Solid state drives (SSDs) and the next generation 3D Flash

Solid state drives (SSDs) have made enormous progress in recent years and an end to the boom is not in sight. This applies in particular to technological advancements by Flash chip manufacturers and the opening up of new fields of application. In addition, SSDs are increasingly replacing hard disk drives (HDDs) at the lower end of the storage capacity range. This article gives an insight into the topics of SSD and Flash card, which in principle have the same structure.
1. TECHNOLOGY: FROM SLC FLASH TO QLC FLASH

NAND Flash, which features cost-efficient production and scalability, has established itself as the memory cell in SSDs. It’s not so long ago that the single-level cell (SLC) was considered state-of-the-art because it is long-lasting and robust. However, for more than ten years now, multi-level cell (MLC) technology has also established itself among customers in the industrial market segment. This is because the price difference to SLC is constantly decreasing and MLC-equipped SSDs can be built with higher capacities.

With multi-level cell (MLC) / triple-level cell (TLC) Flash, the ability to shrink reached its limit at 15 nanometer (nm) structures. The effort using firmware to correct errors in order to obtain and also maintain an unambiguous bit state of 0 or 1 was no longer justifiable in purely physical terms and no longer made sense.

SSDs equipped with 2D TLC NAND Flash were almost never a topic in the industrial market segment, because the write performance is too low and these memories were only suitable for booting, for example. This has changed considerably with the mass production of 3D NAND Flash.

In the Flash area, there has never been such a fast development to the next stage of the Flash cell. As soon as the 32-layer 3D Flash MLC/TLC was on the market, the 64-layer 3D TLC came, and today the 96-layer 3D TLC is standard. The first major customers have already received SSDs with 3D quad-logic cell (QLC).

QLC stores four bits per cell and thus twice the number of states compared to TLC technology with “only” eight states. The capacity per cell is thus doubled. The fact is that there has never been such a rapid advancement in the NAND Flash area as with the introduction of the 3D generation. The SLC NAND Flash dominated when it came to longevity and security. Then followed MLC NAND Flash, which was able to replace SLC in industrial applications in many segments. The main topics of discussion were write and erase cycles, and lifetime. Since the 3D NAND Flash is used in SSD, the pace of innovation has increased significantly. For a short time there was 3D MLC NAND, followed by 32-layer 3D TLC (BiCS2), and in 2018 SSDs with 64-layer 3D TLC (BiCS3) were standard. This year, SSDs with 96-layer 3D (BiCS4) will become more and more common, and probably by the middle of 2020 be replaced by QLC-based SSDs. In addition, we are already talking about 3D NAND Flash, which will make the leap from 96 layers to 128 layers in 2020 and go into series production in 2021. It therefore remains extremely exciting, and the development from 2D SLC to MLC and TLC was pretty boring in comparison.

2. THE PRINCIPLE OF AN SSD / FLASH CARD

The controller alone does not determine the speed of an SSD. An important aspect is the “intelligence” of the firmware, which decides where and how the data are written. How much effort was put into this and which algorithms were programmed and how? How big is the cache to move data, to package it and write it to the Flash storage area? There is so-called RAM-less SSD, where the manufacturer omits the cache for price reasons. This can be seen by a comparison between data sheets, because without cache the SSDs are somewhat slower. Here is a rough list of the firmware and its contents:

- wear leveling (dynamic and static)
- error detection code/error correction code (EDC ECC) handling
- write amplification factor (WAF)
- low power density check (LPDC)
- bad block management
- garbage collection (GC)
- over-provisioning (OP) handling
- end-to-end power handling

The internal structure of an SSD with the essential parts: controller with firmware, Flash storage and RAM (cache).
3. TRICKS OF THE MANUFACTURERS

The large number of manufacturers does not make it easy for the user to select the right SSD. There are now a huge number of various SSDs that are more or less suitable for different applications. It starts with the capacity: One manufacturer specifies the full Flash sizes, i.e. how much Flash is really installed, another gives only a part and implicitly takes the rest as reserve, also called over-provisioning (OP). OP is an additional Flash area used, among other things, by the garbage collection (GC) to speed up internal handling and extend the lifetime of the SSD.

