; (HTTPS://WWW.EXTREMETECH.COM/CATEGORY/COMPUTING)

How Do SSDs Work?

By Joel Hruska (https://www.extremetech.com/author/jhruska) on May 3, 2021 at 8:30 am

Comments (https://www.extremetech.com/extreme/210492-extremetech-explains-how-do-ssds-work#comments)

Here at ExtremeTech, we've often discussed the difference between different types of NAND structures — vertical NAND (http://www.extremetech.com/computing/195536-samsung-850-evo-ssd-high-performance-durable-3d-nand-hits-the-mainstream) versus planar, or multi-level cell (MLC) versus triple-level cells (http://www.extremetech.com/computing/103873-ocz-to-launch-lower-cost-triple-cell-ssd-despite-endurance-performance-trade-offs) (TLC) and quad-level cells (https://www.extremetech.com/computing/274966-intel-launches-first-consumer-qlc-nand-ssd) (QLC). Now, let's talk about the more basic relevant question: How do SSDs work in the first place, and how do they compare with newer technologies, like Intel's non-volatile storage technology, Optane (https://www.extremetech.com/tag/optane)?

To understand how and why SSDs (https://zdcs.link/Gb7jM) are different from spinning discs, we need to talk a little bit about hard drives. A hard drive stores data on a series of spinning magnetic disks called platters. There's an actuator arm with read/write heads attached to it. This arm positions the read-write heads over the correct area of the drive to read or write information.

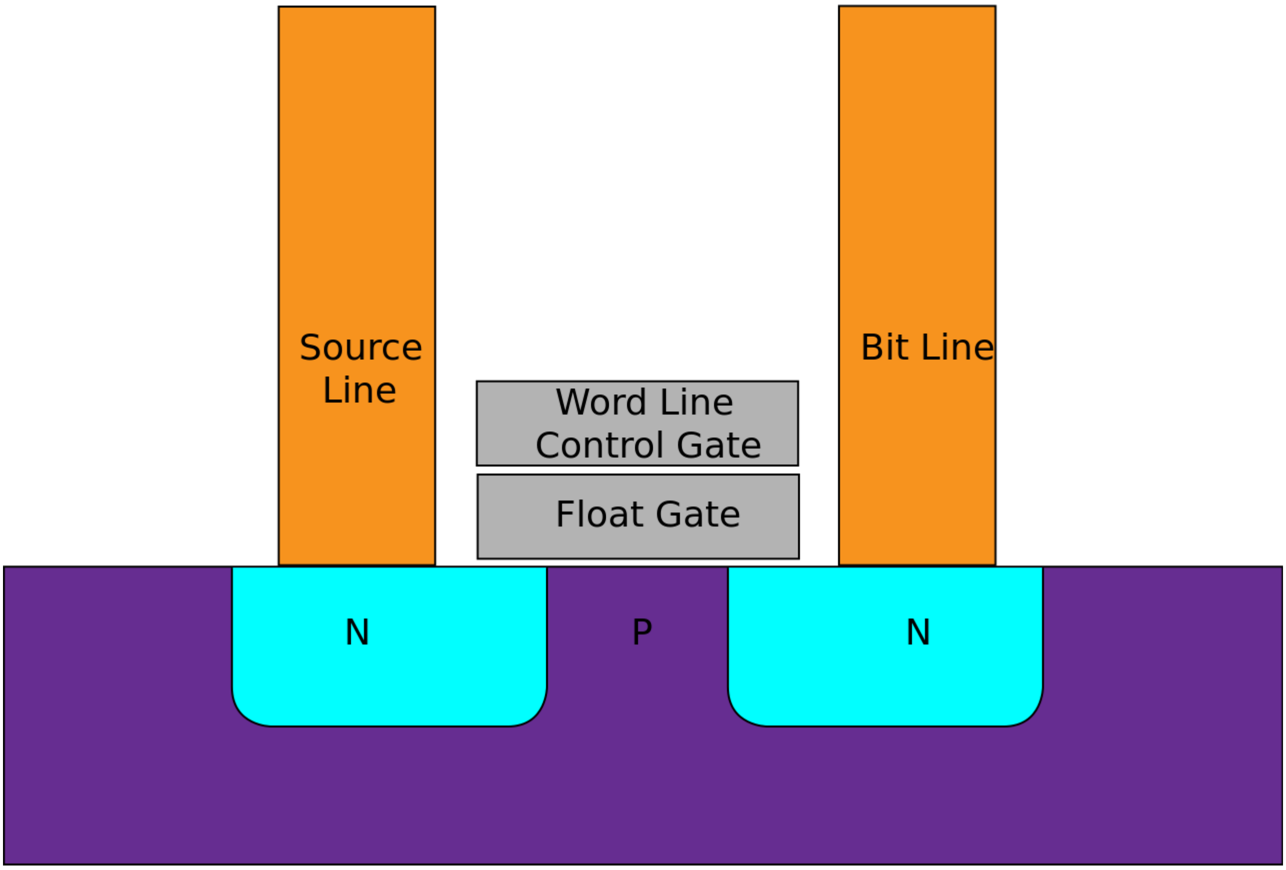
Because the drive heads must align over an area of the disk in order to read or write data, and the disk is constantly spinning, there's a delay before data can be accessed. The drive may need to read from multiple locations in order to launch a program or load a file, which means it may have to wait for the platters to spin into the proper position multiple times before it can complete the command. If a drive is asleep or in a low-power state, it can take several seconds more for the disk to spin up to full power and begin operating.

