Deep Dive Series

The Gorilla Guide to...®
Express Edition

Innovations in Solid-State Storage Media and Systems

Inside this book:

- Discover how solid state media density is leading to a capacity revolution
- Learn why storage performance is no longer being held back by legacy technology
- Find out how storage innovations are leading to a renaissance in enterprise storage

Take A Quick Walk Through the IT Jungle!

Compliments of tegile

ActualTech Media

THE GORILLA GUIDE TO...®

Innovations in Solid-State Storage Media and Systems

EXPRESS EDITION

Scott D. Lowe
ActualTech Media

Compliments of

Printed in the United States of America

First Printing, 2017

ActualTech Media
Okatie Village Ste 103-157
Bluffton, SC 29909
www.actualtechmedia.com

Table of Contents

Introduction

Heading into late 2017 and early 2018, there are exciting happenings afoot in the world of flash storage. With constant innovation in the space, flash storage manufacturers and storage solution providers can continue to meet the growing demands of their customers for storage solutions that are efficient and cost-effective. This paper discusses the current and immediate future state of the flash storage market, with trends broken down into three core groups:

- Density
- Performance
- Cost

While not always considered in these specific terms, customers are continually on the lookout for storage solutions that can help them run increasingly large workloads with more demanding performance requirements at prices that make running such workloads on all-flash systems more feasible than in the past.

Density

The constant march forward in the world of technology has provided us with such laws as Moore's Law and Kryder's Law, which both predict an ongoing rate of advancement in technology. Kryder's Law was formulated in 2005 and initially applied to magnetic media, which, at the time, was far surpassing even Moore's wildest prediction with regard to areal density. Since then, both Moore's Law and Kryder's Law have been used to describe the scenario that is demonstrated by the intense march forward in flash innovation, with constantly improving methods being developed to place more, both in quantity and in density, flash cells in smaller spaces, resulting in modern flash disks that can store massive amounts of data in a 2.5" drive form factor.

There have been multiple methods by which flash makers have managed to increase information density:

- Increasing the amount of data that is able to be stored in individual flash cells; a development that has resulted in what we call Triple Level Cell (TLC) and Quad Level Cell (QLC) NAND types.

- Changing the overall structure of the flash storage medium itself by flipping it on its side and stacking cells.

Both developments will be discussed in the following sections.

Individual Cell Density

The first method of increasing information density involves changing the structure of the individual flash memory cells themselves which has produced a series of increasingly dense flash storage. The market started with Single Level Cell (SLC) NAND flash storage which then progressed to Multilevel Cell (MLC) storage. In MLC, you are able to store twice the amount of data as compared to SLC. While SLC is faster and eminently durable, the cost is very high on a cost per gigabyte basis. MLC systems managed to reduce this cost by half, but at the price of some performance and durability. However, while many lamented the potential for flash storage to wear out and die, thereby destroying company data, this never really happened. Flash has proven to be far more reliable than initially thought, and this reliability has pushed researchers to develop even more dense flash cells.

The Myth of Flash Death

For years, flash media analysts and designers talked about the long-term unreliability of flash media as being one of the key challenges that customers needed to worry about. The problem, they said, was that flash storage media simply could not stand up to disk over the long term, particularly in environments in which data changed at a rapid pace.

There have been different metrics developed to measure flash storage durability. Initially, the discussion evolved around the number of program/erase (PE) cycles that individual flash storage cells could withstand. As flash manufacturers implemented new techniques to spread out the data write impact across entire drives (wear leveling), PE cycles became difficult to measure. To that end, in most modern systems, you will see flash durability measured in drive writes per day (DWPD). You may see drives that can handle just one DWPD while others can withstand ten or more.

The DWPD metric takes into consideration wear leveling and other durability techniques and is also a bit easier for end users to grasp. They can figure out how much of their data changes each day and do some easy math to ensure that they get flash media that can withstand their intended usage patterns.

With all of this said, one reality in flash is that the predictions of constant flash death never really came to fruition. Sure, some disks ended up worn out, but there was hardly the disaster that was generally predicted and, today, flash media remains as reliable as ever.

It's a shame that early flash developers chose *multilevel cell* as the name of the next generation of flash beyond SLC. *Dual level cell* would have been perfect and created a bit less long-term confusion. You see, as time went on, flash manufacturers discovered that they could continue this density trend, and triple level cell (TLC) NAND flash became available on the market.

