NVMe: The Protocol for Future SSDs

When do you need NVMe?

You might have heard that Non-Volatile Memory Express or NVM Express (NVMe) is the next must-have storage technology. Let's look at what NVMe delivers. NVMe is a communications *protocol* for high rate storage systems that runs on top of the PCIe *physical layer* and allows host hardware and software to fully exploit the performance of high-end SSDs. Everyone needs to get it right now, right? Well, not so fast.

NVMe Background

First, some terminology. The PCIe physical layer consists of the silicon driver and receiver on either end of a wire that meets the voltage and signaling rate of the PCIe standard. NVMe is the protocol or language that is communicated over those wires. Those wires could also carry the PCIe *protocol*, but NVMe is a much more suitable way to communicate information pertaining to storage systems. It's no surprise that the development of PCIe and NVMe came from the same group of sponsoring companies.

NVMe was designed to address the limitations of existing storage interfaces and protocols like PATA, SATA, SCSI and SAS. The origins of these protocols go back over 30 years and have evolved based on the limited abilities of traditional mechanical hard disk drives. These legacy interfaces have lasted this long due to the need for forward and backward compatibility in both hardware and drivers. NVMe was designed to be extensible to last just as long as these traditional interfaces that it will replace. The full performance *potential* that NVMe has to offer is just that, potential that is not yet realized. Intel recently released their new Optane SSD drives based on the revolutionary 3D-XP memory technology. 3D-XP is orders of magnitude faster than NAND Flash and NVMe has ample performance headroom not only for the current generation of Optane but for many generations of newer, faster memory technologies.

NVMe enables systems to take advantage of lower data latency and internal parallelism of random access, memory-based storage. NVMe's reduction in I/O overhead and performance improvements give it the ability to process multiple, long command queues, compared to being able to process just one queue with previous logical device interfaces. NVMe advances also include being able to use only a single message for 4KB transfers compared to two. This increases the speed for servers processing a lot of concurrent disk I/O requests.

High-level comparison of AHCI and NVMe[5] storage protocols		
	AHCI (SATA)	NVMe
Maximum queue depth	One command queue; 32 commands per queue	65535 queues;[29] 65536 commands per queue
Uncacheable register accesses (2000 cycles each)	Six per non-queued command; nine per queued command	Two per command
MSI-X and interrupt steering	A single interrupt; no steering	2048 MSI-X interrupts
Parallelism and multiple threads	Requires synchronization lock to issue a command	No locking
Efficiency for 4 KB commands	Command parameters require two serialized host DRAM fetches	Gets command parameters in one 64-byte fetch

Source: www.wikipedia.org

System performance historically was limited by the performance of the storage - not the storage interface, but rather the storage media. In a mechanical hard disk the media performance is limited by the mechanics. No HDD is capable of saturating a SATA-III interface; the best ones come in at about 25%, which is the performance limit of SATA-I. SATA-II and –III were scaled up to use a fraction of the potential performance of NAND Flash, but the SATA protocol is the limiting factor once you reach certain throughput levels. High throughput means that a larger number of transactions can occur, but the SATA protocol does not allow enough pointer space or queue size to handle the performance of an SSD. Since SATA is a bridge between PCIe (whose protocol is also a storage bottleneck) why not simply eliminate both SATA and PCIe? Which is, of course, exactly what NVMe is.

So, why not move to NVMe?

The improvements in the storage interface, such as quadrupling of SATA bandwidth, isn't the reason that SSDs are so much faster than HDDs. SSDs are faster because the latency or response time of the storage media improved by 1000x from a mechanical drive to an SSD.

The observable system performance, what the user can see and feel, is limited by a bottleneck. For the longest time that bottleneck has been the mechanical hard drive. Once that was removed the next bottleneck was the SATA-I interface. Now with SATA-III SSDs the bottleneck today might be some other part of the system, perhaps the GPU for graphics performance or the CPU or memory for processing performance, and in some cases, it might also be the storage system. If you are running storage intensive operations, then NVMe might be exactly what you need. But if the performance bottleneck is elsewhere, or if you don't want to burden yourself with the higher power that comes with NVMe performance, then maybe NVMe is not what you are looking for right now.