From the very beginning, it was clear that hard drives couldn't possibly match the speeds at which CPUs could operate. Latency in HDDs is measured in milliseconds, compared with nanoseconds for your typical CPU. One millisecond is 1,000,000 nanoseconds, and it typically takes a hard drive 10-15 milliseconds to find data on the drive and begin reading it. The hard drive industry introduced smaller platters, on-disk memory caches, and faster spindle speeds to counteract this trend, but there's only so fast drives can spin. Western Digital's 10,000 RPM VelociRaptor family is the fastest set of drives ever built for the consumer market, while some enterprise drives spun as quickly as 15,000 RPM. The problem is, even the fastest spinning drive with the largest caches and smallest platters are still achingly slow as far as your CPU is concerned.

How SSDs Are Different

"If I had asked people what they wanted, they would have said faster horses." — Henry Ford

Solid-state drives are called that specifically because they don't rely on moving parts or spinning disks. Instead, data is saved to a pool of NAND flash. NAND itself is made up of what are called floating gate transistors. Unlike the transistor designs used in DRAM, which must be refreshed multiple times per second, NAND flash is designed to retain its charge state even when not powered up. This makes NAND a type of non-volatile memory.



(http://www.extremetech.com/wp-content/uploads/2015/07/Flash_cell_structure.svg_1.png)

Image by Cyferz at Wikipedia (https://commons.wikimedia.org/wiki/File:Flash_cell_structure.svg), Creative Commons Attribution-Share Alike 3.0.

The diagram above shows a simple flash cell design. Electrons are stored in the floating gate, which then reads as charged "0" or not-charged "1." Yes, in NAND flash, a 0 means data is stored in a cell — it's the opposite of how we typically think of a zero or one. NAND flash is organized in a grid. The entire grid layout is referred to as a block, while the individual rows that make up the grid are called a page. Common page sizes are 2K, 4K, 8K, or 16K, with 128 to 256 pages per block. Block size therefore typically varies between 256KB and 4MB.

One advantage of this system should be immediately obvious. Because SSDs have no moving parts, they can operate at speeds far above those of a typical HDD. The following chart shows the access latency for typical storage mediums given in microseconds.

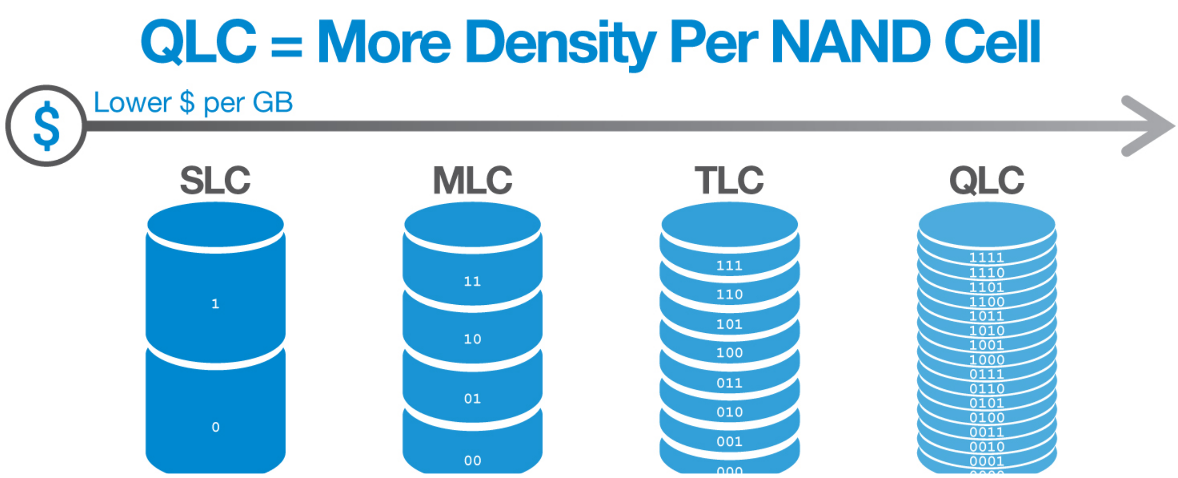
	SLC	MLC	TLC	HDD	RAM
P/E cycles	100k	10k	5k	*	*
Bits per cell	1	2	3	*	*
Seek latency (μs)	*	*	*	9000	*
Read latency (μs)	25	50	100	2000-7000	0.04-0.1
Write latency (μs)	250	900	1500	2000-7000	0.04-0.1
Erase latency (μs)	1500	3000	5000	*	*
Notes	* metric is not applicable for that type of memory				
Sources	P/E cycles [20] SLC/MLC latencies [1] TLC latencies [23] Hard disk drive latencies [18, 19, 25] RAM latencies [30, 52] L1 and L2 cache latencies [52]				

