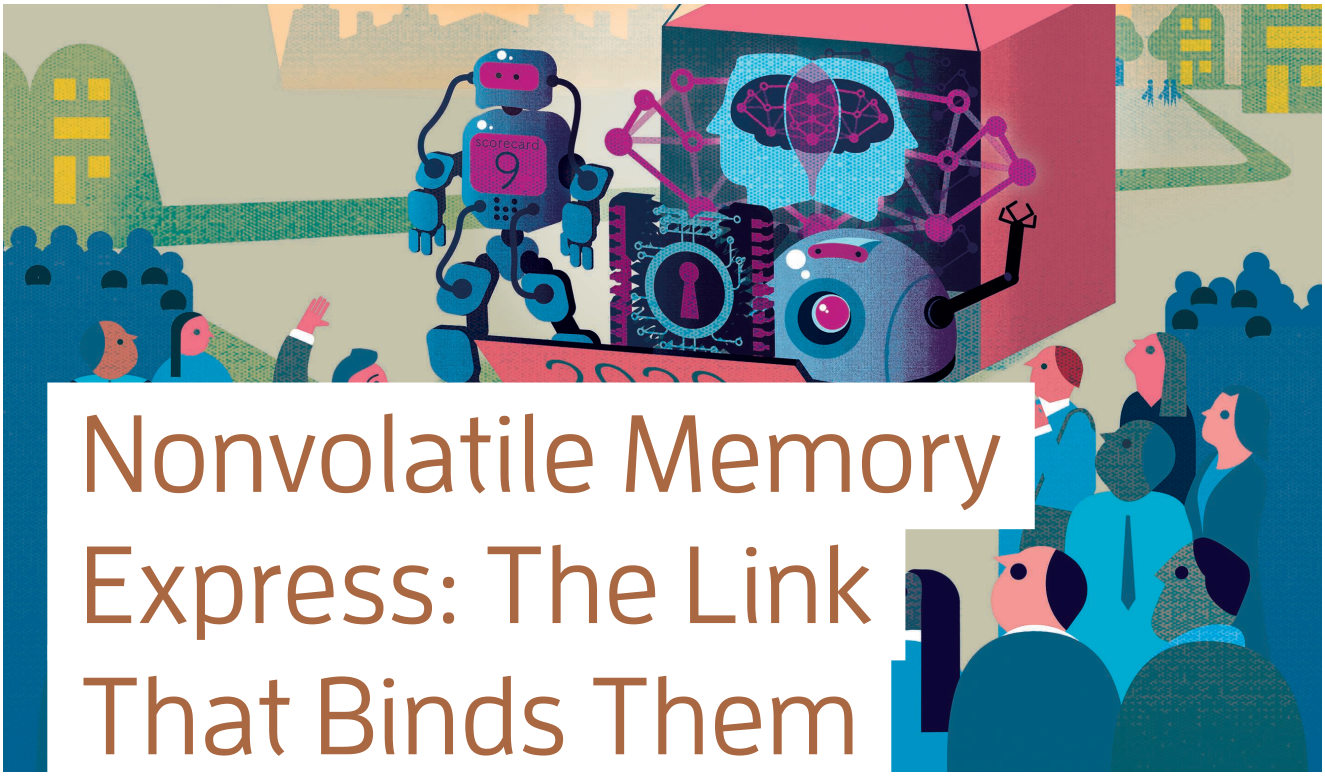


Nonvolatile Memory Express: The Link That Binds Them

Thomas Coughlin, Coughlin Associates

Within the next decade, nonvolatile memory will dominate storage and computing-network infrastructures. This article examines this technology and predicts its impact.

The Nonvolatile Memory Express (NVMe) interface has become an essential element in enterprise, client, and cloud computing. It provides high data rates and low-latency connections between storage and computing networks through the peripheral-component interconnect express (PCIe) serial computer-expansion bus. Data can also be transported by NVMe over fabric (NVMe-oF), such as Fibre Channel, Infiniband, and Ethernet and TCP/IP networks. As NVMe-based systems proliferate, network storage built with NVMe-oF will become dominant within the next five years. NVMe-oF enables new models for enterprise storage, including the use of emerging NVM technologies, disaggregated and composable computing, and infrastructure that incorporates various schemes to bring processing closer to storage.

BACKGROUND

Digital storage and memory are important elements in any computing system. As processing power has grown, so has the demand for storage capacity and performance. The interfaces used to connect digital storage devices to computing resources have changed with the evolution of both technologies. The serial advanced-technology attachment (SATA), small computer system interface (SCSI), serial attached-SCSI (SAS), and parallel interfaces served storage needs for many decades. However, with the rising deployment of solid-state drives (SSDs) that use flash memory and other NVM technologies, achieving peak performance from the SAS and SATA interfaces, which were designed to operate with the expected mechanical latencies of hard-disk drives (HDDs), became an issue. In addition, since SSDs were basically a bunch of semiconductor chips on a circuit board, they didn't need the same form factors as HDDs. Getting the best performance and real estate advantages with SSDs required new storage interfaces and designs.

