JEDEC SSD Specifications Explained

Alvin Cox, Seagate
Chairman, JC-64.8

JEDEC SSD Standards

- JESD218, *Solid State Drive (SSD) Requirements and Endurance Test Method*
- JESD219, *Solid State Drive (SSD) Endurance Workloads*

What do these standards include?

- SSD Requirements
  - SSD Definitions
  - SSD Capacity
  - Application Classes
  - Endurance Rating
  - Endurance Verification
- SSD Endurance Workloads
  - Client
  - Enterprise

Scope of JESD218

- Define JEDEC requirements for SSDs, classes of SSDs, and the conditions of use and corresponding endurance verification requirements.
- The standard is sufficient for the endurance and retention part of SSD qualification.
- Developed for SSDs with NAND NVM.

Reference Documents

- JESD22-A117, *Electrically Erasable Programmable ROM (EEPROM) Program/Erase Endurance and Data Retention Stress Test*
- JESD47, *Stress-Test-Driven Qualification of Integrated Circuits*
- JEP122, *Failure Mechanisms and Models for Semiconductor Devices*
- JESD219, *Solid State Drive (SSD) Endurance Workloads*

Key Definitions

- Endurance failure
  - A failure caused by endurance stressing.
- Endurance rating (TBW rating)
  - The number of terabytes that may be written to the SSD while still meeting the requirements.

Key Definitions

- Erase block
  - The smallest addressable unit for erase operations, typically consisting of multiple pages.
- Page
  - A sub-unit of an erase block consisting of a number of bytes which can be read from and written to in single operations, through the loading or unloading of a page buffer and the issuance of a program or read command.

Key Definitions

- Program/erase cycle
  - The writing of data to one or more pages in an erase block and the erasure of that block, in either order.
- Retention failure
  - A data error occurring when the SSD is read after an extended period of time following the previous write.

Key Definitions

- Solid state drive
  - A solid state drive (SSD) is a non-volatile storage device. A controller is included in the device with one or more solid state memory components. The device should use traditional hard disk drive (HDD) interfaces (protocol and physical) and form factors.

Key Definitions

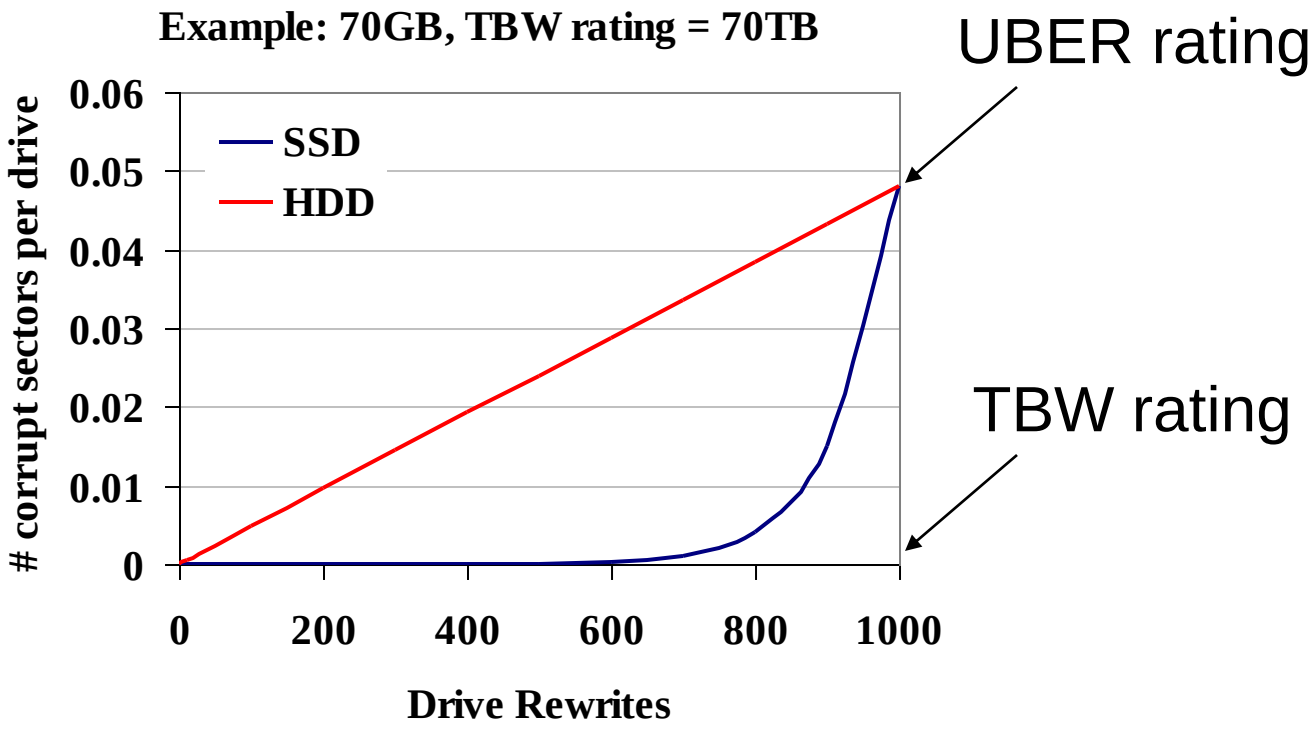
- Unrecoverable Bit Error Ratio (UBER)
  - A metric for the rate of occurrence of data errors, equal to the number of data errors per bits read.