The bottom line is that the incredible performance improvements realized by moving from mechanical storage to solid state are not going to be repeated by moving from SATA to NVMe. If the application doesn't need the additional storage performance, then no benefit will be realized.

Why NVMe will take over … eventually

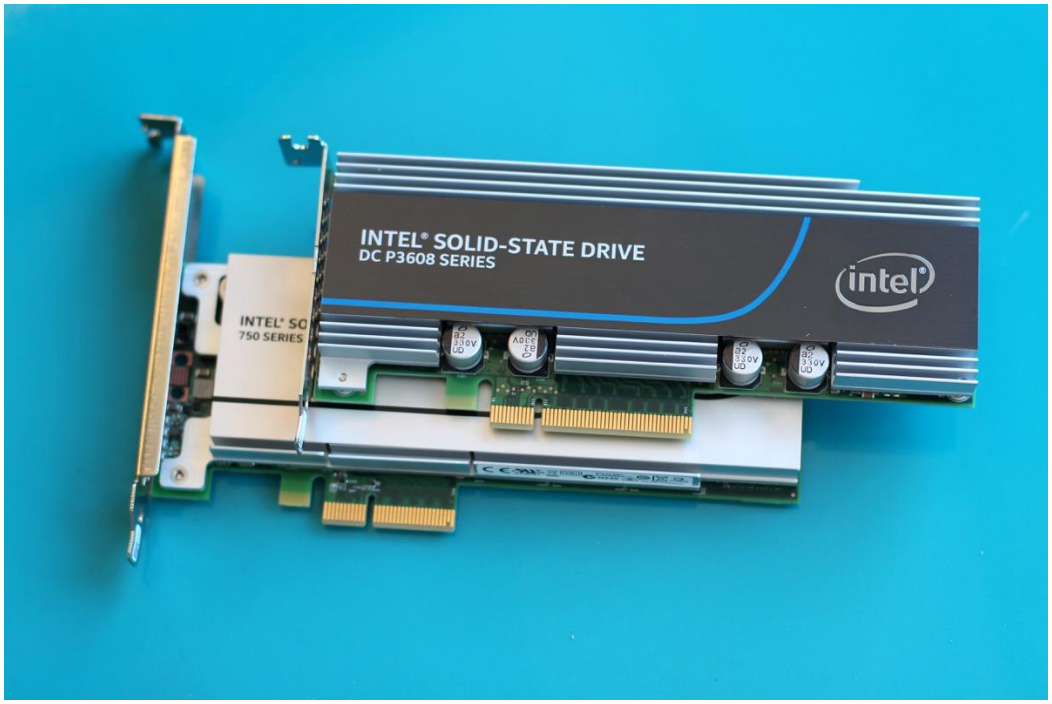
NVMe will become the de facto standard not because of the performance that it offers, but rather because of standardization and broad use. These transitions that are not performance driven can take a very long time to complete. Today traditional storage requires an ACHI (SATA) or SAS controller to connect mechanical hard disks to the PCIe bus, and NVMe is the solution that eliminates the bridge and instead makes the direct connection. But like many entrenched standards the SATA interface is well established, has multiple IP suppliers and is widely available with market based pricing.

A quick survey of SSDs indicates NVMe prices are coming in line for entry level SSDs, but the high-performance potential of NVMe comes at a price. Looking at a range of 512GB SSDs shows the SATA and NVMe M.2s to be about the same price, yet still more expensive than the 2.5". The premium is due to the performance and headroom of the M.2.