Here is an example: Manufacturer A specifies 256 GB, manufacturer B 240 GB and manufacturer C 200 GB, although the SSD has a Flash area, i.e. storage area, of 256 GB. With three-dimensional (3D) Flash equipped SSDs, it is possible that the capacity actually usable by the user is somewhat smaller than the capacity specified. Among other things, one of the reasons for this is that an area of the Flash used is utilized for internal handling. Therefore, a practical test with almost real field conditions is always an advantage to actually analyze the performance and lifetime of an SSD.

4. SSD IN PRACTICE

In practice, a suitable SSD for appropriate use in the application must be selected. This gives the probability of achieving the longest possible lifetime for the SSD. In some applications, the focus might be placed on other factors, so that lifetime moves somewhat into the background.

4.1. APPLICATION

As already described, a test of the SSD under field conditions that will exist later is recommended. A rough distinction according to the application is made by manufacturers using the terms enterprise and client. Enterprise SSDs are to be seen more in the direction of data center and server application and come with a larger Flash area as a reserve in order to deliver a more stable performance over a longer period of time compared to client SSDs. Especially with storage arrays, this is important to maintain no or little latency degradation during peak loads. An additional distinction lies in the usage data: A client SSD must run eight hours/day at an average of 40°C, while 24/7 at an average of 55°C can be required from an enterprise SSD (JEDEC standard). Data retention in the switched-off state should be one year for a client SSD and three months for an enterprise SSD.

4.1.1. SSD with 2D and 3D Flash

For a long time, the two-dimensional (2D) NAND Flash cell (MLC and TLC) has determined the Flash memory market and contributed to the success of the SSD through its inexpensive manufacture, and scaling. However, this technology reached its physical limit with structures at around 15 nanometers (nm). Its successor is the three-dimensional (3D) NAND Flash, known as skyscraper.

2D NAND Flash, also called planar NAND, is constructed as a floating gate cell. With structures smaller than 15 nm, the number of bit errors increases and thus the effort regarding error compensation (ECC/EDC) because not enough electrons can be held in the floating gate cell. The space is simply too small and there are too many interferences.

In contrast to 2D NAND, where memory cells are stacked horizontally on cards, 3D NAND is stacked vertically using multiple layers. This results in higher density, better lifetime and faster read/write operations with lower power consumption. Production of 3D NAND Flash is costlier and more complex, but the cost per gigabyte (GB) is lower. Because it packs so many vertical cells into small width and length dimensions, 3D NAND has far more capacity than 2D NAND within the same length and width dimensions. 3D NAND are currently manufactured as TLC and are comparable in performance to 2D NAND built as MLC. QLC Flash is also produced as 3D NAND and can store four bits, i.e. 16 states (2x2x2x2).

3D NAND should not be confused with the so-called 3D XPoint technology from Intel/Micron, which is more likely to be based on resistive random access memory (ReRAM). ReRAM can store one bit by changing the electrical resistance of a conductive dielectric. The V-NAND Flash produced by Samsung is similar to the 3D NAND Flash described here.

4.1.2. Floating gate versus charge trap Flash cell

NAND Flash technology can be built from floating gate cells and charge trap cells. Almost all NAND Flash manufacturers such as Samsung, SK Hynix, WD and Toshiba now prefer to use charge trap cells for 3D NAND Flash. This technology has a higher storage density per chip area and thus larger capacities can be produced. The only manufacturer producing 3D on floating gate is probably Micron.

Unlike floating gate cell, which uses polycrystalline silicon for storage, charge trap cell uses a silicon nitride insulator. This means that the electrons can no longer jump from cell to cell or influence each other. This technology allows a thinner oxide layer and a lower programming voltage to be used for writing the cell, which in turn increases the speed of the read and write operations.
4.2. ENDURANCE

Endurance, respectively lifetime or write performance, is an important criterion when selecting an SSD. Incorrect selection of an SSD can result in subsequent costs that might arise. While endurance does not play a role when using an SSD as boot medium, a very high endurance is extremely important for an SSD that has taken over the task of a data logger, for example.