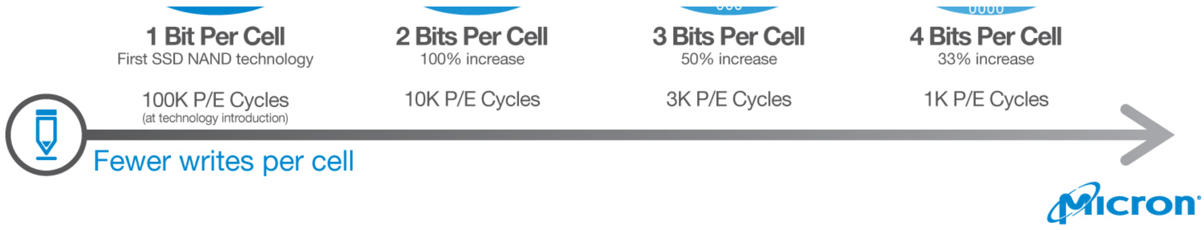
(http://www.extremetech.com/wp-content/uploads/2015/07/SSD-Latency.png)

Image by CodeCapsule (http://codecapsule.com/2014/02/12/coding-for-ssds-part-2-architecture-of-an-ssd-and-benchmarking/)

NAND is nowhere near as fast as main memory, but it's multiple orders of magnitude faster than a hard drive. While write latencies are significantly slower for NAND flash than read latencies, they still outstrip traditional spinning media.

There are two things to notice in the above chart. First, note how adding more bits per cell of NAND has a significant impact on the memory's performance. It's worse for writes as opposed to reads — typical triple-level-cell (TLC) latency is 4x worse compared with single-level cell (SLC) NAND for reads, but 6x worse for writes. Erase latencies are also significantly impacted. The impact isn't proportional, either — TLC NAND is nearly twice as slow as MLC NAND, despite holding just 50% more data (three bits per cell, instead of two). This is also true for QLC drives, which store even more bits at varying voltage levels within the same cell.





(https://www.extremetech.com/wp-content/uploads/2018/08/qlc_nand_density.jpg)

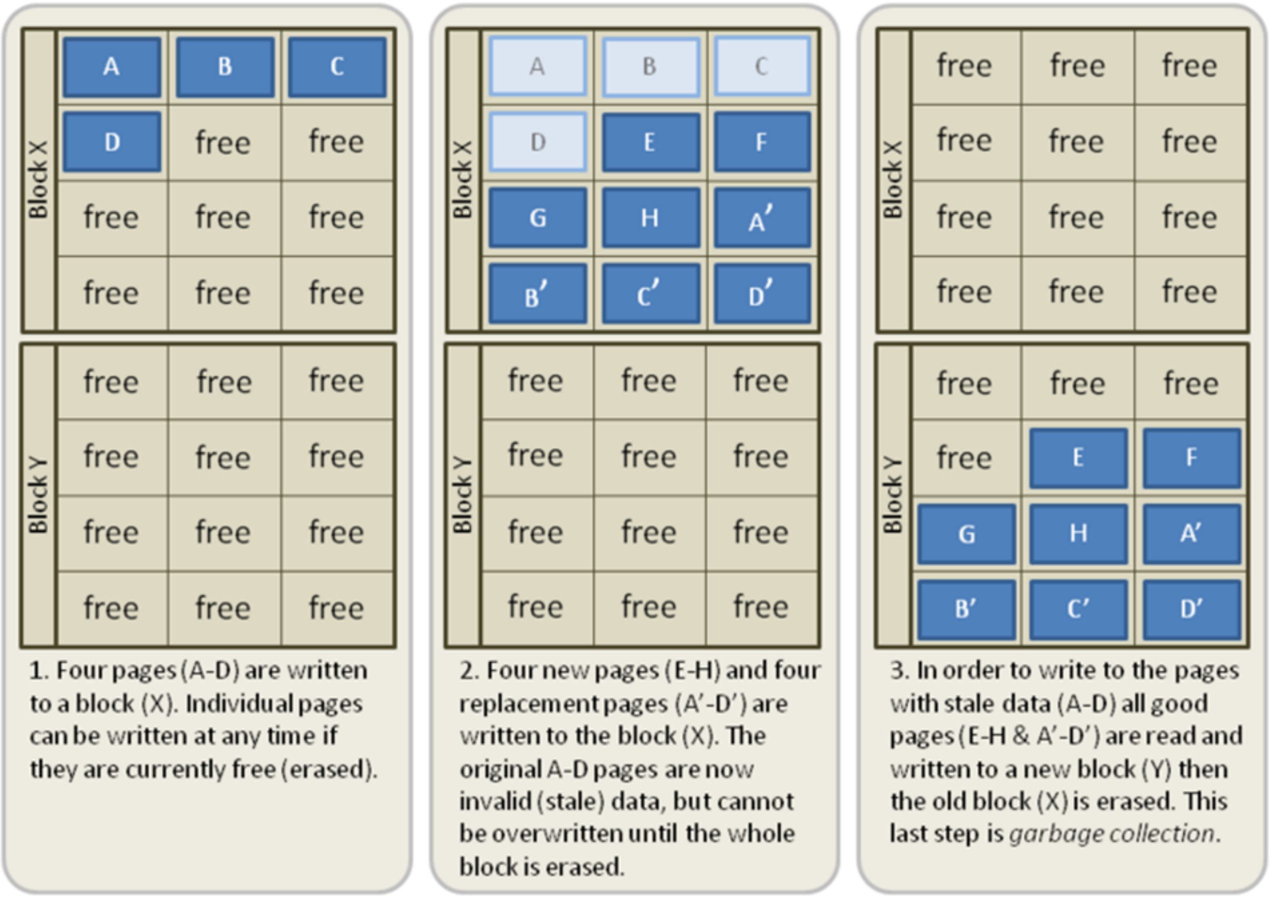
The reason TLC NAND is slower than MLC or SLC has to do with how data moves in and out of the NAND cell. With SLC NAND, the controller only needs to know if the bit is a 0 or a 1. With MLC NAND, the cell may have four values — 00, 01, 10, or 11. With TLC NAND, the cell can have eight values, and QLC has 16. Reading the proper value out of the cell requires the memory controller to use a precise voltage to ascertain whether any particular cell is charged.