As the name implies, triple level cell flash media allows three bits of data to be stored in each cell rather than just the single bit you get with SLC. To better understand how this works, let's take a quick dive into understanding how data is stored in flash media. In short, in the world of SLC, a cell is either on or it's off – a value of 0 (programmed) or 1 (erased) – as determined by measuring voltage values in the media.

Figure 1. Voltage differential in SLC

With MLC, things must get a bit more nuanced as there are twice the bit values which include 00, 01, 10, and 11, with 00 being the fully programmed value and 11 being fully erased. 01 and 10 are considered to be

in partially programmed and partially erased states, respectively.

As you may have guessed, MLC adds some complexity to the equation since there is more sensitivity required to determine cell values. There is also a need to accommodate a phenomenon known as Stress Induced Leakage Current, which is a relatively rare, but potential issue with the way NAND flash stores data values. This is accommodated by using more comprehensive error-correcting code (ECC) mechanisms which keep watch over cell values and help systems recover when something goes awry. In addition, as the density of NAND cells increases, it takes longer to place individual cells through program and erase cycles, which further reduces the overall performance of particularly dense media.

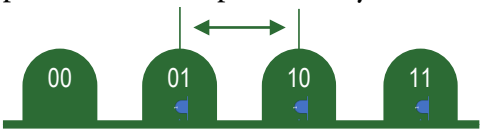

Figure 2. Voltage differential in MLC

With this additional overhead, MLC isn't quite as speedy as SLC, but it's still blazing fast compared to disk and has, for years, enjoyed its spot as a mainstream technology.

TLC NAND

MLC was a major advancement in the world of flash and quickly emerged as the most popular type of flash due to its much lower cost than SLC and its eventually-proven durability and reliability. However, as user demand continued its inexorable march toward more and bigger workloads, MLC is being augmented by even denser NAND flash media types that can hold more data.

One such advancement has been the creation of Triple Level Cell (TLC) flash media, which can store 50% more data than MLC while occupying the same amount of physical space. TLC accomplishes this feat by enabling even more granular voltage variations in individual data storage cells. TLC values include 000, 001, 010, 011, 100, 101, 110, and 111, with the specific values being read as a result of miniscule voltage differentials. Again, 000 is considered fully programmed and 111 is considered fully erased with the six values in the middle being various partially programmed and partially erased values. It takes more time for a read request to be fulfilled since the media needs to hone in on a specific voltage value, and, with

more from which to choose, this drill down process takes longer, with the end result being slower reads.

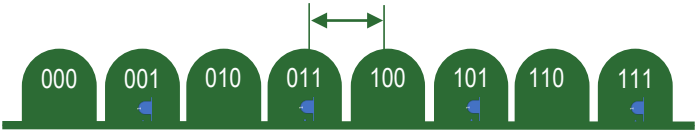

Figure 3. Voltage differential in TLC

Additionally, as was the case in the transition from SLC to MLC, MLC to TLC also results in less overall durability in addition to the performance hit. Again, though, it's still fast, but just not as fast as MLC. When it was first introduced, TLC flash was touted as being very useful in read-centric workloads. This makes sense since reads don't put flash media through the destructive program/erase process. Fewer programming cycles equate to a longer life.

On the durability front, in order to get to exacting voltage to dictate a cell value, electrical current needs to be applied to the media for longer periods. This is one reason for the speed decrease in TLC, and it also contributes to the shorter lifespan since the voltage application actively damages the underlying media. This damaging process is an unfortunate reality in all NAND flash, but is exacerbated by cramming more

data into the same space, as we've done by moving from SLC to MLC to TLC.

However, for workloads that don't place massive write stress on the media, TLC flash has become a popular way to further reduce storage costs while maintaining high levels of performance. But, even this isn't always enough.

QLC NAND

TLC wasn't the end of the line, though. Recently, Toshiba and Western Digital announced advancements in NAND flash manufacturing that leverage Quad Level Cell (QLC) technology, which stores four bits per cell, or double that of MLC. Potential values in such cells are 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, and 1111, and the voltage differential is miniscule even when compared with TLC, in which it was already quite small.

As you may have guessed, the performance values for QLC are worse than that of TLC due to the number of actual read operations that need to be completed in order for a cell's value to be derived. Further, QLC's

durability is far lower than that of even TLC, making QLC suitable primarily for read-centric operations or devices in which data does not change often. For those times when writes do need to take place, QLC's write speed will also be lower than what is experienced with TLC.