Digital Object Identifier 10.1109/MC.2019.2943470
Date of current version: 22 November 2019

PREDICTIONS

There is great interest in adopting NVMe because of its improved performance compared to the SATA interfaces. That functionality is required to process all of the information generated by the Internet of Things (IoT) industry and consumer applications as well as autonomous and automated systems powered by various types of artificial intelligence (AI) algorithms that run on large data sets. The introduction of NVMe enables big changes in the storage technology that underlies modern computing. The projected developments include the following:

- *Prediction 1:* Improving NVMe-interface performance will follow the aggressive PCIe road map that forms the foundation of the technology.
- *Prediction 2:* NVMe will replace the SATA and SAS in most SSDs within the next few years.
- *Prediction 3:* NVMe will drive the advancement of NAND flash technology by enabling new NAND tiering techniques and programming functions that increase endurance, enable computational computing, and facilitate more memory-like access to data.
- *Prediction 4:* New SS storage form factors using various types of NVMe interfaces will enable advanced client and data-center storage capabilities.
- *Prediction 5:* NVMe-oF networking will become the dominant storage hardware in the near future, including for hybrid systems containing HDDs and SSDs.
- *Prediction 6:* The increasing use of emerging memory technologies, such as magnetic random-access memory (MRAM), resistive RAM (ReRAM), and phase-change memory (PCM), will help to provide the performance that future generations of NVMe devices will support.

I will discuss the basis of these predictions as well as technical challenges and risks in the balance of this article. Let's look at each of the forecasts.

Prediction 1

NVMe works through PCIe. PCIe is the interface backbone of many computers, and as a consequence, it enables the operation of graphics-processing units, general graphics cards, and storage and memory. The ubiquity of PCIe and the growing performance of the interface facilitate the growth of PCIe-based storage devices. The PCIe road map calls for continued performance increases, as shown in Figure 1. While PCIe 4 is ramping up in products today, PCIe 5 will be in devices during the next couple of years, and by the middle of the next decade, PCIe 6 will deliver 256-GB/s data rates across 16 lines.

NVMe was developed with the performance of SS storage in mind. The growth of NAND flash-based SSDs and new NVM technologies that perform better than SSDs and nearly as well as volatile dynamic RAM (DRAM) and static RAM (SRAM) led to the idea of a persistent memory (PM) layer that could support load/storage and direct access to memory applications as well as block-based data applications common for HDDs and most NAND flash SSDs. The PM layer enables fundamental changes in computer architecture as well as denser and more power-efficient system memory. Figure 2 illustrates the new memory/storage hierarchy by comparing data-access times for different memory and storage layers.

NVMe supports common data-center management capabilities, including formatting, secure erasing and sanitization, firmware updates, self-monitored health and status reporting, log pages, list devices, get/set features, and namespace management. In addition, it supports new commands and feature sets. Traditional and NVMe-specific capabilities can be used in open source and proprietary software to create storage-system management tools and data-analysis functions. NVMe drives can be incorporated into redundant-array-of-independent-disks systems.

Prediction 2

As NVMe SSDs have increased in volume, their prices have reached parity with, and will shortly fall below, SATA SSDs that have the same storage capacity. Since, with time, NVMe and its underlying PCIe road maps will support more features and faster data, NVMe SSDs will have nearly replaced new SATA SSDs in client systems and data centers by 2023. Likewise, as the NVMe specification develops and the price of NVMe SSDs declines, NVMe SSDs will displace SAS SSDs even though SAS SSDs have enterprise features that are useful to customers. This process will likely be a bit slower than the SATA displacement, but by 2025, NVMe SSDs will probably have replaced SAS SSDs.

Prediction 3

NAND flash-memory development is currently focused on the growth of the number of cell layers and bits that can be stored per memory cell. Currently, NAND flash products are shipping with up to 96 layers, but by 2020, the number will increase to approximately 128, and in the future, hundreds will be used.