Reads, Writes, and Erasure

One of the functional limitations of SSDs (http://www.extremetech.com/tag/ssds) is while they can read and write data very quickly *to an empty drive*, overwriting data is much slower. This is because while SSDs read data at the page level (meaning from individual rows within the NAND memory grid) and can write at the page level, assuming surrounding cells are empty, they can only erase data at the block level. This is because the act of erasing NAND flash requires a high amount of voltage. While you can theoretically erase NAND at the page level, the amount of voltage required stresses the individual cells around the cells that are being re-written. Erasing data at the block level helps mitigate this problem.

The only way for an SSD to update an existing page is to copy the contents of the entire block into memory, erase the block, and then write the contents of the old block + the updated page. If the drive is full and there are no empty pages available, the SSD must first scan for blocks that are marked for deletion but that haven't been deleted yet, erase them, and then write the data to the now-erased page. This is why SSDs can become slower as they age — a mostly-empty drive is full of blocks that can be written immediately, a mostly-full drive is more likely to be forced through the entire program/erase sequence.

If you've used SSDs, you've likely heard of something called "garbage collection." Garbage collection is a background process that allows a drive to mitigate the performance impact of the program/erase cycle by performing certain tasks in the background. The following image steps through the garbage collection process.



(http://www.extremetech.com/wp-content/uploads/2015/07/Diagram-1.png)

Image courtesy of Wikipedia

Note in this example, the drive has taken advantage of the fact that it can write very quickly to empty pages by writing new values for the first four blocks (A'-D'). It's also written two new blocks, E and H. Blocks A-D are now marked as stale, meaning they contain information the drive has marked as out-of-date. During an idle period, the SSD will move the fresh pages over to a new block, erase the old block, and mark it as free space. This means the next time the SSD needs to perform a write, it can write directly to the now-empty Block X, rather than performing the program/erase cycle.

The next concept I want to discuss is TRIM. When you delete a file from Windows on a typical hard drive, the file isn't deleted immediately. Instead, the operating system tells the hard drive it can overwrite the physical area of the disk where that data was stored the next time it needs to perform a write. This is why it's possible to undelete files (and why deleting files in Windows doesn't typically clear much physical disk space until you empty the recycling bin). With a traditional HDD, the OS doesn't need to pay attention to where data is being written or what the relative state of the blocks or pages is. With an SSD, this matters.

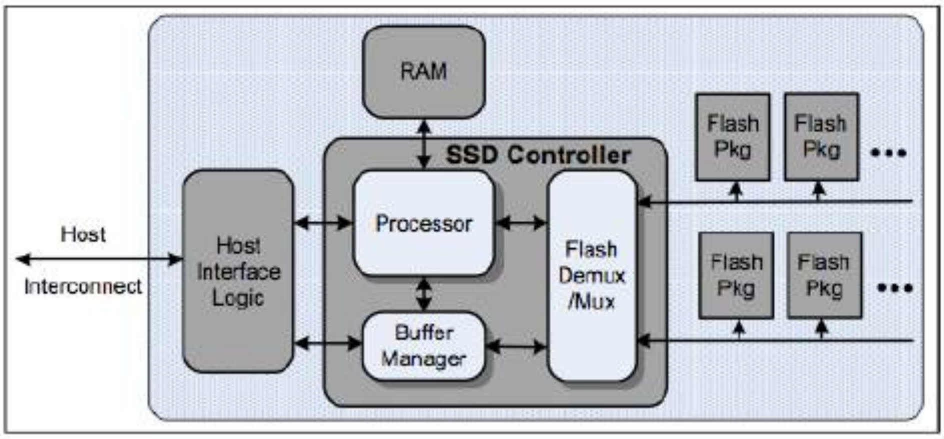
The TRIM command allows the operating system to tell the SSD it can skip rewriting certain data the next time it performs a block erase. This lowers the total amount of data the drive writes and increases SSD longevity. Both reads and writes damage NAND flash, but writes do far more damage than reads. Fortunately, block-level longevity has not proven to be an issue in modern NAND flash. More data on SSD longevity (http://www.extremetech.com/computing/201064-which-ssds-are-the-most-reliable-massive-study-sheds-some-light), courtesy of the Tech Report, can be found here.