However, you should expect to see the use of QLC increase as manufacturers find ways to make it more reliable. Its density potential is double that of MLC, and the cost per gigabyte will be quite low compared to other technologies.

Cell Size

In addition to making it possible to store more data in individual NAND cells, flash manufacturers have also taken to decreasing the size of the individual cells, and this comes at the cost of reliability. As the cells shrink in size, they aren't able to last as long as larger cells.

There is a practical limit as to how small cells can be shrunk before it results in a product that is simply unusable. There comes a point at which intercell interference becomes so great that the product simply doesn't work.

However, manufacturers are incentivized to keep trying to get more storage from the same amount of physical space, so they keep trying. Moreover, every time a new advancement is announced, there are cries of "this is as far as it can go" and "this isn't reliable enough for the enterprise." In response to these cries, a funny thing happens: manufacturers make it better, getting it to a point where it *is*, in fact, reliable enough for the enterprise and storage solution providers to build application-centric solutions around new advancements.

However, physics is an indomitable force. It compels manufacturers to be on the constant lookout for new and innovative ways to increase capacity capabilities while also continuing to drive down the cost of flash.

3D/Vertical NAND (3D NAND or V-NAND or 3D V-NAND)

A few years ago, the first vertically stacked NAND-based flash storage products were introduced to the market; a form factor that is, today, becoming the standard. As manufacturers began to hit the real physics-imposed limits of cell size reduction, they needed a way to keep increasing overall flash density

without running into interference and other limits. At the same time, there was a desire to continue to shrink the physical size of disks shipped to customers and to use less material to further bring down the cost.

Over the past few years, manufacturers have adopted the real estate development adage that it's cheaper to build up than out. In real estate terms, it's an admission that land is more expensive than building material, so it's far less expensive to build a skyscraper than it would be to build the equivalent square footage just one story tall.

While it sounds very simple – building up rather than out – it's actually quite complex to make such a change. Just as you can't build a skyscraper with the same architectural method you use for a single-story building, you can't use the same techniques in 3D NAND as you for a single layer, or planar NAND. It's not just a matter of flipping NAND cells on their sides. You need to add new structures to enable 3D NAND to work.

At the most fundamental level, 3D NAND is exactly like 2D NAND in that electrons trapped in cells are used to determine that cell's value; but beyond that, it

looks quite different, with new materials added to enable the 3D nature and far more dense capacity. 3D NAND is described by the number of layers that it features, and these layers require the addition of some insulation. Specifically, material not capable of conducting electricity is placed between the layers to prevent interlayer interference from taking place. Within each layer, the media is still based on MLC, TLC, and QLC flash technology, but there is some interesting potential that is not possible with 2D. 3D NAND is often more reliable than its 2D counterpart, and it is often faster as well.

Today, we see 3D NAND with dozens of layers of capacity, with companies introducing newer products with an ever-increasing number of layers. Recently, the market saw the announcement of a 96-layer TLC-centric device that boasts massive density.

You should expect to see the continued transition from 2D to 3D NAND take place very quickly. Although 3D NAND is more complex to produce, the end result is a lower cost per gigabyte and less overall physical space required for storage.

Performance

Capacity in storage systems is obviously a critical item, but application owners are just as interested – and are often *more* interested – in the performance of storage systems. As companies continue to work with larger and larger data sets, and as they consume data from all kinds of sensors, and as the Internet-of-Things (IoT) continues to generate bulk data, it all has to go somewhere, and it all has to be analyzed; and all of this needs to happen *fast*. Real-time data analysis and data ingestion are the new black.

However, even with the raw speed of flash, it's been held back by 70s-era technology. Controllers have been based on technologies that have their roots in the introduction of the SCSI interface in 1979. This is where Non-Volatile Memory Express (NVMe) comes into play. More than just a physical interface, NVMe can unleash the full power of even the most powerful flash systems.

Non-Volatile Memory Express (NVMe)

Let's take a step back for a second. What is storage and what is RAM? Storage systems hold your persistent data and, on demand, load that data into RAM where

it is right next to the processor and operating on media that is orders of magnitude faster than even the fastest flash. In other words, the entire point of storage is to simply maintain data until it's needed by RAM and, upon request, get necessary data loaded into RAM as quickly as possible.

Flash gets data from storage into RAM far more quickly than is possible with spinning disks, but it's still hindered by middleware components that hold it back. Controllers, including SAS and SATA controllers, sip data from storage and place it into RAM. What if rather than sipping storage using a single straw, you could put 64,000 straws in a cup and pull data through them at full speed?