$$UBER = \frac{number\ of\ data\ errors}{number\ of\ bits\ read}$$

Key Definitions

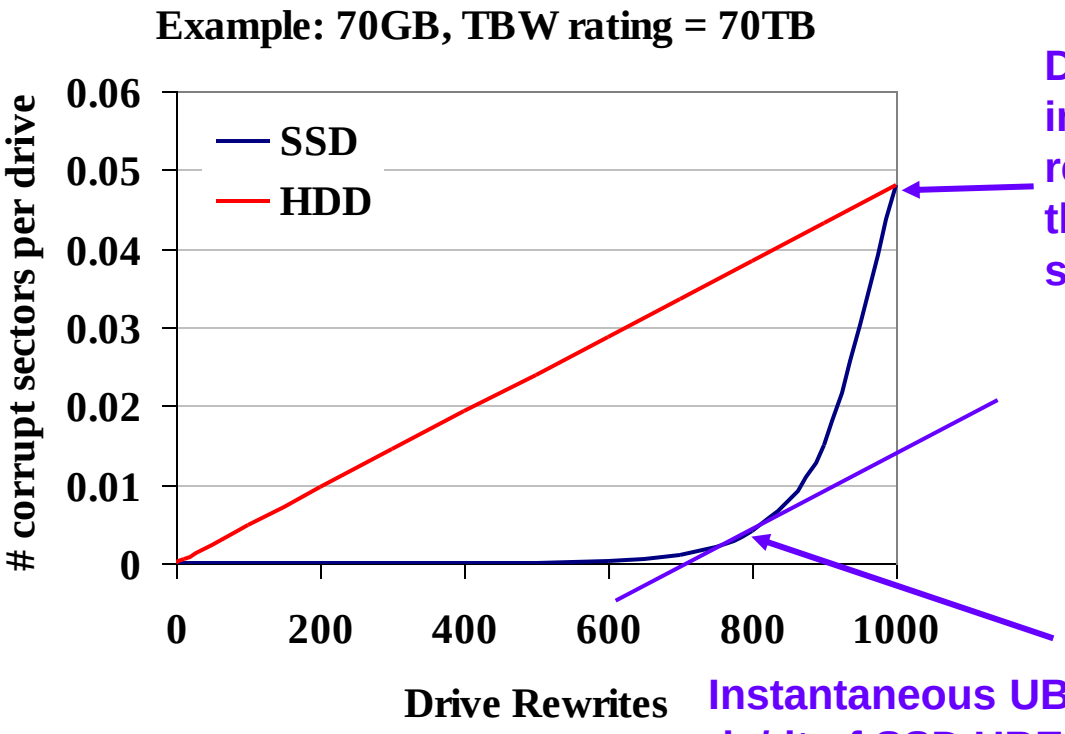
- Write amplification factor (WAF)
  - The data written to the NVM divided by data written by the host to the SSD.

UBER of HDDs and SSDs



Material submitted by Intel

UBER equality question



Does allowing the limit to intersect the constant line result in a much worse UBER at that point in time than the specification limit?