The last two concepts we want to talk about are wear leveling and write amplification. Because SSDs write data to pages but erase data in blocks, the amount of data being written to the drive is always larger than the actual update. If you make a change to a 4KB file, for example, the entire block that 4K file sits within must be updated and rewritten. Depending on the number of pages per block and the size of the pages, you might end up writing 4MB worth of data to update a 4KB file. Garbage collection reduces the impact of write amplification, as does the TRIM command. Keeping a significant chunk of the drive free and/or manufacturer over-provisioning can also reduce the impact of write amplification.

Wear leveling refers to the practice of ensuring certain NAND blocks aren't written and erased more often than others. While wear leveling increases a drive's life expectancy and endurance by writing to the NAND equally, it can actually increase write amplification. In other to distribute writes evenly across the disk, it's sometimes necessary to program and erase blocks even though their contents haven't actually changed. A good wear leveling algorithm seeks to balance these impacts.

The SSD Controller

It should be obvious by now SSDs require much more sophisticated control mechanisms than hard drives do. That's not to diss magnetic media — I actually think HDDs deserve more respect than they are given. The mechanical challenges involved in balancing multiple read-write heads nanometers above platters that spin at 5,400 to 10,000 RPM are nothing to sneeze at. The fact that HDDs perform this challenge while pioneering new methods of recording to magnetic media and eventually wind up selling drives at 3-5 cents per gigabyte is simply incredible.



SSD Logic Components

(http://www.extremetech.com/wp-content/uploads/2015/07/2006640.jpg)

A typical SSD controller

SSD *controllers*, however, are in a class by themselves. They often have a DDR3 or DDR4 memory pool to help with managing the NAND itself. Many drives also incorporate single-level cell caches that act as buffers, increasing drive performance by dedicating fast NAND to read/write cycles. Because the NAND flash in an SSD is typically connected to the controller through a series of parallel memory channels, you can think of the drive controller as performing some of the same load-balancing work as a high-end storage array — SSDs don't deploy RAID internally but wear leveling, garbage collection, and SLC cache management all have parallels in the big iron world.

Some drives also use data compression algorithms to reduce the total number of writes and improve the drive's lifespan. The SSD controller handles error correction, and the algorithms that control for single-bit errors have become increasingly complex as time has passed.

Unfortunately, we can't go into too much detail on SSD controllers because companies lock down their various secret sauces. Much of NAND flash's performance is determined by the underlying controller, and companies aren't willing to lift the lid too far on how they do what they do, lest they hand a competitor an advantage.

Interfaces

In the beginning, SSDs used SATA ports, just like hard drives. In recent years, we've seen a shift to M.2 drives — very thin drives, several inches long, that slot directly into the motherboard (or, in a few cases, into a mounting bracket on a PCIe riser card. A Samsung 970 EVO Plus drive is shown below.

(https://www.extremetech.com/wp-content/uploads/2019/01/540134-samsung-ssd-970-evo-plus.jpg)

NVMe drives offer higher performance than traditional SATA drivers because they support a faster interface. Conventional SSDs attached via SATA top out at ~550MB/s in terms of practical read/write speeds. M.2 drives are capable of substantially faster performance into the 3.2GB/s range.

The Road Ahead

NAND flash offers an enormous improvement over hard drives, but it isn't without its own drawbacks and challenges. Drive capacities and price-per-gigabyte are expected to continue to rise and fall respectively, but there's little chance SSDs will catch hard drives in price-per-gigabyte. Shrinking process nodes are a significant challenge for NAND flash — while most hardware improves as the node shrinks, NAND becomes more fragile. Data retention times and write performance are intrinsically lower for 20nm NAND than 40nm NAND, even if data density and total capacity are vastly improved. Thus far, we've seen drives with up to 96 layers in-market, and 128 layers seems plausible at this point. Overall, the shift to 3D NAND has helped improve density without shrinking process nodes or relying on planar scaling.

