the following two steps.

a) Write once, sequentially, to the full LBA Address Space using random data. If additional TBW ratings are being determined with non-random data (see 5.3.3), non-random data shall be used for this step instead of random data.

b) Create Free LBA Space from the Maximum LBA contiguous to the lower LBA address corresponding to the PreCond%full value if the PreCond%full value is less than 100%.

2. Test Phase

The test phase stresses the drive in a manner that is representative of use during long-term operation in a client environment. The test phase comprises multiple runs of a Test Trace. The Test Trace has the same LBA footprint for each iteration. During the test phase, a Test Trace is run repeatedly.

5.1 Client Endurance Workload Traces

A Master Trace is a command trace from which all Test Traces within the SSD_Capacity range of the Master Trace in the trace library are derived. It comprises drive input/output operations captured over extended period of time in a single user PC with an installed operating system that supports trim commands. The Master Trace incorporates all I/O operations to a representative SSD, including those occurring during shutdown/boot and hibernate/restore activity. Operations include the following:

a) write commands;
b) trim commands;
c) flush cache commands.

Trim applies to the ATA command set. The equivalent of the trim command is defined in other interface standards (e.g., SCSI Unmap; NVMe Deallocate). If a device using a command set other than ATA is being tested, the relevant form of trim for the device's command set should be used by the test software.

This Master Trace should not be used to verify the endurance of SSD's utilized in cache applications since the workload may vary from the activity performed in the client workload Master Trace.

If the Master Trace has an LBA Range = M, and it is required to create a Test Trace

with LBA Range = T, then the total length of Cold LBA Ranges within the Master Trace is compressed (or expanded) by M-T in order to create the Test Trace. If the total length of all the Cold LBA Ranges within the Master Trace is C, then the length of each individual Cold LBA Range, of length > 20MB, shall be reduced by a fraction (M-T)/C, or increased by a fraction (T-M)/C, as appropriate. During reduction or expansion, the change in length of an individual Cold LBA Range should be a multiple of 4KB.

A Test Trace contains all commands that are present in the Master Trace. The LBA Range of a Test Trace almost spans the LBA address range of the drive to which it relates. A Test Trace for a specified SSD_Capacity parameter is derived from the Master Trace with the use of compression or expansion if required. The Test Trace may combine consecutive trim commands found in the Master Trace into a single trim command.

5.2 Client workload data patterns overview

The workload trace files only contain the ATA I/O commands. The data payload content of the LBAs is not included in the traces due to the size of such a file not being practical and due to the need for the data to change within the LBA during repeated writes. Some drives may include data compression or reduction techniques. Some encryption software may cause data to become essentially 100% random. Due to these variables, random data shall be used as the standard TBW rating condition. Based on examination of drives in field applications, for additional TBW ratings of drives that incorporate data compression or reduction techniques, the non-random data shall have entropy of approximately 50%.

Random data
Random data shall be generated according to the description for the enterprise workload (see 4f). A local copy of the test trace file may be written to the SSD in place of some random data as required by the testing method.

Non-random data
Non-random data patterns should be created using the non-random data file created by the Perl code located in Annex B.1. The non-random data file generated by this Perl code has approximately 50% entropy. A random byte offset shall be used to ensure that the data selected from the non-random data file has sufficient randomness to exceed the SSD_Capacity. It is acceptable to substitute a few bytes in each sector with metadata (such as the LBA number) or other information in place of the non-random data in those bytes if the addition of such information provides enhanced test interpretation capability. A local copy of the test trace file may be written to the SSD in place of some non-random data as required by the testing method.

Trim

To verify an additional TBW rating in a non-trim-enabled environment, two options exist:

a) disable the trim command function on the drive; or
b) modify the Client workload to remove trim commands

6 Non-Standard TBW Ratings

A minimum sample of two drives shall be tested with the non-standard conditions to be rated (i.e., PreCond%full < 100%, trim commands disabled, and/or non-random data pattern) to establish the steady-state write amplification factor (WAFnonstd). One, two, or three of these variables may be combined in a single test, depending on the desired test configuration. The ratio of the steady-state WAF of the drive family verified with the standard conditions described in 5.1 (WAFrnd) to the steady-state non-standard conditions WAF (WAFnonstd) of the drive is used to extrapolate the TBW with non-random data:

WAFstdWAFnonstd×TBWstd=TBWnonstd

7 Reference Documents

- Viking whitepaper: AN0029 - AN0029 - SSD Application Classifications
- Viking SSD Product Datasheets
  http://www.vikingtechnology.com/products/ssd/ssd.html
- JEDEC Document JESD218: Application classes for client & enterprise SSDs
- JEDEC Document JESD219: Client & enterprise SSDs workloads

8 About Viking Technology

Viking Technology develops and delivers innovative high-technology products that optimize the value and performance of our customers' applications. Founded in 1989, Viking Technology has been providing Original Equipment Manufacturers (OEMs) with industry leading designs, engineering, product support and customer service for 20 years. For more information visit http://www.vikingtechnology.com.

9 Revision History

11/29/17		Initial release. Change # to AN0032
		Add more reference documents. Add more detail throughout.

10 Terms and Definitions

10.1 Cluster:

The minimum # of contiguous LBAs allocated by the host operating system.

10.2 Cold LBA Range:

An LBA Range not referenced by a write or trim in a trace.

10.3 Command trace:

A trace of I/O commands.

10.4 Compression:

The process of scaling LBAs associated with commands in the Master Trace to produce a Test Trace with a smaller LBA Range.

10.5 Entropy:

The non-compressibility of a data pattern when in file format such that 100% entropy has no reduction in file size and 50% entropy has approximately a 50% reduction in file size when converting to a zip file.

10.6 Expansion:

The process of scaling LBAs associated with commands in the Master Trace to produce a Test Trace with a larger LBA Range.

10.7 Footprint:

A set of LBAs.

10.8 Free LBA Space:

A set of LBAs within LBA Address Space, which are not currently allocated for data storage by the host OS.

10.9 LBA Address Space:

The full set of user addressable LBAs on the device.

10.10 LBA Range:

The set of contiguous LBA addresses bounded by the lowest and highest LBA of an addressed structure.

10.11 Maximum LBA:

A value specifying the LBA number corresponding to the last LBA in the LBA Address Space.

10.12 Master Trace:

A command trace typical of a period of continuous drive operation in a typical client environment, from which Test Traces are derived.

10.13 PreCond%full:

The % of clusters in LBA Address Space that are assigned to valid data on the device under test by the preconditioning process.

10.14 SSD_Capacity:

Measured in GBs and may have a useable capacity and a total raw NAND capacity on the SSD

10.15 Test Trace:

A command trace comprising I/O operations to a drive under test that is run during the test phase of a client workload.

10.16 Trim:

A command from the host (e.g. Windows OS) that identifies data that is no longer used by the operating system, allowing the memory space to be considered Free LBA Space by the drive.

Global Locations				
US Headquarters	Canada Office	Texas Office	India Office	Singapore Office
2950 Red Hill Ave. Costa Mesa, CA 92626 Main: +1 714 913 2200 Fax: +1 714 913 2202	500 March Road Ottawa, ON K2K 0J9 Canada	1201 W. Crosby Road Carrollton, TX 75006 USA	A 3, Phase II, MEPZ- Special Economic Zone NH 45, Tambaram, Chennai-600045 India	No 2 Chai Chee Drive Singapore, 109840
For all of our global locations, visit our website under global locations. For sales information, email us at sales@vikingtechnology.com				

viking TECHNOLOGY

A RF, Optical, Microelectronics and Memory Company