A guide to flash storage and SSD fundamentals
Flash storage is the technology of the moment, providing high-performance random I/O capabilities far in excess of what can be achieved with mechanical hard drives.

But what is going on inside a flash drive? Why are writes much more troublesome than reads in flash? Why are flash drives' lifetimes limited? And what are flash storage makers doing to overcome these issues?

In this article, we look at exactly what flash storage is, how it is managed at controller level and some of the clever work that storage makers do to get the best out of solid state.

Flash deconstructed

When we talk about flash storage, we usually mean Nand flash, which is solid-state memory made of millions of Nand memory gates on a silicon die.

Flash technology recently reached its 30th birthday and manufacturers continue to push the boundaries of density on a single chip, which now extend into three dimensions with technology such as V-Nand from Samsung.

Flash is similar to system memory in that there are no moving parts, but it has the additional property that its contents are not lost when power is turned off.

Data is stored in cells, which gives us the terminology used to describe the main forms of flash, namely SLC, MLC and TLC.

SLC stands for single-level cell in which each memory cell records only a single value (of two states) – on or off, 0 or 1, based on the voltage of the cell. MLC, or multi-level cell, is capable of storing up to four states representing two bits of data – 00, 01, 10 or 11. TLC – triple-level cell – stores three bits in a single cell, using the eight states from 000 to 111.

Flash devices such as solid state disks (SSDs) are Nand chips packaged with additional circuitry and firmware known as a controller, which is responsible for managing the reading and writing process, as well as other ancillary tasks.

Flash reads and writes

Cells on flash storage are arranged into pages (typically 4KB or 8KB in size) and further grouped into blocks of around 128KB to 256KB, with some checksum data. The exact size depends on the flash manufacturer and product line.

The properties of Nand flash are such that a single value in a cell can be changed from “1” to “0” but not the other way around without reformatting the entire block, a process known as a program-erase (P/E) cycle.

As a result, writing data to flash in place requires the reading of an entire block from flash and into the memory of the controller, updating it with new data, erasing the existing block and writing the data back to the flash device. This
inefficient multi-stage process is known as write amplification, where each write operation to flash requires more than one physical write I/O.

Write amplification is a problem for flash devices because Nand chips are degraded slightly with every write operation and so devices have a finite number of P/E cycles. SLC Nand has a P/E cycle count of around 100,000 per block, but MLC can be as low as 5,000 per block of data.

The finite lifetime of flash means that writing data back in place repeatedly (for example, a file or database column rewritten multiple times) can very quickly result in a device failure. For this reason, flash drive manufacturers have employed techniques in the controller to mitigate the shortcomings of flash lifetime.

**Wear levelling**

Wear levelling is one technique that flash drive manufacturers use to improve device endurance or lifetime. Rather than storing data in the same place, wear levelling distributes write I/O blocks across the entire flash device, always writing to a new empty page. The result is more even wear across all Nand cells and increased device lifetime.

In addition to MTBF (mean time between failures), manufacturers also quote a figure known as DWPD (device/drive writes per day), which provides a measure of how many complete drive writes can be sustained over a fixed period (usually three to five years) before the device can be expected to fail.

DWPD figures vary greatly, from less than one to as high as 50, depending on whether the device is for the consumer or enterprise market. Naturally, devices with higher endurance attract a higher price.

**Value in the controller**

Controller circuitry and firmware performs the task of managing I/O back and forth from the Nand chips. Flash drive suppliers have invested significantly in optimising the firmware to work with Nand to deliver improved product lifetimes.

**Garbage collection**

As we have seen, flash device architectures store data in pages, which are grouped together in blocks for P/E cycles.

As techniques such as wear levelling distribute write I/O across an entire device, blocks start to fill up with pages of both in-use (or valid) and invalidated data that has been moved elsewhere in the device.

To re-use these invalidated pages, the entire block must be erased. A process called garbage collection manages the movement and consolidation of valid pages between blocks, allowing an entire block to be erased for subsequent re-use.

The effectiveness of the garbage collection process can have a direct effect on the performance of flash. When data is initially written to an SSD, the contents are placed on empty or partially filled blocks and very fast write times result.

But, at some point, the controller needs to start reclaiming pages for re-use and when this occurs, devices can experience a dip in performance, sometimes called the "write cliff".
The quality of algorithms used to perform garbage collection has a direct impact on performance – yet again demonstrating the importance of controller features.