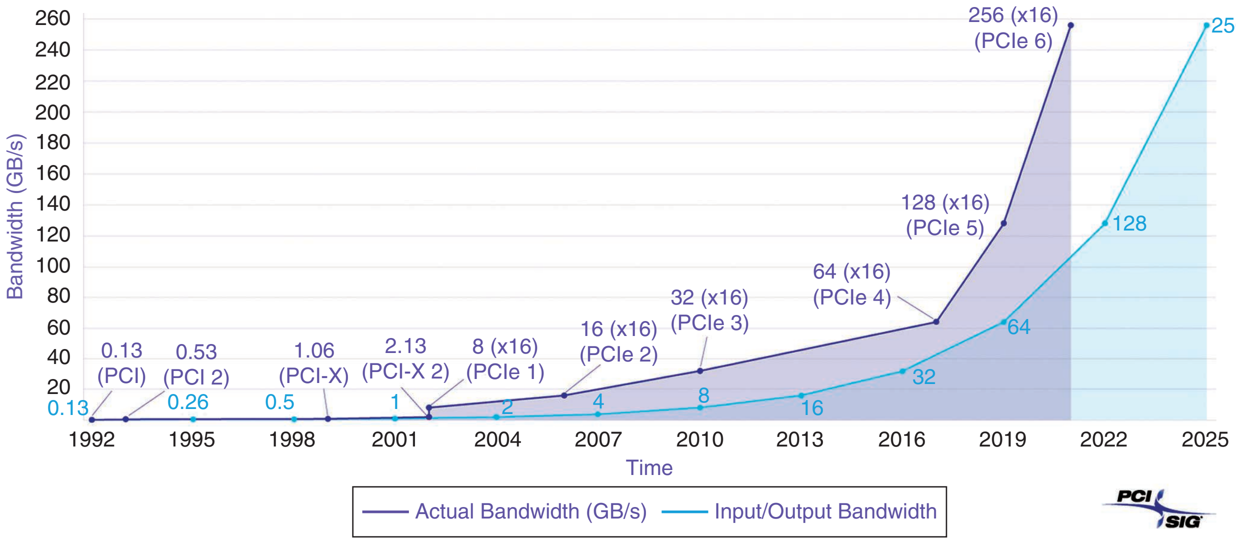


FIGURE 1. The PCIe bandwidth road map.[1] The bandwidth doubles every three years. (From [1], used with permission.)

Lithographic feature sizes posed development issues that were show stoppers for planar flash. The problem was resolved by going to the 3D flash design, which had relaxed space constraints. However, 3D flash requires more processes to create the memory cell layers and, thus, more process time per wafer, resulting in considerably more expensive NAND flash factories. The yield and production volume of 3D NAND flash memory must compensate for the increased fabrication cost so that the price per gigabyte will continue to decline and drive the technology's adoption.

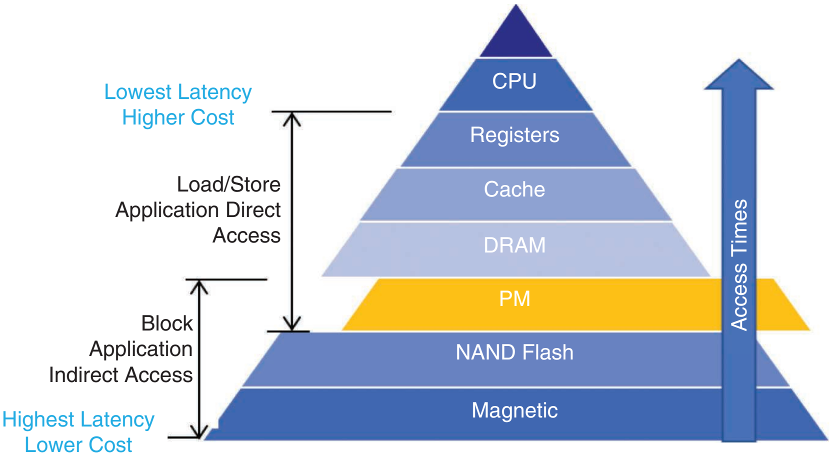


FIGURE 2. The memory/storage hierarchy with the PM layer.[2]

However, as the number of layers increases, wafer production costs rise, limiting the advances toward lower prices per gigabyte. For that reason, NAND flash manufacturers are pushing to get more bits per memory cell. Currently, many NAND flash products ship with two bits per cell (MLC) or three bits per cell (TLC), and the industry is striving to include more for most client and enterprise applications. There are problems with increasing the bit number per cell that relate to the decrease in the signal-to-noise ratio for detecting the individual bits. Thus, the error-correction algorithms used to distinguish additional bits per cell are more complex and take longer to execute, decreasing device performance. In addition, as the bit total rises, the number of times that the memory cells can be erased and rewritten declines (that is, the flash endurance erodes).

Managing the endurance (wear) of multibit memory cells requires new controller functions. The slower performance and lower endurance of four- and five-bit-per-cell 3D flash memory is leading to tiering concepts where MLC and even single-level-cell flash are used in combination with quad-level-cell (QLC) flash inside an SSD to improve performance and reduce wear. Programming models associated with the NVMe interface will play an important role in enabling the use of higher bit-per-cell NAND flash during the next few years and enable the price per gigabyte to decline.

New software is needed to take full advantage of NVMe storage. Application middleware and operating systems are no longer bound by file-system overhead, as in traditional storage, to run persistent transactions. The Storage Networking Industry Association (SNIA) Solid State Storage Initiative (SSSI) PM programming model enables connectivity beyond storage (including memory), networking, and processing. Figure 3 shows the SNIA-SSSI model with various PM access modes. The PM model supports block-based and file-storage innovation as well as capabilities that tap the performance of new NVMs. For block and file access, data are read and written using RAM buffers, with the status explicitly checked by software that controls whether context switches or polling is used for data transfers. Volume and PM modes enable load/store where data are loaded into or stored within processor registers. The processor makes the software wait for data transfers, and errors generate exceptions rather than having explicit status-checking. Those new programming functions will become widespread during the next few years.