4.2.1. Terabytes written (TBW) / Drive writes per day (DWPD)

In SSD data sheets there is information about write performance under the heading endurance. This is specified in terabytes written (TBW) or drive writes per day (DWPD). The TBW value indicates the sum of the write performance in terabytes until the SSD has reached the end of its lifetime. An example: With a value of 100 TBW, 100 terabyte (100,000 gigabyte) of data can be written to the SSD. Depending on the application, this value can tend to change downwards. For example, many small data packets and quantities require much more SSD performance than larger data packets. The following is an example of TBW value from the data sheet of an SSD:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>256GB</th>
<th>512GB</th>
<th>1TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endurance</td>
<td>100 TBW</td>
<td>200 TBW</td>
<td>400 TBW</td>
</tr>
</tbody>
</table>

The DWPD value indicates how often the SSD can be written with the same amount of data, based on the capacity of the SSD, within three or five years each day until the end of its lifetime is reached. Example: DWPD = 1.9 / 5 for a 240 gigabyte (GB) SSD would mean that this 240 GB SSD can be written every day with three times the amount of data (i.e., 1.9 x 240 GB = 456 GB) for five years until the manufacturer’s warranty expires. In total this would be: 456 GB x 365 days x five years = 0.9 petabyte (PB). See the following example from an SSD data sheet:

<table>
<thead>
<tr>
<th>Capacities</th>
<th>Endurance(^1) (Petabyte written)</th>
<th>DWPD (for 3 yrs.)</th>
<th>DWPD (for 5 yrs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>240GB</td>
<td>0.9</td>
<td>3.4</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Decisive criteria for endurance in practice are the following:
- right choice of capacity
- evaluation of the endurance in the data sheet and comparison with an application; determination of performance of the SSD
- definition of operating temperature range of the SSD

Achievable lifetime of the SSD can be analyzed using appropriate tools. In a use case, the SSD should be operated similarly to how it will be used later. The self-monitoring, analysis and reporting technology (SMART) values can then be read out or the lifetime can be displayed as a traffic light function by means of a diagram.

For example, the lifetime can be reduced in an application where a lot of small data (<4 KB) are to be written. Here, it makes sense to pack the data first and then write it.

Due to the structure of an SSD, it can happen that the firmware has to move the incoming data several times until it finally finds its place in the Flash storage. This process is called write amplification factor (WAF). The higher this WAF is, the more the Flash cells wear out, and the lifetime declines rapidly.

Simply put, WAF can be expressed as follows:

WAF = \( \frac{\text{bytes written to Flash storage}}{\text{bytes written from host}} \)

By means of SMART values: \( \text{WAF} = 1 + \frac{\text{attribute 248}}{\text{attribute 247}} \)

Once the WAF has been determined and TBW for the corresponding capacity is taken from the data sheet, the approximate write performance of the SSD is then calculated: Write performance = TBW / WAF. In the example above and an assumed WAF of 3, this would mean: 100 TBW / 3 = 33 TBW, i.e. 33,000 gigabyte (GB). If one divides this value by five years and the number of days, then this would still result in a good value of approximately 18.5 GB as write performance per day. Nevertheless, this value is significantly lower than the one given in the data sheet.
4.2.2. Steady state or write saturation

If you only consider write and read performance mentioned in manufacturer’s data sheets, you may be disappointed in practice. The reason for this is that the SSD gradually slows down during operation. This aspect is called steady state or write saturation. It is, unfortunately, an entirely normal process due to the internal structure, the internal handling of the SSD firmware and the ambient temperature at which the SSD operates. If, for example, the manufacturer has not installed reserve capacity Flash or if the firmware is not optimally programmed, the write speed can settle at around 30 to 50% of the value given in the data sheet. This means, for example, that of the 450 MB/s specified, which the SSD will probably reach at the beginning, in practice it will drop to 250 to 300 MB/s..

4.2.3 Temperature behavior

If self-heating and high ambient temperatures are now added, then less than 100 MB/s may possibly be achieved. Some manufacturers have installed throttling so that the SSD does not further self-heat during operation.

There are SSDs that have additional temperature protection. At a temperature greater than or equal to 75°C, the protective mechanism throttles the SSD down and does let it run freely again until the temperature has dropped. Above 87°C, the controller blocks all write and read operations.