This is where the power of NVMe comes into play.

Flash storage is far different than disk in how data is read. With disks, you had to wait for a read head to get to just the right spot on a disk before you could read the data, so having a single operations queue made sense. Flash, though, has no such limitations. You can read from many locations at once. Imagine how much faster you can pull data from storage if you could read 64,000 locations at the same time. Inside

that lone SAS or SATA queue, you saw support for 256 or 32 commands, respectively. Each NVMe queue supports up to 64,000 commands. In theory, if you completely max out a system and fill every single one of the 64,000 queues with 64,000 commands, your storage system would be fulfilling 4,096,000,000 commands.

It's safe to say that NVMe is the market's answer to shedding the last vestiges of spinning disk-based storage support in favor of a solution that is made for modern solid-state storage systems.

NVMe leverages the PCI-e bus, which has far more capability than other buses in the system. In fact, the PCI-e bus has been used for flash storage for quite some time, but the traditional PCI-e connector hasn't lent itself to some of the administrative needs around storage. First, it's tough to hot swap a PCI-e card, and second, there are only a few PCI-e connectors in a server whereas storage devices often need a dozen or more.

In addition to more queues and using the PCI-e bus (which resides right next to the memory bus), NVMe also reduces the number of CPU instructions that are

required to satisfy a storage request which further increases the storage system's throughput potential, as measure in IOPS. This IOPS increase does not necessarily carry with it a corresponding increase in storage-centric CPU cycles, either. In an era in which Intel-based CPUs are powering the commodity hardware trend, this is incredibly valuable as it increases the overall workload density capability of servers and storage hosts.

SAS-3 currently tops out at 12 Gb/s, and the SAS-4 specification bumps this to 22.5 Gb/s, although currently there are no reasonable devices that support this standard. NVMe's reliance on the PCI-e bus is a limitation for throughput as well, but it scales to an aggregate of 16 Gb/s combined throughput across devices.

Beyond the Rack

The industry also recognizes the fact that storage isn't typically trapped on just a single host. Most organizations have multiple storage systems bound together with some kind of a storage transport protocol, such as iSCSI or Fibre Channel. However, these protocols are, by their nature, limited by legacy

command sets and operating systems still using SCSI internally to address such systems. With the emergence of NVMe-over-Fabrics, storage can move beyond rackscale to include thousands of NVMe devices. Through this effort, companies will be able to more easily solve both their capacity and performance challenges.

DIMM-based Flash/NVDIMMs

Why don't we just keep everything in memory and skip storage altogether? Well, if we could, we would. Memory is *far* faster than even the fastest solid state persistent media currently available. The problem, though, is one of volatility. In short, regular memory is considered volatile. The only time RAM can actually store data is when an electric current is being applied to the memory chips, which are typically delivered in a DIMM form factor. While a system is up and running, DIMMs are great - but as soon as you turn off the power, whatever data was on the DIMM is gone forever. Moreover, the cost of actual memory on a per-gigabyte basis far exceeds that of current flash media, making it unaffordable for mass storage.

There is a middle-ground, though. As mentioned earlier in this document, the goal of storage is to dump contents as quickly as possible to RAM. The location of storage matters when it comes to how fast this can happen. If data is stored on an array that is milliseconds away, it takes a lot of time to move a lot of data. What if, however, you could move data from one DIMM socket to another?

That's where Non-Volatile Dual In-line Memory Modules (NVDIMM) come into play. NVDIMM modules feature solid state storage in a DIMM form factor. Although these DIMMs aren't used as traditional memory, they do reside in DIMM sockets in servers equipped with a BIOS that can support storage in a DIMM socket.

At present, there are two types of NVDIMMs on the market:

- NVDIMM-F. The F variant is all flash storage with no actual memory on the board. This is basically an SSD in a DIMM form factor.
- NVDIMM-N. The N variant combines flash storage with actual RAM. Behind the scenes, the DIMM can flush the contents of the RAM portion of the device to the non-volatile flash segment in order to maintain the data.

NVDIMM-F devices appears to underlying operating systems as local block-based storage devices, so their use has no impact on ongoing operations. They can work with existing applications and processes. Because there is no need to work through a bus - even a fast bus like PCI-e - latency is significantly reduced, although, beyond that, performance characteristics are much like any other NAND storage device.