Instantaneous UBER for the spec limit is when dx/dt of SSD UBER = dx/dt of constant UBER = Specification limit (i.e., slope of tangent line to curve equals that of a constant UBER over life slope)

UBER

- Lifetime-average is the standard in reliability (may be more familiar by its equivalent name, cumulative % fail). Instantaneous slope is not used as the basis for qualification.
- Lifetime average is in the draft because it is the most accurate value to use and the most consistent with prior practice.

UBER determination

- Although the UBER concept is in widespread use in the industry, there is considerable variation in interpretation. In this JESD218, the UBER values for SSDs are lifetime values for the entire population.
  - The numerator is the total count of data errors detected over the full TBW rating for the population of SSDs, or the sample of SSDs in the endurance verification.
  - A sector containing corrupted data is to be counted as one data error, even if it is read multiple times and each time fails to return correct data.
  - The denominator is the number of bits written at the TBW rating limit, which aligns to the usual definition of errors per bit read when the read:write ratio is unity.

WAF

- Write amplification factor (WAF)
  - The data written to the NVM divided by data written by the host to the SSD.
- An SSD usually writes more data to the memory than it is asked to write.
- The nature of the workload plays a significant role.

WAF

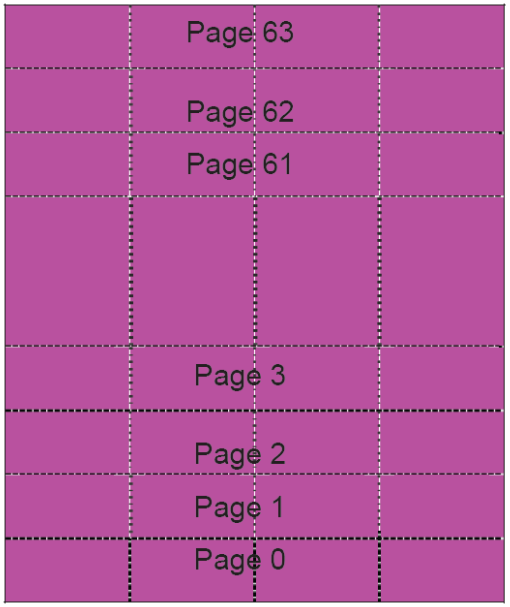
- Factors that impact WAF:
  - Sequential versus random
  - Large transfers versus small ones
  - Boundary alignment
  - Data content/patterns (especially for SSDs using data compression)

WAF example

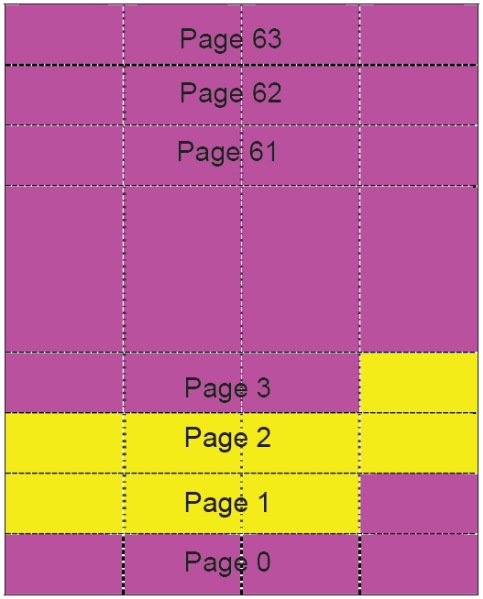
- Because NAND can only be written to after having been erased, and the granularity of erasure is coarse (referred to as the Erase Block), some NAND management algorithms can result in a large amount of data being written to NAND for a modest amount requested to be written by the host. The multiplication factor that describes how much larger the ultimate write to the NAND becomes is known as write amplification. For example, if a host write of 4KB results in a write to the NAND of 16KB, then the write amplification is 4.

WAF example

The figure shows one NAND block comprised of 64 pages. For this example, assume that each page is 2KB in size (four sectors) giving a total of 256 sectors per block. This block has valid host data in all of its pages. Assume that the host writes only some pages in this block with new data as illustrated on the next slide.