<table>
<thead>
<tr>
<th>SSD</th>
<th>256GB</th>
<th>Thermal band</th>
<th>read</th>
<th>write</th>
<th>note</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&gt;=75°</td>
<td>250MB/s</td>
<td>60MB/s</td>
<td>Thermal throttle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;=85°</td>
<td>150MB/s</td>
<td>20MB/s</td>
<td>Thermal throttle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;=87°</td>
<td></td>
<td></td>
<td>Emergency shutdown, controller blocked input/output</td>
</tr>
</tbody>
</table>

4.2.4 Housekeeping

Due to the structure of an SSD with NAND Flash and its structure as page and block, with new data a block must first be erased before it can be written to. This task is integrated in the firmware and is called garbage collection (GC). As a result of carrying out this internal task, a little rest is repeatedly needed by the SSD. The effect is apparent when recording the write speed.
4.3. **SOLDERABLE SSDS**

Embedded multimedia card (eMMC) ICs or Universal Flash Storage (UFS) ICs are solderable SSDs. Both components are structured similarly to an SSD, but integrated in a compact IC package. Solderable SSDs are mainly used in embedded applications, in the automotive industry or in smart home solutions.

The eMMC is currently offered in version 5.1 with high read/write speed and optimized firmware, among other things, to improve lifetime. Figure 4 shows the size comparison between eMMC and a 1-euro coin. Thanks to the compatibility of the packages, the individual ICs can be easily exchanged (e.g., to increase the capacity from 16 GB to 32 GB or 64 GB).

For many years, Avnet Integrated has been very successfully using eMMC ICs as Flash solution on its embedded modules. The additional advantage is multi-partitioning and programming as enhanced user data area (EUDA) SLC. This allows users to realize their own system in a very flexible way.

If the speed of the eMMC is too slow, UFSs are used, but they are not pin-compatible. The new UFS 3.0 ICs feature a data rate of 2,300 MB/s. In comparison, eMMC 5.1 delivers 400 MB/s. The UFS ICs come with 3D NAND Flash.

5. **POWER / POWER FAILURE**

Power is a topic of great interest for industrial applications. On the one hand, the question is how stable is the power supply and, on the other hand, the effect of a sudden failure of the supply voltage. The internal construction of an SSD and its programming with incoming data can cause certain problems, which can even lead to a loss of the data. Industrial and enterprise SSDs now have relatively good protection in an emergency and integrate a low-power detector on board. In the event of a voltage drop, the SSD controller no longer accepts any further command and tries to save the data that is currently being transferred between the controller, cache and Flash. When the power is restored, the controller attempts to safely restore the data and ensures that the SSD resumes its function (Figure 5).

Some manufacturers go a step further and place capacitors on the board of the SSD to maintain the internal voltage a little longer and thus to safely write data that is currently in the DRAM cache. In summary, it should be noted that a clean and stable power supply is very important for secure operation of an SSD, and can recover data.

6. **FORM FACTORS**

SSDs today come standard in 2.5-inch packages and are used wherever either an HDD has been used previously or you simply want to go for this flexible, easily interchangeable form factor. The 2.5-inch SSDs are hot-swappable. The M.2 form factor is available in several versions; the four-digit number indicating the width and length. The 2280 form factor with a width of 22 mm and a length of 80 mm is currently the favorite. In addition, there are 2230, 2242, 2260 and currently 22110 for larger capacities from 4 TB. M.2 with 2280 and 2242 are produced and used in large quantities. M.2 is not hot-swappable.

The 2.5-inch form factor will also be used in future applications. The situation is different with mini-SATA (mSATA), which was accepted by the market with a delay and is clearly heading in the direction of M.2 for new developments.
7. **INTERFACES**

The interface is the electrical and logical signaling between the SSD and the CPU. It defines the maximum bandwidth, minimum latency, expandability and hot-swap capability of the SSD. There are three basic interface options:

- Serial ATA (SATA)
- Serial Attached SCSI (SAS)
- Non-volatile memory express (NVMe)

Today, SATA is used as standard in almost all systems, but is reaching its limits when it comes to better performance or higher capacities. This is where SAS is used. If high speeds are required, NVMe comes into play. NVMe is based on PCI Express (PCIe) and is used in servers. With newer boards, this interface is used more and more in industrial and gaming applications.