Storage drives	Interface	Performance	Price/GB
AiC x8 Enterprise	NVMe	~40Gb/s	5.0x
U.2 x4 Enterprise	NVMe	~18Gb/s	2.5x
M.2 x4 Prosumer	NVMe	~20Gb/s	2.0x
M.2 x4 Entry level	NVMe	~12Gb/s	1.0x
M.2 SATA	SATA-III	~3Gb/s	1.0x
2.5" SSD Prosumer	SATA-III	~4.5Gb/s	1.5x
2.5" SSD Entry level	SATA-III	~3Gb/s	0.9x

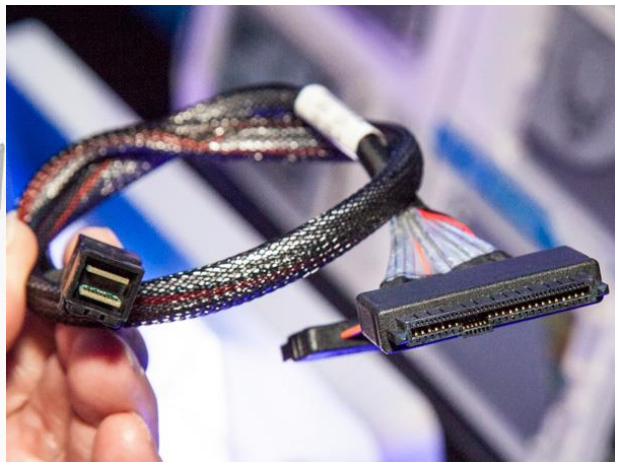
NVMe SSD form factors

NVMe drives come in three basic form factors, the Add-in card (AiC), the 2.5” (U.2) and the M.2. Since NVMe protocol runs on the PCIe physical layer, the connections will be nearly identical to PCIe. The AiC cards resemble standard expansions cards, the M.2 resembles mini-PCIe, and the 2.5” is meant to address the modularity requirements of storage systems. The AiC is high capacity and can be designed to support any number of PCIe lanes. These cards are generally for high IOPS applications, but you need to dedicate PCIe slots for the cards and they are not hot swappable. Below are a 4 lane and an 8 lane AiC from Intel.



A smaller version of this kind of implementation is the M.2 form factor, commonly measuring 22mm wide and a length of between 42mm and 110mm. The M.2 socket resembles that of an SO-DIMM for memory, so the card lays parallel to the main board. Of course, there are other socket orientations available, but the most common is the one used in laptop computers.

To round out the offerings is a 2.5” or 3.5” form factor that uses a U.2 cable with U.2 NVMe interface. These drives are intended for enterprise and datacenter applications and are generally available in high capacities and high write-endurance models. Similarly, the U.2 cable and host connector are significantly larger than SATA. So far there aren’t any low performance NVMe 2.5” drives, and with good reason, since the 2.5” SATA is the most cost effective solution available and is expected to be for quite some time.



NVMe Cross-Over

With NVMe prices that started rather high and have fallen quickly, it’s no surprise that pundits are predicting when it will be cheaper than SATA, but that’s not the way it works. NVMe will never be significantly cheaper than SATA. Prices for NVMe are high because it’s being sold based on its performance. When NVMe competes with SATA it will be competing based on $/GB. Since both use the same NAND Flash behind the controller there is no reason for NVMe to ever cost less than a SATA drive. This is of course different for DRAM technology since the memory chips for a given DRAM are a completely different design and subject to different market forces.

Robust NAND Devices

Today we have a range of NAND Flash devices with different capabilities and requirements. The oldest and most robust NAND technology is SLC (single level cell), then MLC (multi-level cell), now TLC and QLC are emerging along with 3D-NAND architectures. These newer technologies are more cost effective per bit, but at the same time have lower reliability, and endurance at the component level. Even today SLC is still manufactured for systems requiring the highest possible uptime and ruggedness.

The newest NVMe controllers are being developed to work with the latest NAND devices, meaning TLC, QLC, and 3D-NAND. It is unlikely that emerging NVMe drives will support the more reliable MLC and SLC NAND technologies. Applications that require a more stable and robust storage solution would be best to stick with proven MLC NAND, which will limit the availability of cost effective NVMe drives.

Conclusion

The well-established SATA storage industry is cost effective and entrenched for most applications. SATA drives are available with reliable MLC or SLC NAND Flash components. The higher performance NVMe will always sell at a premium whereas SATA-level performance will approach the cost of SATA drives. NVMe is not a disruptive technology that will significantly change the economics of storage systems - it is a performance-based solution for those willing to pay for it.

Sources

https://en.wikipedia.org/wiki/NVM_Express

https://en.wikipedia.org/wiki/PCI_Express