With the slowdown in processor performance growth, it has become popular to utilize “accelerators” that are located away from the CPU to improve results and increase the composability of the compute/network/storage system. *Computational storage* refers to increasing the computing power close to or in storage devices to improve overall system performance. In a composable infrastructure, computing resources can also be located near or within storage hardware. Those resources can run services in a virtualized environment close to the data. The services could include basic data management, deduplication and encryption, and more sophisticated applications. With the processing taking place closer to the data, the latency and power consumption would decline since information wouldn’t need to move between the storage device(s) and a CPU. File system and storage-management functions could be offloaded from dedicated servers to the storage devices without server intervention. Even more advanced functions could be added to enable application-specific code to be executed at the storage devices on behalf of clients and servers. Early examples of local-compute functions include database searches, data mining, and some image-processing applications involving the built-in processor in a group of HDDs. More recently, SSDs have been constructed that use increased internal processing to perform various functions.[4]

The work on standards for the various approaches to computational storage is underway at the SNIA, where the Computational Technical Working Group was formed within the SSSI. Early efforts to increase the intelligence of storage devices go back, at least, to the active-disk concept.[5] Various types of computational-storage devices are available today, including the Newport Platform from NGD Systems,[6] Eideticom field-programmable gate array that provides virtualized offload-processing services in an NVMe environment,[7] and the Lightbits NVMe hardware accelerator that enables data reduction and protection and NVMe/TCP acceleration.[8]

Computational storage using NVMe devices and NVMe-oF networks will become common during the next decade. The demand for more energy-efficient and higher-performance storage combined with computing will help to drive

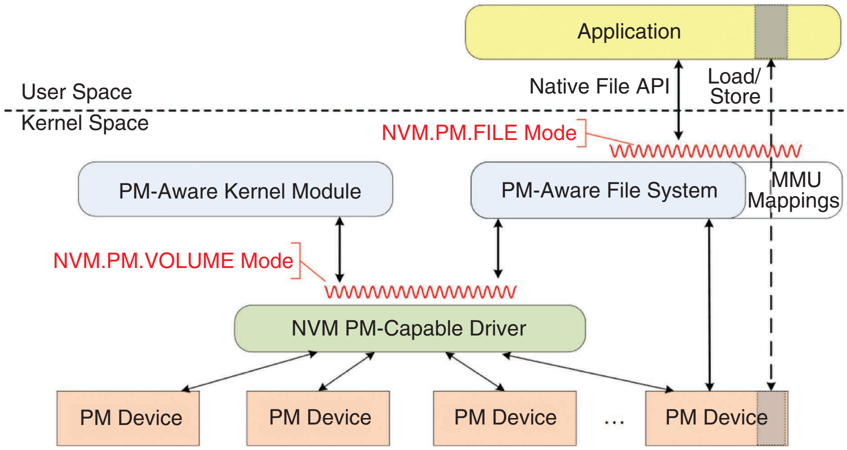


FIGURE 3. The SNIA-SSSI PM programming model.[3] API: application programming interface; MMU: memory-management unit. (From [3], used with permission.)

the development of in-memory computing, which is likely to make its appearance during the 2030s. Near-memory computing will also be enabled by new protocols, such as the Compute Express Link Interconnect (CXL) that runs on the PCIe physical bus and will be used for acceleration and near-memory computing within ten years. Since NVMe and CXL both use the PCIe physical layer, I predict that they will work closely together and probably merge as part of a larger suite of PCIe protocols during the late 2020s.

Prediction 4

New SS storage form factors using various types of NVMe interfaces will enable advanced client and data-center storage capabilities.

NVMe SSDs are available in a great many form factors, including PCIe add-in cards, M.2, and ball-grid arrays (BGAs). They offer higher-density storage than the older SSDs that mimicked the shape of HDDs. In addition, NVMe is available in physical formats, such as U.2, that are roughly the same shape as 2.5-in HDDs. The new form factors enable smaller SSDs in client devices, including BGAs that are essentially the size of a conventional semiconductor-chip package and contain the SSD controller and NAND flash memory chips. For instance, Toshiba's BG4, introduced in January 2019,[9] is a 16 × 20 mm NVMe PCIe 3.0 X4-interface BGA that uses a 96-layer 3D TLC NAND to provide up to 1 TB of capacity. As the number of 3D NAND layers increases and QLC flash comes into use, products with this form factor will deliver several terabytes of capacity, likely up to 10, within several years.

Prediction 5

The popularity of NVMe storage will spur advances in NVMe-oF that will lead to new options for storage and computing architectures. It is likely that NVMe-oF will displace other common fabric-storage networking