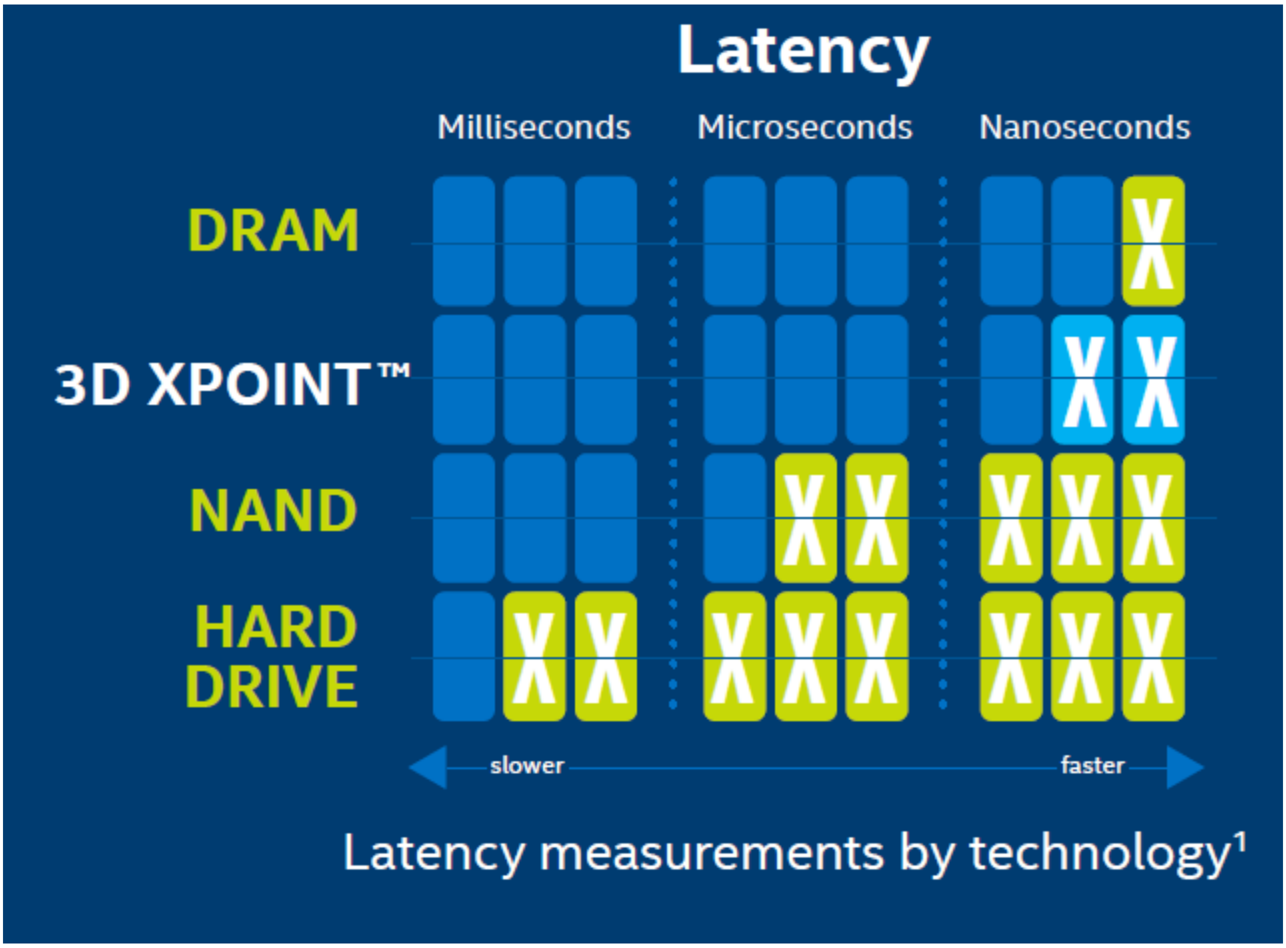
Thus far, SSD manufacturers have delivered better performance by offering faster data standards, more bandwidth, and more channels per controller — plus the use of SLC caches we mentioned earlier. Nonetheless, in the long run, it's assumed NAND will be replaced by something else.

What that something else will look like is still open for debate. Both magnetic RAM (http://www.extremetech.com/computing/193065-mram-manufacturer-everspin-teams-up-with-globalfoundries-to-build-magnetic-memory) and phase change memory (http://www.extremetech.com/extreme/182096-ibm-demonstrates-next-gen-phase-change-memory-thats-up-to-275-times-faster-than-your-ssd) have presented themselves as candidates, though both technologies are still in early stages and must overcome significant challenges to actually compete as a replacement to NAND. Whether consumers would notice the difference is an open question. If you've upgraded from an HDD to an SSD and then upgraded to a faster SSD, you're likely aware the gap between HDDs and SSDs is much larger than the SSD-to-SSD gap, even when upgrading from a relatively modest drive. Improving access times from milliseconds to microseconds matters a great deal, but improving them from microseconds to nanoseconds might fall below what humans can really perceive in most cases.

Optane Retrenches in the Enterprise Market

From 2017 through early 2021, Intel offered its Optane memory as an alternative for NAND flash in the consumer market. In early 2021, the company announced it would no longer sell Optane drives in the consumer space, (https://www.extremetech.com/computing/319228-intel-kills-its-consumer-facing-optane-products) except for the H20 hybrid drive. H20 combines QLC NAND with an Optane cache to boost overall performance while reducing drive cost. While the H20 is an interesting and unique product, it doesn't offer the same kind of top-end performance Optane SSDs did.