WAF example



In NAND, before programming a page it has to be first erased and the erase granularity in NAND is in blocks. Therefore to program the 8 sectors shown in yellow in Figure 2 above, one possible approach is to use a read/modify/write algorithm as follows:

1. Copy entire block (page 0 to page 63) to DRAM.
2. Modify pages 1, 2 and 3 with the new data that the host wants to write. DRAM now has the block as shown in the figure with the new Host data.
3. Erase the block in the NAND.
4. Program the block with the data from DRAM. (This is equivalent to writing 256 sectors in the NAND.)

With this implementation, a host write of 8 sectors resulted in a NAND write of 256 sectors. The write amplification for this example is 32 (256/8).

Material submitted by Intel

WAF example

Note that had the example above used a larger transfer that spanned the entire erase block (or sequential transfers that could be internally buffered in the SSD to fill the entire erase block) that the write amplification would essentially be 1 since the entire erase block would have new data to be written. The nature of the workload has substantial impact on resulting write amplification and in general small random writes tend to create the largest write amplification values.

SSD capacity

- SSD Capacity in Gbytes = (User-addressable LBA count – 21168) / 1953504
  - Same value as IDEMA for HDDs except no 50GB limit
  - Requested by OEMs for ease of implementation (HDD or SSD)
  - This version for 512 byte sectors
  - 4k sector version in development

Application classes

- Current application classes:
  - Client
  - Enterprise
- Application classes attributes:
  - Workload
  - Data retention
  - BER

Endurance rating

- Establish a rating system for comparing SSDs.
- Provides unique rating for application class.
- Rating based on a user-measurable interface activity: TBW.
- TBW = TeraBytes Written.
  - "Decimal" value to be consistent with user interfaces.

Endurance rating

The SSD manufacturer shall establish an endurance rating for an SSD that represents the maximum number of terabytes that may be written by a host to the SSD, using the workload specified for the application class, such that the following conditions are satisfied:

1) the SSD maintains its capacity;
2) the SSD maintains the required UBER for its application class;
3) the SSD meets the required functional failure requirement (FFR) for its application class; and
4) the SSD retains data with power off for the required time for its application class.

This rating is referred to as TBW. Requirements for UBER, FFR, and retention are defined for each application class.

SSD endurance classes and requirements

Application Class	Workload	Active Use (power on)	Retention Use (power off)	Functional Failure Rqmt (FFR)	UBER
Client	Client	40°C 8 hrs/day	30°C 1 year	≤3%	$\leq 10^{-15}$
Enterprise	Enterprise	55°C 24hrs/day	40°C 3 months	≤3%	$\leq 10^{-16}$

Temperatures and data retention

- Tables show # weeks retention as a function of active and power-off temperatures.
- Numbers are based on Intel's published acceleration model for the detrapping retention mechanism (the official JEDEC model in JESD47 and JEP122 for this mechanism).

Client

Power Off Temperature	Active temp 25	30	35	40	45	50	55
55	1	1	2	2	3	5	8
50	2	2	3	4	6	9	15
45	4	4	5	7	10	17	27
40	7	8	10	14	20	31	52
35	14	16	20	26	38	61	101
30	28	32	39	52	76	120	199
25	58	65	79	105	155	244	404

Enterprise

Power Off Temperature	Active temp 25	30	35	40	45	50	55
55	0	0	0	0	1	1	2
50	0	0	0	1	1	2	4
45	0	1	1	1	2	4	7
40	1	1	2	3	4	7	13
35	2	2	3	5	8	14	25
30	3	4	6	10	16	28	50
25	7	9	12	20	33	58	101

Material submitted by Intel

Endurance rating example

Using the appropriate workload, the SSD manufacturer may determine the relationship between host writes and NAND cycles, the latter being the number of p/e cycles applied to any NAND block, and use this relationship to estimate the SSD endurance rating. If the SSD employs more than one type of NAND component with different cycling capabilities, then a separate relationship should be obtained for each type of NAND. If operating the SSD to the desired TBW is impractical because time required would be excessive, then the relationship between NAND cycles and host writes should be extrapolated. In performing the extrapolation, any nonlinearities in SSD operation, such as those resulting from a reduced cycling pool at extended cycles, should be accounted for.

Endurance rating example

The estimated endurance rating is the TBW such that

f(TBW) < NAND cycling capability (1)

where f(TBW) expresses maximum NAND cycles as a function of TBW. 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 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.