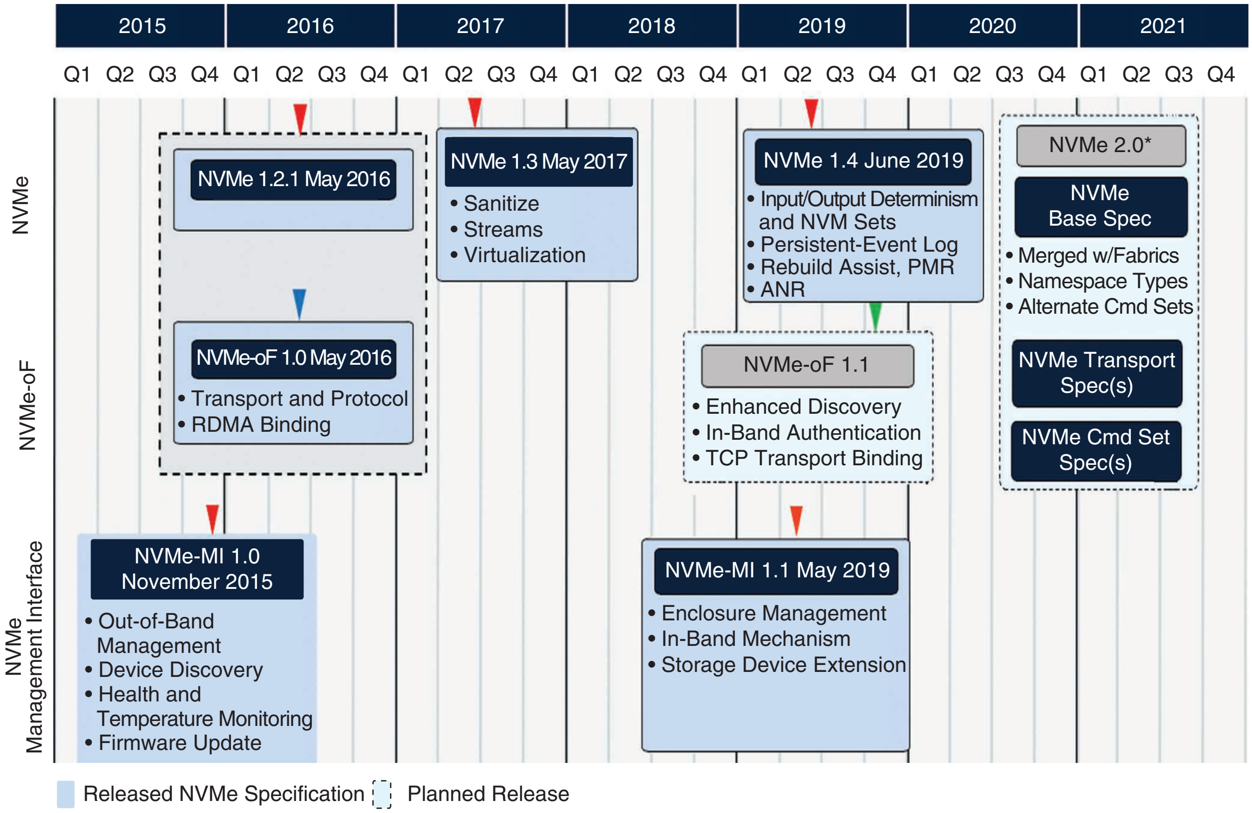


FIGURE 4. The NVMe road map.[10] RDMA: remote direct-memory access. (From [10], used with permission.)

approaches, such as the Internet SCSI (iSCSI). Figure 4 displays the NVMe road map, which includes the introduction of various feature sets, for example, the development of NVMe-oF where NVMe commands are transported across different types of storage networks. NVMe devices will also function as remote memory that is accessible through a storage fabric.

NVMe is a memory-mapped PCIe transport model, while fabrics are message based and may include shared memory. According to NVM Express Inc., a *binding* is a specification that enables reliable data, command, and response delivery between a host and NVM subsystems for NVMe transport. The binding may exclude or restrict functionality based upon the NVMe transport's capabilities. NVMe data transport can work across several fabric technologies, including the Fibre channel, InfiniBand, RDMA over converged Ethernet, TCP, iWARP, and possible future systems.

NVMe-oF will become the dominant network storage fabric of the 2020s, a development that has already begun. Western Digital introduced an HDD array with an NVMe network interface as part of its OpenFlex composable-infrastructure platform in 2018.[11] NVMe will likely become a universal storage interface used in all manner of systems and a broad range of applications within the next few years. It is even possible that native NVMe HDDs will eventually displace SATA and SAS if those protocols become obsolete.

SNIA's Object Drive Technical Working Group is developing a specification for native NVMe-oF drives[12] that will support SSDs with Ethernet ports using standard 8639 and 1002 (enterprise and data center SSD form factor) connectors and designs. In data centers, SSD form factors with native NVMe-oF will be commonplace. Faster NVMe devices with lower latency will result in NVMe-oF storage networks that have fewer delays, as shown in Figure 5. Faster PCIe buses and storage technologies in the NVMe SSDs will yield performance improvements.

Prediction 6

Tomorrow's NVMe devices will incorporate technologies that are faster than today's NAND flash memory, including Intel's Optane [3D cross point (XPoint)], PCM, enhanced NAND flash such as Samsung's Z-NAND, and