The actual storage on an NVDIMM device can be NAND flash, or it can be any other successor technology currently under development. There are a number of next-generation solid state storage mediums under development, each with the goal of further closing the performance gap between storage and DRAM.

Intel 3D XPoint/Optane

As the performance gap between DRAM and storage continues to shrink, people are referring to these emerging products as *memory class storage*. In a perfect world, a device would exist that has the performance characteristics of DRAM with the non-volatility characteristics of NAND - and all at the price of disk. Imagine a world in which boot times were a thing of

the past. You would just shut systems down and, upon power up, they'd be right back at the state they were in when you shut them down.

Imagine if you could store *all* your data on a medium that has the performance capabilities of DRAM. That's the point of storage class memory. One of the most talked-about entrants in next-generation solid state storage is Intel. The company, with the help of partners, has created a new class of storage dubbed '3D XPoint' which Intel is combining with new controllers and software and selling under the 'Optane' brand.

As the name implies, 3D XPoint has some physical resemblance to 3D NAND in that it doesn't exist solely in a 2D planar world. 3D XPoint is comprised of memory cells that are each coupled with a *selector*. By running different voltages through the selector via cross-stitched wires—which, by the way, also enable 3D XPoint's ability to address individual memory cells—the memory cells in this design are programmable by simply varying the voltage sent to its selector. This is all accomplished without the need for transistors, too.

Further, Optane is touted as having more overall durability than NAND flash as well. Intel's initial product offerings boast up to 30 drive-writes-per-day (DWPD), which is *far* more than we see from many existing NAND technologies. At 375 GB (Intel's P4800X device) and 30 DWPD, more than 12 petabytes of data can be written through the device each day. On the performance front, Intel shows latency of less than 10 microseconds and over 500,000 IOPS of throughput.

In a world in which data is being ingested on a more regular basis and from more devices and where data may be more transient, the increased durability, lower latency, and high throughput of Optane isn't something that organizations can ignore. Increasing use of Internet of Things (IoT) services will demand this kind of storage in order to be able to maintain the growth of the market. Customers will need storage that can read and write data at similar speeds and that, upon analysis, can be discarded without fear that the underlying media lacks the reliability to support constantly changing information.

Cost

Throughout this paper, we've talked about the need for manufacturers to constantly reduce the cost of their storage devices, which, in turn, allows array developers to sell more storage capacity to clients for less money. Again, as data sets get bigger, the ability to store that data in a way that doesn't break the bank becomes increasingly important too.

Hybrid All Flash Arrays

While cost and size are important aspects of data storage, performance is also critical. To that end, the era of hybrid storage began a few years ago. In the original generation of hybrid storage, storage array manufacturers coupled a bit of flash storage with a lot of spinning disk and combined that with powerful software that helped to determine what would run where. Through this combination, customers were able to get flash performance for critical workloads while maintaining disk-like economics for workloads where capacity was king.

Today, flash/disk hybrid storage is still alive and well, but all-flash storage has become the holy grail, particularly as the cost of flash has plummeted.

Moreover, as we get into things like NVMe, disk simply can't keep pace at all. As a result, we are seeing the emergence of all-flash hybrids.

This may sound like an oxymoron, but it's not! In all-flash hybrid systems, the fastest-of-the-fast flash is used as the performance tier while a slow type of flash is used for the capacity side. You may have enterprise grade MLC flash used for workloads that require massive performance and QLC-centric 3D NAND storage used for capacity. If you have varied workloads that demand different performance or capacity characteristics, all-flash hybrid storage may be the answer you've been looking for.

Data Services

Although things like deduplication and compression used to be optional, in the world of flash, they're core capabilities that mean the difference between life and death for flash-centric storage companies. Even though the cost of flash, in general, continues to drop, raw flash still does not equal raw disk when it comes to cost. And, as mentioned, there's always more data to store, so customers demand more storage that doesn't continue to cost them more money.

Data deduplication and compression efficacy are dependent on the underlying workloads. These features have been around for a while, but they continue to drive flash penetration forward, so they're included here as key capabilities for modern storage systems.

Summary

The discussion in this paper revolves around what is happening in the world of NAND-based flash today. There are all kinds of innovations taking place on a continuous basis, and we're sure that this paper will require updating on a regular basis to keep up. In fact, we hope it does! The constant development and innovation we're witnessing in the world of storage is moving the entire IT organization forward as workloads are able to enjoy the massive performance capabilities inherent in modern solutions and companies are able to enjoy the continuous cost reductions that result from these advancements.