Endurance rating example

In that case, WAF would be a constant (for a given workload), and

$f(TBW) = (TBW \times 2 \times WAF) / C$

where C is the SSD capacity and the factor of two is the guard band for the wear leveling effects. The SSD endurance rating would then become

$TBW < (C \times \text{NAND cycling capability}) / (2 \times WAF)$

In the more general case, WAF may not be a constant. More extensive characterization would be needed to determine f(TBW) in equation (1) before estimating the endurance rating.

Endurance rating example

The NAND cycling capability is obtained from component qualification data. 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 cycles which is generally not available to third parties. 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. The guard band for wear leveling effects (two in this example) may be measured from similar SSD data or estimated from the design of the wear leveling scheme.

Endurance verification

Verification goals

- Reasonable test time (1,000 hours)
- Reasonable sample size
- Applicable to multiple designs
- Extendable to an SSD “family”

Produce the right answer!

Direct versus extrapolation

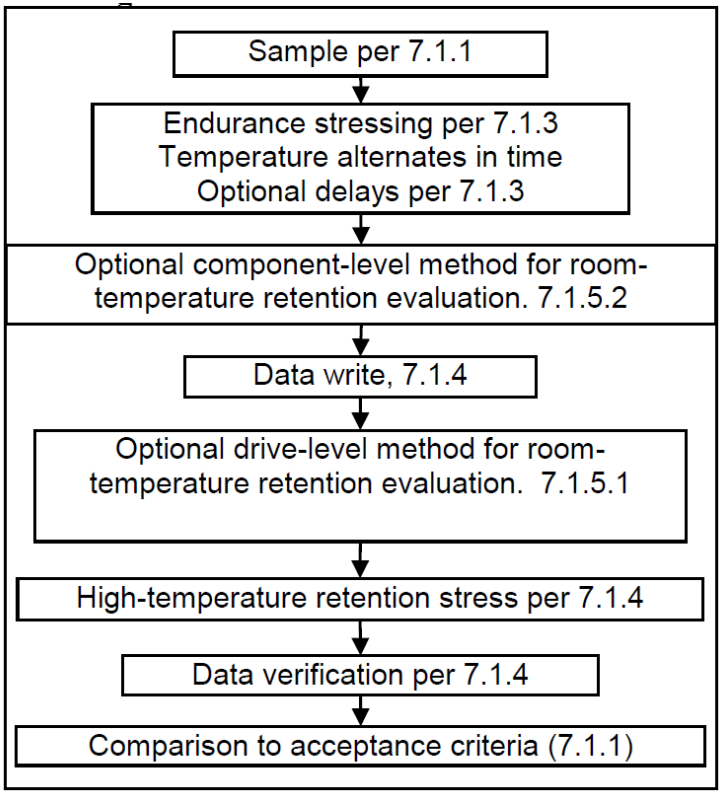
- Two methods described in the specification:
  - Direct
  - Extrapolation
- Direct method runs the SSD to 100% of P/E cycles.
- Extrapolation method requires:
  - Knowledge of component characteristics
  - Information normally available only to the SSD manufacturer