Optane will remain in-market in the enterprise server segment. While its reach is limited, it's still the closest thing to a challenger that NAND has. Optane SSDs don't use NAND — they're built using non-volatile memory believed to be implemented similarly to phase-change RAM — but they offer similar sequential performance to current NAND flash drives, albeit with better performance at low drive queues. Drive latency is also roughly half of NAND flash (10 microseconds, versus 20) and vastly higher endurance (30 full drive-writes per day, compared with 10 full drive writes per day for a high-end Intel SSD).



(https://www.extremetech.com/wp-content/uploads/2016/10/Optane1.png)

Intel Optane performance targets

Optane is available in several drive formats and in as a direct replacement for DRAM. Some of Intel's high-end Xeon CPUs support multi-terabyte Optane deployments and support a mix of DRAM and Optane that provides a server with much more RAM than DRAM alone could, at the cost of higher access latencies.

One reason Optane has had trouble breaking through in the consumer space is that NAND prices fell dramatically in 2019 and stayed low through 2020, making it difficult for Intel to effectively compete.

Check out our ExtremeTech Explains (http://www.extremetech.com/tag/extremetech-explains) series for more in-depth coverage of today's hottest tech topics.

Now Read:

- The Worst CPUs Ever Built (https://www.extremetech.com/computing/274650-the-worst-cpus-ever-made)
- Happy 40th Anniversary to the Original Intel 8086 (https://www.extremetech.com/computing/270926-happy-40th-anniversary-to-the-original-8086-and-the-x86-architecture)
- The Myths of Moore's Law (https://www.extremetech.com/extreme/223022-the-myths-of-moores-law)

Nvidia Reportedly Prepping Two New RTX 3050s, One with Just 4GB of VRAM

JOSH NOREM

Read More

You Might Also Like

Powered By ZergNet (https://www.zergnet.com/privacy-widget)

Tagged In

storage (https://www.extremetech.com/tag/storage)
extremetech explains (https://www.extremetech.com/tag/extremetech-explains)
SSDs (https://www.extremetech.com/tag/ssds) SSD (https://www.extremetech.com/tag/ssd)
NAND (https://www.extremetech.com/tag/nand) hard drives (https://www.extremetech.com/tag/hard-drives)
NAND flash (https://www.extremetech.com/tag/nand-flash)
solid state storage (https://www.extremetech.com/tag/solid-state-storage)

Popular in the Community

Conversation 51 Comments

We are currently updating our comments feature - during this transition, older comments may not be visible, but will return in the coming days. Thank you for your patience!

Commenting as **Guest** Connect with disqus Log in Sign up

What do you think?

Sort by **Best**

Dromo • 8 February, 2019

mistake here
". If you've upgraded from NAND to an SSD and then upgraded to a faster SSD,"
Upgraded from a HDD ??

Reply 7

1 reply

Hikari . • 12 February, 2019

[quote]When you delete a file from Windows on a typical hard drive, the file isn't deleted immediately. Instead, the operating system tells the hard drive it can overwrite the physical area of the disk where that data was stored the next time it needs to perform a write. This is why it's possible t...**See more**

Reply 8 1

3 replies

Daniel WhatThe Heck • 9 December, 2019

I think its important to note that M.2 is a form factor and connector, not the actual interface standard. M.2 supports SATA, USB, and NVMe. There are plenty of SATA M.2 cards out there that will disappoint the uninformed consumer expecting to get the speeds of an NVMe drive.

Reply 4

1 reply

trparky • 6 February, 2019

The only problem with Intel's 3D XPoint (Intel Optane) is that it's expensive. Sure, it's fast; I have no doubt. But to get anything close to a decently sized drive you practically have to sell your kidney.

NAND may be slightly slower but it wins in the capacity and price war when you compare it to ...**See more**

Reply 3

3 replies

JackNaylorPE • 26 April, 2019

Id like to hear more about the real world benefits of SSDs. Running an Office Suite script that finishes a series of 132 tasks twice as fast is meaningless when each of those 132 tasks requires a human to key in several keystrokes for each one. If ya look at SSD reviews and the tests performed m...**See more**

Reply 3

8 replies

IraMinor • 29 July, 2019

Another advantage of 3D-Xpoint is that it is byte addressable, no block stuff to deal with. My next SSD will be 3D_Xpoint, M.2, PCIe 4 or 5. In Windows I eliminated the paging and hibernate files. Never turn off the SSD or go to sleep. I shut the system down at night because booting it back up is s...**See more**

Reply 2

Mete Can Karahasan • 11 February, 2019

@Joel Hruska:
Thank you for citing Tech Report on their crazy long study. I had an actual friend who was eager to review ssds. One look at their findings and I knew how high the stakes could be.

Reply 1

Ron Dunbar • 6 October, 2020

Windows 10 is soon to have a feature that informs you that a PC's SSD will soon need last rites. It would be interesting to hear just how the OS makes that determination. Don't get me wrong - I think it's a great feature, considering that SSDs do wear out. I simply have an academic interest in how ...**See more**

Reply 1

1 reply

G **Guest** • 9 December, 2019

This comment was deleted.

