swissbit®

Application Note

AN2108en

Comparing Specifications

[1] This work is licensed under the Creative Commons License “Attribution 4.0 International”. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Contents

1 Abstract

Whether a flash storage medium is suitable for the intended use case can be determined in advance using the data sheet. In addition to the operating temperature range, the endurance and performance are the most important factors for most applications. For this purpose, the manufacturers usually list values in the data sheet using a selection of common test methods. Depending on the manufacturer, however, different test methods are used, which means that the resulting values are not always easy to compare with products from other manufacturers.

The typical methods, programs and terms that are used for such measurements are explained in this document to provide a way to better compare the characteristics of different SSDs and match them with the intended use.

2 Performance

For performance, it is important to understand which access pattern is used for reading and writing. In addition, the distribution pattern of the previous write accesses has a very large influence on the subsequent read and write speeds.

Basically, all storage media should undergo preconditioning, where all storage addresses are written at least once, before measuring performance. Because NAND flash has no fixed assignments between the physical and logical addresses, the current mappings are updated in management tables, which are stored in the flash. With preconditioning, all logical addresses are assigned to physical addresses in the flash. In the case of read or write access to a logical address, the associated physical address must first be found or, after a write access, the old address must be invalidated. If preconditioning is skipped, the first write access to the logical address is faster than subsequent accesses, resulting in better performance values. Accordingly, it should be noted in the data sheet if the write test was performed on a fresh, out-of-box device.

2.1 Sequential Access

The simplest access pattern in performance tests is sequential access. The medium is filled with data in ascending order over the entire address space, and the data is transferred from the host in large chunks of at least 128KiB. The data itself consists of random data (i.e., data with maximum entropy) to ensure the measurement result is not falsified with (rare) integrated compression storage media. Even with uncompressed media, data that only consists of binary zeros should be avoided. This is because some flash controllers recognize the zeros and, instead of writing the zeros to the flash, simply delete the entries of these memory addresses in the management tables. This means that for a read access, zeros are returned without having to read them from the flash.

The value under *Sequential Write* in the data sheet is therefore the average write speed over the entire capacity, the maximum value instead of the average value, or, if noted accordingly, the average from a section of the total capacity. The problem is that the arithmetic mean value over the entire capacity cannot be compared with reduced ranges or maximum values. This is because many media today have a fast pSLC cache that can accommodate several percent of the total capacity before writing to the slower TLC or QLC starts.

The sequential write speed achieves the highest value of all write tests because the internal administration overhead is minimal and the garbage collection is not used. This value is, therefore, a good orientation for applications in which mainly large files, such as photos or videos, are written.

Corresponding to the sequential write access, there is also a value for the *sequential read* access. Like sequential writes, the medium is filled in ascending order over the entire address space (preconditioned). As a result, the physical distribution in the flash is also sequential, which makes it easy to search for the addresses in the tables. Read accesses are therefore fast, especially since reading is also carried out efficiently in large data chunks.

Here, too, average values over the entire drive cannot be compared with sections or maximum values. The possible existence of a pSLC cache that still contains data from the previous write access also accelerates the read accesses.

2.2 Random Access

In applications with many small files or databases, as well as in multi-user systems or in servers, the accesses to the storage medium are mostly randomly distributed. In addition, the amount of data read and written is rather small. For such applications it is better to use the values for *Random Write* and *Random Read*. Unless otherwise stated, these values refer to data chunks of 4 KiB that are written and read randomly across the entire medium. Each 4 KiB segment is written or read exactly once. The average value for the entire medium is then reported. As with any speed test, the entire medium was previously written completely sequentially in order to have a known and repeatable start condition.

Here, too, a medium will perform better when writing and reading if less than 100 % of the medium is accessed or the maximum value is given instead of the arithmetic mean. Choosing a segment size of more than 4 KiB in this test also increases the throughput. An access size of 4 KiB is typically chosen because this is the smallest standard block size of the file system (i.e., it is the smallest unit that can be read or written). Server applications, therefore, mainly generate this access size.

The performance information for random accesses is usually given in the unit *Input/Output Operations Per Second* (*IOPs*). Because this is the number of 4 KiB chunks transferred per second, multiplying this value by 4 KiB results in the performance in KiB/s.

Random 4 KiB accesses only achieve a slow speed. On the one hand, the administrative effort increases sharply due to the small chunk size on the entire route between the host and the NAND flash. On the other hand, the garbage collection[2] is fully engaged, which reduces performance as searching for the physical address in the administration tables is time-consuming.

This test – if carried out with 4 KiB over the entire capacity – represents the absolute lowest performance limit for read and write accesses. Because no access pattern is slower, this test shows the minimum performance of the medium.