NVMe protocol is the successor to advanced host controller interface (AHCI). So far, SSDs have speeds of about 500 MB/s. In comparison, SSDs with NVMe are two to three times faster because the SSD is connected directly via PCIe and the bottleneck is bypassed by SATA / SAS. The NVMe SSDs are available in two versions: Gen3 x2 and Gen3 x4. This stands for Generation 3 and either two or four lanes. With two lanes the read speed is theoretically 2 GB/s maximum and with four lanes 4 GB/s maximum.

8. **TOOLS**

Lifetime is an essential factor for an SSD used in industrial systems. There are numerous tools on the market – such as Crystaldiskinfo, HD Tune, HDD Health or Hard Disk Sentinel, which can be used – for a fee or free of charge – to predict lifetime of SSDs. The best choice is a tool provided by the SSD manufacturers that can almost always look deeper into the complex depths of an SSD with its interleaving of the Flash and programming by means of firmware.

9. **FUTURE OF THE SSD**

The SSD, no matter in which form factor, is experiencing a real boom and finding a place in an increasing number of innovative applications. SSDs in 2.5-inch form factor, which have now replaced smaller HDD capacities in almost all areas, play a large part in this regard. With the transition from 2D NAND to 3D NAND Flash, manufacturers are producing ever larger SSDs and Flash cards, and are in extreme competition with HDD. If, for example, a 16 GB or 32 GB capacity of the SSD is sufficient in the application, then this can only be found with third party manufacturers but no longer with the major Flash memory manufacturers. Newer SSD generations are only available with a capacity of 256 GB or more. But then, however, at a price that is close to the previous 128 GB SSD. Hence, this gives the user twice the capacity at almost the same price.

While in 2018, the 3D NAND SSD was mainly with 64-layer Flash, in 2019, the production of products with 96 layers will be increased. SSDs equipped with 3D QLC Flash will also be delivered. The SSD will also conquer new fields of application, be it in 2.5-inch or M.2 form factor, as Flash card or as eMMC or UFS.

With regard to the interface, a replacement of SATA and SAS in the direction of NVMe is in the pipeline. NVMe is a standard that is related to the PCIe bus and thus achieves a much higher bandwidth than SATA or SAS. New developments in 2.5-inch or M.2 form factors should already be going in the direction of NVMe.

Flash memory manufacturers are, of course, already developing the next technologies, such as phase-change memory (PCM), resistive random access memory (ReRAM) and magnetoresistive random access memory (MRAM)/spin transfer torque (STT). It remains exciting which technology will assert itself in the mass market. The new storage class memory (SCM) should close the gap between a fast DRAM and a non-volatile Flash. Although a DRAM is fast, it does not hold data after the supply voltage is removed. A Flash is non-volatile, however it is slow.
Solid state drives (SSDs) and the next generation 3D Flash

Author: Dipl. Ing. (FH) Robert Herth
Senior BDM Manager Storage, Avnet Integrated

Abbreviations:
solid state drive (SSD)
hard disk drive (HDD)
single-level cell (SLC)
multi-level cell (MLC)
triple-level cell (TLC)
quad-level cell (QLC)
error detection code/error correction code (EDC/ECC)
write amplification factor (WAF)
low power density check (LPDC)
garbage collection (GC)
over-provisioning (OP)
resistive random access memory (ReRAM)
terabytes written (TBW)
drive writes per day (DWPD)
self-monitoring, analysis and reporting technology (SMART)
embedded multimedia card (eMMC)
Universal Flash Storage (UFS)
enhanced user data area (EUDA)
mini-SATA (mSATA)
Serial ATA (SATA)
Serial Attached SCSI (SAS)
non-volatile memory express (NVMe)
PCI Express (PCIe)
advanced host controller interface (AHCI)
phase-change memory (PCM)
resistive random access memory (ReRAM)
magneto-resistive random access memory (MRAM)
spin transfer torque (STT)
storage class memory (SCM)
FOR WORLD CLASS INDUSTRIAL EMBEDDED SYSTEMS, SPEAK TO AVNET INTEGRATED. 
FIND OUT MORE: AVNET.COM/INTEGRATEDSYSTEMS