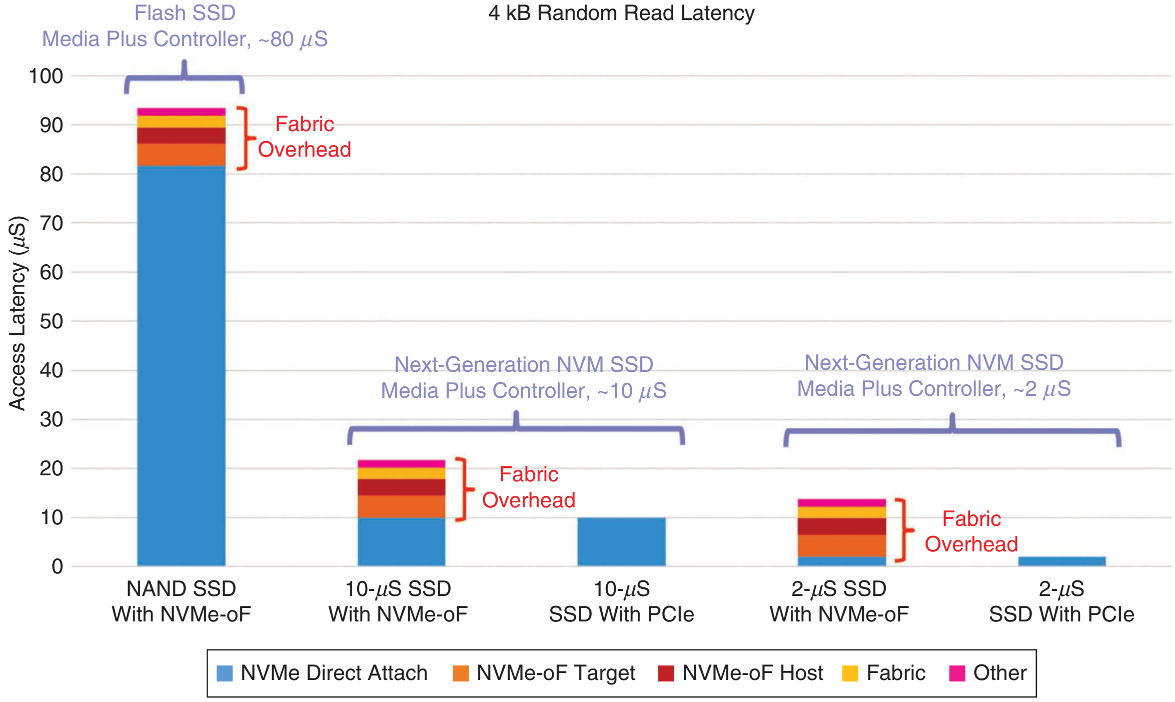


FIGURE 5. The improved 4-kB random-read latency for NVMe-oF with faster NVMe SSDs.[13] (From [13], used with permission.)

perhaps MRAM and ReRAM. They may be used by themselves or in combination to provide optimal trade-offs between performance and cost.

New NVM technologies will add to the suite of memory options, providing designers with more alternatives for their systems. Figure 6 presents the latency, endurance, and data retention for common VMs (SRAM and DRAM) compared with PCM, ReRAM, and MRAM. MRAM has the capability to replace or complement DRAM and SRAM. In particular, an MRAM technology called *spin-orbit torque*, which is currently in development, provides latency and data-rate performance that could match that of SRAM. MRAM cells occupy much less real estate on silicon wafers than their SRAM counterparts, which will yield a higher NVM capacity per chip by the mid- to late 2020s.

Demand from a spectrum of applications will drive NVM technologies' annual petabyte shipment growth, as shown in Figure 7. Replacing VMs with NVMs will reduce the energy consumption in NVMe storage devices and systems as well as embedded products. During the next decade, it will become common for those devices and systems to use encryption to protect the data in NVMs from unauthorized access. In 2019, all of the major semiconductor foundries were offering MRAM and some ReRAM options for embedded-device memory. By the mid 2020s, such devices will be conventional.

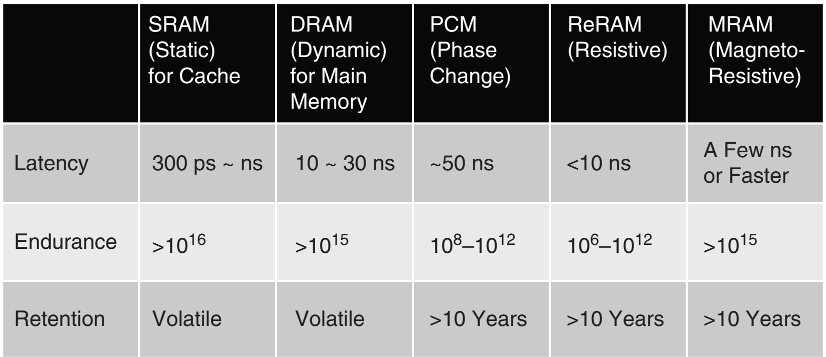
IMPACT

NVMe will have a great impact on businesses and what they do with data. It will enable applications that require fast information delivery, including real-time analytics and various emerging systems. It will become the dominant interface for client and enterprise storage devices, rapidly displacing most of the competing storage interfaces, and the price of NVMe hardware will drop quickly. There are even signs that NVMe could become the single storage interface for devices, including HDDs, within the next decade.

The data rate and latency of NVMe and NVMe-oF will enable the growth and development of IoT applications as well as data processing for various AI systems and big science and engineering modeling. NVMe opens the door to using NVM technologies in new ways, driving their growth and resulting in more energy-efficient and higher-performing storage devices and systems. The technologies will be key elements in creating composable, disaggregated storage infrastructure in data centers during the 2020s. The impact of NVMe and NVMe-oF is already widespread. Fibre channel and InfiniBand have been popular NVMe storage fabrics; starting in 2018, NVMe through the Ethernet and TCP/IP networks began to appear. In 2019, significant products were introduced that used various NVMe-oF approaches, including Fibre channel and Ethernet/TCP/IP-based systems.