2.3 CrystalDiskMark

Performance values for sequential and random accesses is often based on measurements with CrystalDiskMark. This test program, which is available for Windows only, has several advantages. It's free, easy to use, only takes a few minutes to run, and is non-destructive. It is file-based and does not overwrite existing data. The performance measurement with CrystalDiskMark can, therefore, easily be retrieved with already existing media and compared with other data sheets. It should be noted that only values that have been measured by CrystalDiskMark with an identical major release number should be compared.

[2] see application note AN2101 for more information about garbage collection

The big disadvantage of CrystalDiskMark, however, is the short runtime. The maximum file size is limited and, in the default setting, is only 1 GiB. Storage architectures that have a pSLC cache or, in the case of enterprise SSDs, a large DRAM, achieve very good values with CrystalDiskMark. The massive drop in speed only occurs with larger amounts of data when the switch to TLC/QLC takes place.

2.4 Burst Mode

The *Burst Mode*, or *Burst Transfer Mode*, describes the maximum speed that the medium can achieve. This occurs when the flash controller buffer is used for writes or reads, and the flash is not accessed. Assuming a sufficiently fast flash controller, this speed corresponds to the maximum transmission speed of the interface to the storage medium (e.g., 600 MB/s with SATA-III). This value has no practical relevance as it is only reached for fractions of a second.

3 Endurance

The endurance of a flash storage medium is defined by the number of erase cycles per flash block, after which the medium can hold the data for a defined period of time at a defined temperature without being powered. A typical combination is 3000 erase cycles per block, after which the flash can keep the data at +40 °C for at least one year. The firmware of the storage medium ensures that all flash blocks age, or wear, evenly. As with the performance measurements, the access pattern used when writing also impacts the endurance. For measuring endurance, preconditioning with the complete sequential writing of the medium takes place first. Then the erase counters of the medium are retrieved, and the test program is started. The amount of data that is transferred from the host to the storage medium is then measured. From the amount of data transferred and the increase in the number of consumed erase cycles, the amount of data that can be achieved over the entire service life can be calculated. This value is reported as *Terabytes Written* (*TBW*) in the data sheet.

Often the *Drive Writes per Day* (*DWPD*) are also provided in the data sheet. This value is calculated by dividing the TBW by the capacity of the medium and then by a defined number of days.

Example: A 512 GB medium reaches 560TBW in a certain test until the data retention has reduced to one year when the system is switched off.

$$\frac{560\,\mathrm{TBW}}{512\,\mathrm{GB}\cdot 3\,\mathrm{years}} = 1\,\mathrm{DWPD}$$

One DWPD results over an assumed period of three years. Every day for a period of three years, this test can be used to write the amount of data that corresponds to the storage capacity of 512 GB.

3.1 Sequential Write

With sequential writing, the medium is repeatedly filled with large data segments. The garbage collector is not be engaged because flash blocks are automatically released again. In other words, no data has to be moved internally to create new empty blocks. This means that the maximum write amount is normally achieved here because there is no further noteworthy wear-out from writing the user data. In the case of architectures with pSLC cache, however, the pSLC cache can first reach its cycle limit due to the low level of wear-out on the TLC / QLC area, thereby limiting the overall service life.

3.2 Random Write

In the case of random writing, the garbage collector is fully engaged as it has to constantly move user data in order to free flash blocks again. Internally, a multiple of the amount of data that is written to the medium by the host must be moved. Unless otherwise specified, the chunk size is again 4 KiB. This test represents the worst case for the endurance and is considered unrealistic for almost all application scenarios.

3.3 JEDEC Workloads

The JEDEC Solid State Technology Association has created standards for flash media. It has issued test specifications in the JESD218 and JESD219 standards for the service life measurements. It defines test rules for servers and multi-user systems (“Enterprise”) and single-user computers (“Client”), which should reflect typical use cases. The measured values obtained in this way clearly represent more realistic endurance information than the two extremes of sequential and random writing. Lifetime information for Enterprise and Client workloads should be available in every data sheet.

3.3.1 Enterprise

The so-called *Enterprise workload* simulates typical server applications. It consists of 2/3 of 4 KiB chunks. The rest is made up of various larger and smaller data chunks. Multi-user systems also show a certain locality in the access patterns: so, half of the accesses only take place on 5 % of the capacity. In addition to this restriction, all accesses are random, and the test runs for a week.

The JEDEC demands that the written data must have maximum entropy (i.e., not compressible) in the event the storage medium supports data compression.

The Enterprise workload does not benefit from a pSLC cache. Due to the high number of random accesses, it achieves the best values on architectures with small internal administration sizes. Correspondingly, an optimization for the Enterprise workload leads to a deterioration in sequential write performance.

3.3.2 Client

While the Enterprise workload is a synthetic test, the *Client workload* consists of recorded write accesses in a real system. This recording is played back several times, and the endurance of the medium is then extrapolated. In contrast to the Enterprise workload, which only consists of write commands, the Client workload also contains trim and flush cache commands. While the trim commands support the garbage collector and reduce wear, the flush cache commands force data packets that are still held in the RAM of the flash controller to be written to the flash.