**Cutting write times with Trim**

As we have seen, all issues with flash occur when writing to the device. So, if you can cut down on the processes involved in write I/O, it can improve device performance and lifetime.

One technique used to avoid write I/O is Trim. This allows the operating system (OS) to flag blocks of data that have been released from the local file system and to begin the erase process before the next write occurs.

Normally, reads and writes occur at page level, but deletes can only occur at the (larger) block level. In the normal write process, deletes occur at block level, but Trim allows the erase part of the P/E cycle to occur earlier.

Trim is supported by the major OSs and by the SCSI protocol as the Unmap command, which, in turn, is supported by the major hypervisor suppliers.

**Supplier implementations**

Flash devices have very different characteristics to hard drives. As a result, array suppliers have had to either develop new architectures designed around flash, or modify existing products to deal with flash drives.

Techniques include reading and writing in block sizes to match the drive being used, as seen in EMC XtremIO and HP’s 3PAR StoreServ systems.

Meanwhile, Hitachi Data Systems (HDS) designed its own flash module, which consolidates management functions into a custom controller, rather than using commodity SSDs. In a similar way, Violin Memory implements system-level wear levelling across all custom modules in its system, rather than on each drive.

Some of these flash benefits are implemented in hardware, but typically innovations are achieved through architectural design and software. This should come as no surprise, as increasingly we are seeing storage move towards a software-defined world.
Flash storage: Who makes what and where to deploy

This guide walks you through all the key areas of flash storage: all-flash, hybrid or PCIe; MLC vs SLC; speccing for performance and troubleshooting flash

Flash storage is the big news in storage arrays. Driven by the large and unpredictable workloads of server and desktop virtualisation, the low latency of flash storage has risen to the challenge.

But flash storage comes in many forms: in all-flash arrays, in hybrid arrays with spinning disk, as server-side PCIe flash. Then there is the alphabet soup associated with flash – MLC, SLC and now TLC.

So, which flash storage should you choose? What type of flash best suits different virtualisation and high-performance workloads? This ComputerWeekly.com guide walks you through all the key decision points of flash storage.

Flash storage articles

Big storage takes two routes in flash array shakedown
In the first of a series of surveys of the flash storage market, Computer Weekly looked at the responses of the big six suppliers to the flash revolution – to retrofit SSD to existing architectures or buy a startup.

Flash array market roundup: The startups
The startups led the flash array fray from the start and still did when this was written, despite the efforts of the big six. Here we survey the work of the startup vanguard in the flash revolution.

Big storage turns the tide in the hybrid flash array market
The hybrid flash array market – as with most flash storage – has been all about the startups, but did the tide turn with EMC's re-architecting of its VNX line for flash?

PCIe SSD flash storage roundup: The few become fewer
The PCIe flash storage, or server-side flash, market is one dominated by a few, in which the big six array makers don’t get a look-in. Computer Weekly surveyed the speeds and feeds on PCIe flash.

Flash caching software market roundup
Flash storage cuts latency and boosts IOPS, but cannot do the job alone; flash caching software is needed to target the blocks that need it most. Computer Weekly surveyed the key suppliers in this space.

Flash storage fundamentals

MLC vs SLC: Which flash SSD is right for you?
MLC vs SLC: Which flash SSD you choose depends on the performance you need and the price you want to pay, but the differences are not as great as you may think.

Podcast: MLC, eMLC, SLC, TLC – what they are and are good for
The flash storage market is an alphabet soup of three-letter acronyms. Here we run the rule over the type of flash available and what they are good for.
When to use flash SSD instead of SAS or SATA?
When to use flash SSDs instead of SAS or SATA spinning disk is all down to application requirements, says David Boyle, senior consultant at GlassHouse Technologies (UK).

Flash storage performance

Sizing VDI storage for I/O performance in flash arrays
Flash arrays are increasingly being used for VDI storage. But how do customers size flash storage for the I/O requirements of virtual desktops?

Accelerate the performance of virtual machines with flash technology
Learn where to implement flash in your environment to tackle virtual machine performance issues.

You’ve got flash storage, so why isn’t everything faster?
To get the most out of flash storage, you need storage performance monitoring of the whole stack, from servers to LUNs.

Flash storage: The future

What’s wrong with flash storage: And what will come after?
Despite flash storage taking the enterprise by storm, its days are numbered because of an unfavourable combination of technological obstacles and manufacturing economics.