Reply

1 reply

RH • 16 April, 2020

I went from HDD to SSD on my home & laptops a few years ago. My newest laptop I bought last fall, had Nvm2 and WOW! That sucker is REALLY fast. Nice to be able to shut everything down and boot back up in seconds. And, the load time on photoshop is REALLY quick. On the home PC, still using an S...**See more**

Reply

Show More Comments

Powered by OpenWeb

Terms | Privacy | Feedback

ExtremeTech Newsletter

Subscribe Today to get the latest ExtremeTech news delivered right to your inbox.

Email Address... Sign Up

This newsletter may contain advertising, deals, or affiliate links. Subscribing to a newsletter indicates your consent to our Terms of Use (http://www.ziffdavis.com/terms-of-use) and Privacy Policy (http://www.ziffdavis.com/privacy-policy). You may unsubscribe from the newsletter at any time.

More Articles

[(https://www.extremetech.com/mobile/331686-samsung-pauses-galaxy-tab-s8-pre-orders-citing-unexpected-demand)](https://www.extremetech.com/mobile/331686-samsung-pauses-galaxy-tab-s8-pre-orders-citing-unexpected-demand)
Samsung Pauses Galaxy Tab S8 Pre-Orders Citing Unexpected Demand (https://www.extremetech.com/mobile/331686-samsung-pauses-galaxy-tab-s8-pre-orders-citing-unexpected-demand) Feb 15

(https://www.extremetech.com/internet/331675-clever-malware-masquerades-as-windows-11-installer)
Clever Malware Masquerades as Windows 11 Installer (https://www.extremetech.com/internet/331675-clever-malware-masquerades-as-windows-11-installer) Feb 15

(https://www.extremetech.com/computing/331701-amds-milan-x-goes-head-to-head-with-intels-sapphire-rapids)
AMD's Milan-X Goes Head to Head with Intel's Sapphire Rapids in Leaked Benchmarks (https://www.extremetech.com/computing/331701-amds-milan-x-goes-head-to-head-with-intels-sapphire-rapids) Feb 15

(https://www.extremetech.com/gaming/331676-alienwares-qd-oled-gaming-monitor-coming-to-market-at-a-shockingly-low-price)
Alienware's QD-OLED Gaming Monitor Coming to Market at a Shockingly Low Price (https://www.extremetech.com/gaming/331676-alienwares-qd-oled-gaming-monitor-coming-to-market-at-a-shockingly-low-price) Feb 15

(https://www.extremetech.com/gaming/331641-nvidia-reportedly-prepping-two-new-rtx-3050s-one-with-just-4gb-of-vram)
Nvidia Reportedly Prepping Two New RTX 3050s, One with Just 4GB of VRAM

(https://www.extremetech.com/gaming/331641-nvidia-reportedly-prepping-two-new-rtx-3050s-one-with-just-4gb-of-vram) Feb 14

(//www.extremetech.com/)

ABOUT (HTTPS://WWW.EXTREMETECH.COM/ABOUT/)

CONTACT (HTTPS://WWW.EXTREMETECH.COM/ABOUT/)

NEWSLETTERS (HTTP://WWW.PCMAG.COM/NEWSLETTER_MANAGE)

ADVERTISE (HTTP://WWW.EXTREMETECH.COM/ABOUT#ADVERTISE)

MORE FROM ZIFF DAVIS:

MASHABLE (HTTPS://WWW.MASHABLE.COM/) PCMAG (HTTP://WWW.PCMAG.COM/) GEEK (HTTP://WWW.GEEK.COM/)

ASKMEN (HTTP://WWW.ASKMEN.COM/) BESTGIFTS.COM (HTTPS://BESTGIFTS.COM/)

EVERYDAY HEALTH (HTTPS://WWW.EVERYDAYHEALTH.COM/) IGN (HTTP://WWW.IGN.COM/)

OFFERS.COM (HTTP://WWW.OFFERS.COM/) SPEEDTEST.NET (HTTP://WWW.SPEEDTEST.NET/)

TECHBARGAINS (HTTP://WWW.TECHBARGAINS.COM/) TOOLBOX (HTTP://WWW.TOOLBOX.COM/)

WHAT TO EXPECT (HTTPS://WWW.WHATTOEXPECT.COM/)

RSS FEEDS (HTTP://FEEDS2.FEEDBURNER.COM/ZIFFDAVIS/EXTREMETECH)

PRIVACY POLICY (HTTPS://WWW.ZIFFDAVIS.COM/PRIVACY-POLICY)

TERMS OF USE (HTTP://WWW.ZIFFDAVIS.COM/ABOUT/TERMS)

ADVERTISE (HTTPS://WWW.EXTREMETECH.COM/ABOUT#ADVERTISING)

ACCESSIBILITY (HTTPS://WWW.ZIFFDAVIS.COM/ACCESSIBILITY)

DO NOT SELL MY PERSONAL INFORMATION (HTTPS://WWW.EXTREMETECH.COM/CCPA)