Direct versus extrapolation

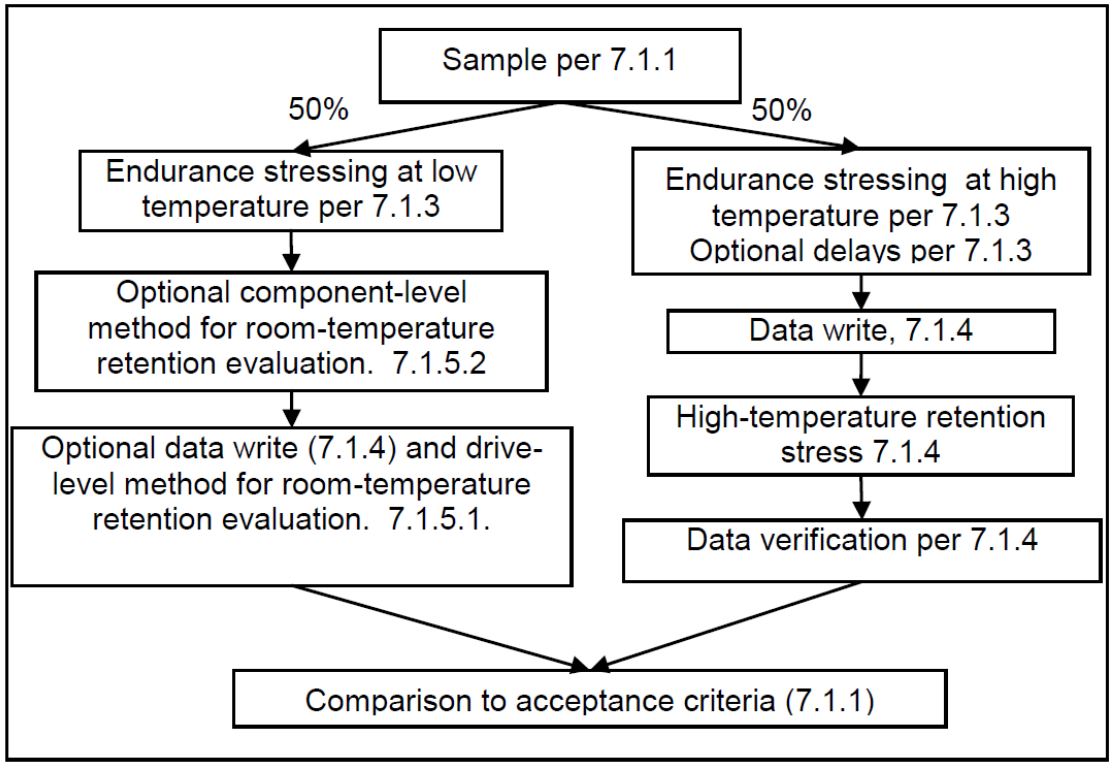
- Both methods:
  - Count only failures based on endurance
  - Include high and low temperature test lots
  - Require sample sizes supporting a 60% statistical confidence level

Temperature lots

- Different temperatures introduce different NAND failure mechanisms.
- It is necessary to test both at elevated and low temperatures.
- Two approaches are acceptable for incorporating both high and low temperatures into the endurance stressing: the ramped-temperature approach and the split-flow approach.
- The preferred temperature measurement is the temperature reported by the SSD if it has that capability (ATA and SCSI statistics).



Ramped-Temperature Approach



Split flow approach

Direct method

- Best method but not likely to be used.
- SSDs are stressed to their stated endurance rating (in TBW) using specified workloads.
- The endurance stressing is to be performed at both high and low temperatures.
- Following this endurance stressing, retention testing shall be performed. Since the retention use time requirements are long, extrapolation or acceleration is required to validate that the SSD meets the retention requirement.

Extrapolation methods

- Extrapolation methods may be used if the direct method would require more than 1000 hours of endurance stress.
- Most of these methods require special access to SSD internal operations or to NVM component information which make these methods possible only for the manufacturer of the SSD.

General requirements for extrapolation methods

1. The SSD must meet the requirements for FFR and UBER for the temperatures and times stated in the table.
2. The FFR and UBER requirements must be met for both low-temperature and high-temperature endurance stressing.
3. Data retention is to be verified under the assumption that the endurance stressing in use takes place over no longer than 1 year at the endurance use temperature and hours per day of use are as specified per application class.
4. Data retention is to be verified both for a temperature-accelerated mechanism (1.1eV) and a non-temperature-accelerated mechanism.
5. All requirements are to be established at 60% statistical confidence.

Extrapolation methods

- Accelerated write rate through modified workload
- Extrapolation of FFR and bad-location trends
- FFR and UBER estimation from reduced-capacity SSDs
- FFR and UBER estimation from component data

Accelerated write rate through modified workload

- The workload in the endurance stress is modified so that more p/e cycles can be performed on the nonvolatile memory in a given amount of time.
- Example of acceptable modified workloads are:
  - 1. A workload with a different ratio of sequential to random writes, or different transfer sizes.
  - 2. A workload which includes proprietary instructions to the SSD to perform internal data transfers, which result in writes that bypass the host.
  - 3. Reduced number of reads.

Extrapolation of FFR and bad-location trends

- The SSD may be stressed to only some fraction of the TBW rating and during the course of the endurance stress, functional failures may occur, as well as a certain number of locations that get marked as ‘bad’. The increase in these two quantities may be plotted as a function of TBW and extrapolated to the TBW rating to obtain estimates of the final levels of FFR and bad locations.