The smallest unit that can be addressed in flash is the page size. Several pages are physically or logically connected to one another, which means that several hundred kilobytes have to be written at once for each write access. The operating systems, therefore, wait several seconds and collect data during this time before they are sent to the storage medium. The flash controller itself waits a few milliseconds to make sure that it has received all the data from the host before starting the write process to the flash. The flush cache commands prevent the controller from waiting. The commands instruct the operating system and the flash controller to write the data received immediately to the flash. As a result, large flash areas, which contain a small amount of user data, are written.

Flush cache commands are used by applications and the file system to protect the data from an unexpected power loss or to create fixed points for the journal of the file system.

As with the Enterprise workload, architectures with small internal administration sizes benefit from the Client workload because it reduces the write overhead with a flush cache command. In contrast to the Enterprise workload, the address space used for the Client workload is limited to almost 60 GB. Systems with a large pSLC cache show a high level of endurance because parts, or even the entire test, can run in the cache, removing the need to transfer to TLC / QLC.

4 Summary

Performance values and endurance information can only be compared between different storage media if exactly the same test method was used. If the general conditions are not evident from the data sheet, it is impossible to assess if the specified values are comparable to other results.

When it comes to life expectancy, the JEDEC workloads specify a precise test sequence so that the values from different manufacturers can be directly compared. You only have to consider whether your own use case corresponds to the Client or the Enterprise workload. Although typical industrial applications do indeed correspond to the write pattern of the Client workload, flush cache commands usually do not occur with this frequency, resulting in a significantly longer service life.

CONTACT US

Headquarters	**Swissbit AG** Industriestrasse 4 9552 Bronschhofen Switzerland	Tel. +41 71 913 03 03 sales@swissbit.com
Germany (Berlin)	**Swissbit Germany AG** Bitterfelder Strasse 22 12681 Berlin Germany	Tel. +49 30 936 954 0 sales@swissbit.com
Germany (Munich)	**Swissbit Germany AG** Leuchtenbergring 3 81677 Munich Germany	Tel. +49 30 936 954 400 sales@swissbit.com
North and South America	**Swissbit NA Inc.** 238 Littleton Road, Suite 202B Westford, MA 01886 USA	Tel. +1 978-490-3252 salesna@swissbit.com
Japan	**Swissbit Japan Co., Ltd.** CONCIERIA Tower West 2F 6-20-7 Nishishinjuku Shinjuku City, Tokyo 160-0023 Japan	Tel. +81 3 6258 0521 sales-japan@swissbit.com
Taiwan	**Swissbit Taiwan** 3F., No. 501, Sec.2, Tiding Blvd. Neihu District, Taipei City 114 Taiwan, R.O.C.	Tel. +886 912 059 197 salesasia@swissbit.com
China	**Swissbit China**	Tel. +886 958 922 333 salesasia@swissbit.com

Disclaimer:
The information in this document is subject to change without notice. Swissbit AG ("SWISSBIT") assumes no responsibility for any errors or omissions that may appear in this document, and disclaims responsibility for any consequences resulting from the use of the information set forth herein. SWISSBIT makes no commitments to update or to keep current information contained in this document. The products listed in this document are not suitable for use in applications such as, but not limited to, aircraft control systems, aerospace equipment, submarine cables, nuclear reactor control systems and life support systems. Moreover, SWISSBIT does not recommend or approve the use of any of its products in life support devices or systems or in any application where failure could result in injury or death. If a customer wishes to use SWISSBIT products in applications not intended by SWISSBIT, said customer must contact an authorized SWISSBIT representative to determine SWISSBIT willingness to support a given application. The information set forth in this document does not convey any license under the copyrights, patent rights, trademarks or other intellectual property rights claimed and owned by SWISSBIT.

ALL PRODUCTS SOLD BY SWISSBIT ARE COVERED BY THE PROVISIONS APPEARING IN SWISSBIT'S TERMS AND CONDITIONS OF SALE ONLY, INCLUDING THE LIMITATIONS OF LIABILITY, WARRANTY AND INFRINGEMENT PROVISIONS. SWISSBIT MAKES NO WARRANTIES OF ANY KIND, EXPRESS, STATUTORY, IMPLIED OR OTHERWISE, REGARDING INFORMATION SET FORTH HEREIN OR REGARDING THE FREEDOM OF THE DESCRIBED PRODUCTS FROM INTELLECTUAL PROPERTY INFRINGEMENT, AND EXPRESSLY DISCLAIMS ANY SUCH WARRANTIES INCLUDING WITHOUT LIMITATION ANY EXPRESS, STATUTORY OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.