NVMe-oF enables in- and out-of-band enclosure management. NVMe/TCP makes NVMe block storage through standard TCP/IP transport possible. It provides the ability to disaggregate (or separate) NVMe SSDs from computing systems without requiring changes to the networking infrastructure. With that disaggregation, storage and computing can be independently scaled in pools to maximize the resource utilization and optimization for specific workloads. Disaggregating HDD storage is common in data centers since the network overhead is small compared to the equipment's millisecond access latencies and low input/output operations per second (IOPS). On the other hand, disaggregating NVMe flash SSDs is challenging, since the network and protocol overhead is much closer to the microsecond latency and IOPS that are possible with NVMe SSDs.

Across moderate distances, NVMe-oF enables the scale-out of NVMe devices that have direct attached storage and latencies of fewer than 10 μs. It avoids unnecessary protocol translations and provides an end-to-end NVMe data-transport model, including the disaggregation of storage and computing. The capabilities of NVMe-based storage devices and systems are propelling a wave of start-up companies that offer new ways to disaggregate storage,



	SRAM (Static) for Cache	DRAM (Dynamic) for Main Memory	PCM (Phase Change)	ReRAM (Resistive)	MRAM (Magneto-Resistive)
Latency	300 ps ~ ns	10 ~ 30 ns	~50 ns	<10 ns	A Few ns or Faster
Endurance	$>10^{16}$	$>10^{15}$	10^{8}–10^{12}	10^{6}–10^{12}	$>10^{15}$
Retention	Volatile	Volatile	>10 Years	>10 Years	>10 Years

FIGURE 6. The comparison of the latency, endurance, and data retention for volatile and emerging NVM technologies.[14] (From [14], used with permission.)

create composable infrastructure (which puts computing, networking, and storage resources together as needed), manage and protect data running across dispersed hardware, and provide computational-storage approaches that will improve computer-system efficiency, increase performance, and save data-center energy.

TECHNOLOGY AND BUSINESS CHALLENGES

Although the need for high-performance storage is growing and driving NVMe and its related technologies, there are some technical and business challenges that need to be overcome for the protocol to reach its full potential. The first involves the historical advancement rate for the PCIe technology that NVMe is based upon. Although the evolution from the fourth generation of the PCIe to the fifth and sixth was expected to happen during increments of three or four years, the move from PCIe 3 to PCIe 4 took seven. According to the PCIe Special Interest Group, the slow transition was due to major changes to the firmware that underlies the PCIe interface; moving from generation four to five and, eventually, six will be much more rapid. Still, history indicates that the actual implementation may be slower than projected.

The next challenge will be determining which NVMe form factors will become the dominant interfaces in the future. At present, there are many similar but not identical form factors in the industry. For maximum proliferation of the NVMe interface, the industry must consolidate similar form factors into a few standardized device configurations. There are other protocols being developed using the PCIe physical interface, including CXL, that will create new compute-near-memory (that is, computational storage) technologies. Since NVMe is being used for accelerator-based computational storage solutions as well, it remains to be determined how it and CXL will work together and when one PCIe-based technology will be used versus another.

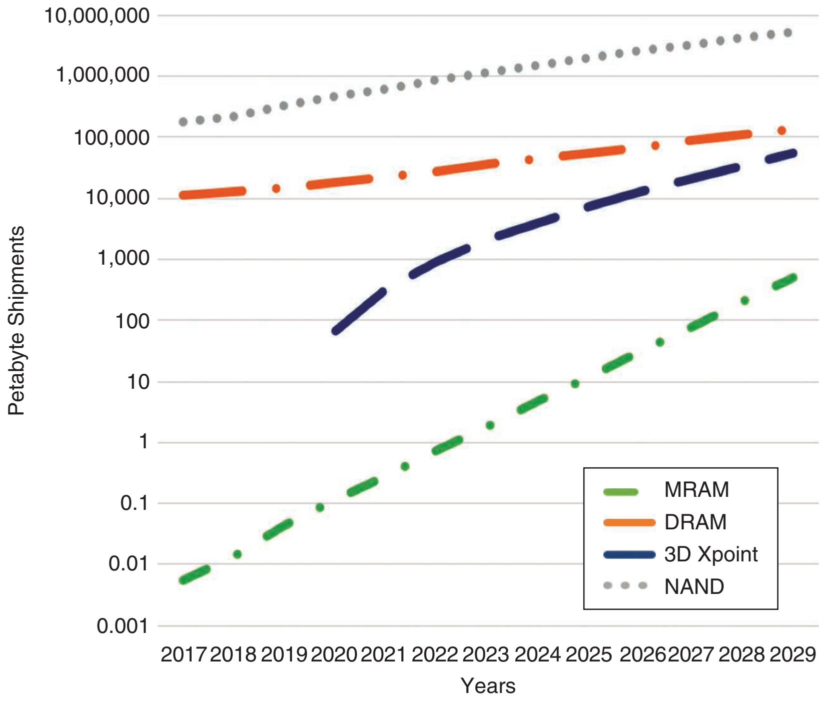


FIGURE 7. The projections for the annual shipment growth of NAND, DRAM, 3D XPoint, and MRAM.[15]

PREDICTION RISKS

The prediction that NVMe will replace SATA looks to be coming true in data centers and client devices, but the transition to SAS is a little less clear. The SAS transition will likely take some time, particularly because it is also used for many HDD interfaces. That creates some risk to the projections for NVMe dominance. In addition, as mentioned in the previous section, other PCIe physical-layer–based interconnects, such as CXL, could be used instead of NVMe-oF for in-box computational storage. Likewise, technologies including those from the Gen-Z Consortium might be used rather than NVMe-oF for out-of-box storage and computing connectivity.