Extrapolation of FFR and bad-location trends

- It is recommended that lognormal or Weibull plotting be used.
- The extrapolated value for FFR must be within the FFR requirements.
- The extrapolated value of bad locations must be lower than can be tolerated by the architecture of the SSD.
- This extrapolation method is not acceptable for verifying UBER.
- Note: The calculation of a 60% confidence limit on the extrapolated values is not straightforward.

FFR and UBER estimation from reduced-capacity SSDs

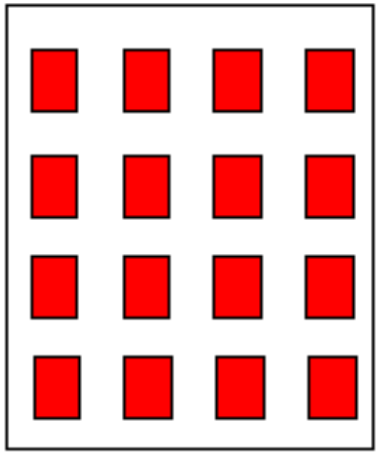
- The capacity of an SSD may be artificially reduced so that some nonvolatile memory components or blocks are not written to, while the remaining ones are written to more extensively than would be the case in the full-capacity SSD.
- The manufacturer is to ensure that the method of capacity reduction does not significantly distort the normal internal workings of the SSD. For example, the number of spare memory blocks may need to be reduced to ensure that the write amplification factor and the ability of the SSD to tolerate a bad blocks does not change.

Reduced-capacity SSD

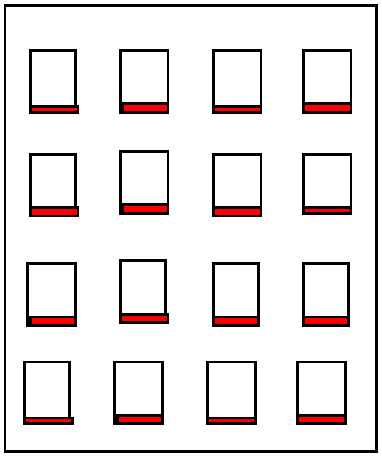
- Qualification family
  - A lower-capacity SSD is used for endurance verification and the data is valid for larger capacities if the qualification family criteria is met.
    - Same nonvolatile memory products, or different nonvolatile memory products that are themselves part of the same component qualification family (defined in JESD47).
    - Same controller and the same firmware
    - Same ratio of TBW specification to capacity

Short stroking

- Term used for HDDs
- Example when used for SSD:



Normal SSD



Short stroked to 20% capacity

FFR and UBER estimation from component data

- With this method, individual nonvolatile memory components are stressed to their target p/e cycles. The temperature, time, and sample size requirements described for the direct method are to be followed, except that the stresses are to be performed on components stressed to the target p/e cycles.
- The calculations of required sample size, measured FFR, and measured UBER shall be scaled according to the number of nonvolatile memory components in the drive and the number of bits written and verified.
- Every effort shall be made to match the stress conditions to those experienced in the SSD, and the criterion for data errors shall be based on the same level of error correction designed into the drive.

Workloads

Endurance workloads

- Client workload is still under development
  - Based on actual captured traces
  - Capacity scaling method being defined
- Enterprise workload specified
  - Uses 100% of SSD user capacity
  - No “typical application” allows use of synthetic trace
  - Full random data to emulate encryption

Enterprise workload

- Start with the SPC profile
  - 4K aligned, all Random
  - 60% Writes, 40% Reads
- In practice, the lack of automated tools for alignment in virtual environments and lack of reporting of alignments within SCSI protocol stacks makes reduction or runts a time consuming manual process, therefore, adjust percentages to add 10% random runt (0.5-3.5K) transfers
  - 4% .5K, 1% each 1K-3.5K
- Address distribution to simulate application usage except making contiguous simplifies test apparatus requirements.

Enterprise workload

- Several workload studies show that less than 5% of the data get >50% of the accesses and 20% of the data get >80% of the accesses.
- Distribution:
  - First drive under test:
    - 50% of accesses to first 5% of user LBA space
    - 30% of accesses to next 15% of user LBA space
    - 20% of accesses to remainder of user LBA space
  - Distribution is offset through the different DUTs.

Enterprise workload

- Random data is used for the data payload. The intent of the randomization is to emulate encrypted data 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.

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%

Questions

Thank You