The growth and use of emerging memories, including Intel's Optane, MRAM, and ReRAM, are still in their early adoption. While there are major commitments to provide NVM technologies for discrete and embedded applications, there are still many technical and business risks associated with them. Thus, their role in developing advanced storage and memory hierarchies in NVMe storage devices and systems is an open question.

NVMe will be a great enabler of new storage and computing architectures that will support

ABOUT THE AUTHOR

THOMAS COUGHLIN is the president of Coughlin Associates. He has more than 37 years of data-storage industry experience through engineering and management positions at several companies. Coughlin received a Ph.D. in electrical engineering. He is a Fellow of the IEEE, president of IEEE-USA, and active with the Storage Networking Industry Association and Society of Motion Picture Television Engineers. Contact him at tom@tomcoughlin.com.

IoT, video, big data, and AI applications by harnessing the full potential of SS storage. Its capabilities will grow with new generations of its underlying PCIe technology as well as software built around the needs and capabilities of modern SS storage. Likewise, NVMe-oF will enable a new network fabric built around the abilities of conventional NAND flash storage as well as emerging NVM technologies.

NVMe and NVMe-oF will be key technologies in creating future data centers that provide disaggregated computing, networking, and storage; enable composable-computing infrastructures; and create near-memory computing architectures using accelerators in or near the stored data. Near-memory computing will reduce power consumption and improve latency. NVMe is poised to be the dominant storage interface during the 2020s in client and embedded devices as well as enterprise applications.

REFERENCES

1. A. Yanes, "PCI-SIG DevCon 2019 update," in *The SCSI Bus and IDE Interface*, 2nd ed. F. Schmidt, Ed. Boston: Addison-Wesley, 1998, p 44.
2. T. M. Coughlin, "Persistent memory, NVM programming model, and NVDIMMs," presented at the IEEE MMST Conf., May 2018.
3. "NVMe programming model," SNIA SSSI, Colorado Springs, CO. [Online]. Available: https://www.snia.org/tech_activities/standards/curr_standards/npm
4. L. C. Quero, Y. Lee, and J. Kim, "Self-sorting SSD: Production sorted data inside active SSDs," in *Proc. 31st Symp. Mass Storage Systems and Technologies (MSST)*, May 2015. doi: 10.1109/MMST.2015.7208281.
5. E. Riedel, "Active disks: A Case for remote execution in network-attached storage," in *Proc. Parallel Data Systems Retreat*, Nov. 1997. [Online]. Available: https://www.pdl.cmu.edu/PDL-FTP/NASD/CMU-CS-97-198.pdf
6. "NGD Systems delivers industry's first scalable NVMe computational storage platform," NGD Systems, Irvine, CA, Mar. 2019. [Online]. Available: https://www.globenewswire.com/news-release/2019/03/06/1748906/0/en/NGD-Systems-Delivers-Industry-s-First-Scalable-NVMe-Computational-Storage-Platform.html
7. "Eideticom & Nallatech announce the NoLoad NVM Express U.2 computational storage platform with compression and peer-to-peer processing capability on Xilinx FPGAs," Eidetic Communications, Calgary, Canada, Aug. 2018.
8. "Lightbits Labs revolutionizes cloud infrastructure with first production NVMe/TCP solution; Provides hyperscale storage experience to private clouds," Lightbits, San Jose, CA, Mar. 2019. [Online]. Available: https://www.businesswire.com/news/home/20190312005220/en/Lightbits-Labs-Revolutionizes-Cloud-Infrastructure-Production-NVMeTCP
9. "Toshiba Memory America unveils 1TB single package PCIe-Gen3 X4L SSDs with 96-Layer 3D flash memory," Toshiba, Tokyo, Jan. 8, 2019. [Online]. Available: https://business.kioxia.com/en-us/news/2019/corporate-20190108-1.html
10. J. M. Hands and C. Brett, "NVM Express specification updates," NVM Express, Inc., Beaverton, OR, Mar. 19, 2019.
11. "Western Digital unveils the future of data infrastructure," Western Digital, Santa Clara, CA, Aug. 7, 2018.
12. "Object drives," SNIA Working Group, Colorado Springs, CO. [Online]. Available: https://www.snia.org/object-drives
13. P. Onufryk, "PCIe fabrics and NVMe," in *Proc. NVMe Developer Days*, San Diego, CA, 2018. [Online]. Available: https://www.nvmedeveloperdays.com/English/Collaterals/Proceedings/2018/20181205_APPL-101_Onufryk.pdf
14. S. Wang, "Emerging memory fundamentals," presented at the Proc. Stanford Emerging Memory and Artificial Intelligence Workshop, Aug. 2019.
15. T. M. Coughlin and J. Handy, "Emerging memories ramp up," Coughlin Assoc., Atascadero, CA, June 2019.