Transitioning to NVMe technology requires a clear understanding and thorough planning; this expert guide is here to help.
In this e-guide:

The benefits non-volatile memory express (NVMe) delivers to flash storage and server connectivity is undeniable.

But on a more practical note, how will enterprises adopt NVMe and how well will it work with existing technologies? What do you need to know to pick the NVMe-based array that provides your organization with the performance improvement it needs and also fits its budget?

In this expert guide, learn:

- The future of the NVMe protocol
- How NVMe's low latency, high throughput and highly parallel I/O will help the enterprise (and how its adoption will rock the enterprise boat)
- 4 NVMe-based array options to consider
- Top 5 reasons to select a hybrid or all-flash array
The present and likely future of the NVMe protocol

Jim O'Reilly, Cloud consultant, Volanto

Nonvolatile memory express started out as one of a handful of protocols to enable a faster connection of a flash drive to a PC via the PCI Express bus. While it is becoming the de facto standard for that in the PC world, in the enterprise world it is all about extending local connection out of the PC and into the world of networked storage via fabric. The goal is to give flash drives in a storage network the same I/O speed as one connected directly to a computer via PCIe. We'll look at what NVMe is, how it works and what its future holds.

The NVMe protocol was a logical consequence of the much higher performance of flash drives. Basically, the old SCSI-based file stack in the operating system couldn't keep up with the very high level of I/O operations. There were just too many interrupts, and the stack itself has thousands of CPU instructions per block of data.

Enterprises needed an offering that reduced interrupts dramatically, freeing up the associated overhead time for productive work in the CPU. In the same way, a method of transferring data that avoided interactions with the host CPUs as much as possible made sense.
All about that data transfer

The NVMe protocol evolved pretty rapidly to address these problems. Direct memory access over the PCIe bus addressed the interaction issue and was a well-tested method for moving data. It changed the transfer mechanism from a "push" system that required transfer requests and acknowledgments to a system that allowed the receiving node to "pull" data when ready. Experience had shown that this approach reduces CPU overhead to just a few percent.

The interrupt system was replaced by a circular queue method, with one set of queues for pending transfers and another for completion statuses. These command queues are parsed using remote direct memory access (RDMA), while the completion queues are handled in blocks of responses, which effectively aggregates interrupts.

One complaint on SAS/SATA/Fibre Channel file stacks was the lack of any sense of priority or source owner. NVMe attacks this elegantly, with 64,000 possible queues, each of which identifies both originator and priority. This permits data to be delivered back to the originating core, for example, or to a specific application. This addressing scheme becomes even more powerful when we look at the extensions of NVMe to fabrics.

The physical side of NVMe has resolved out as either a connection of two PCIe lanes contained in an M.2 connector or a standardized SATA Express (SATAe)
connector. Both of these are defined to allow SATA drives to be connected properly as an alternative to NVMe/PCIe.

M.2 technology allows very compact SSDs to be attached. These M.2 drives leave out cases and other drive items to save space and cost and, as a result, substantial capacities are available in 1.5-inch and 3-inch packages that are just an inch or so wide. We can expect 10 TB or more in these tiny form factors in 2018.

Size, though, is an unexpected benefit of NVMe. The real headline is summed up in announcements made at September’s Flash Memory Summit, where several drives boasting 10 billion IOPS were described. This compares well with the 150 IOPS of a traditional HDD, and is the primary reason that enterprise HDD sales are falling rapidly.

NVMe has made the SAS interface obsolete in the process. SAS is based on a SCSI software stack, and thus the old long-winded fileIO systems. It couldn't keep up with even small configurations of SSDs, and that was when drives achieved just 2 billion IOPS each. Attempts to add RDMA to SAS have been discussed, but SATAe has clearly won the battle.

With SAS fading away into the sunset, what about SATA itself? There is no cost difference involved in the actual PCI or SATA interfaces. They are almost identical electrically, and chipset support for autodetection and common electronics for connection are making any host-side connection trivial.
The only reason to keep SATA is to maintain an artificial price differential for NVMe drives versus SATA drives. The dynamics of pricing in the SSD market are much more sophisticated than in HDDs. Latency, IOPS and drive durability are all considered, as well as capacity, while the concept of enterprise versus commodity drives has become quite fuzzy as large cloud providers use the other differentiators to determine what they need.

Appliance-level redundancy and a no-repair maintenance approach mean the humble single-port commodity drive can move into the bulk storage space, while higher-performance and long-durability drives make up the primary tier. The upshot of this change in differentiators is that there is no long-term need for SATA, so it too will fade away, allowing NVMe/PCIe to be the local drive interface.

What technology will NVMe battle next?

All is not smooth sailing for the NVMe protocol, however. We have Intel's Optane technology to contend with. Intel plans to connect Optane SSDs using an Intel-developed -- and strictly, proprietary -- fabric scheme. This will be based on OmniPath, which will be native to the CPU. Intel's worldview is that this fabric will connect all the crucial modules in a server, including CPUs, graphics processing units and field-programmable gate arrays, as well as memory modules and local SSDs.
Intel drives the CPU business and can make this approach stick. Any alternative interfaces will need translation chips, which brings up the next evolution in NVMe. The idea of extending the NVMe protocol over fabrics has been in the works for some years, even as early as the first NVMe discussions. It’s a sensible extension of the approach, especially as we have huge amounts of experience with RDMA over Ethernet and InfiniBand.

We can expect something of a battle between Intel's OmniPath approach and Ethernet/RDMA. The latter has the huge advantage of experience in the market, a strong technology roadmap, and plenty of installed base and near ubiquity as a cluster connection scheme. With industry leaders such as Mellanox designing translators from OmniPath to Ethernet, I bet we’ll see clusters connecting using a hybrid of the two, with OmniPath internally and Ethernet/RDMA externally to the servers.

This is the future of storage systems. One of these fabric approaches will be anointed, and we’ll move forward to rapidly accept it. The reason is that connecting all the major modules in a cluster into a virtualizable pool of resources that can be directly accessed via the fabric grid will make for some very powerful and versatile systems.
NVMe wins whatever happens. The ring-buffer/RDMA approach has worked so well that all of these various offerings will use it to handle operations. It’s very safe to say that the NVMe protocol is the future of computing.
Move to NVMe technology will depend on need and speed

George Crump, President, Storage Switzerland

Nonvolatile memory express is a storage protocol designed specifically for memory-based storage, enabling technologies like flash to reach their full potential. Since NVMe technology is a change to infrastructure, it would likely take years for a data center to take advantage of it fully. But with each incremental step forward, organizations will see performance improvements.

NVMe is an industry standard that replaces the SCSI storage protocol. It supports thousands more simultaneous commands than SCSI can, and it can more deeply queue those commands. The primary benefit of NVMe technology is more responsive storage thanks to lower latency. It also uses Peripheral Component Interconnect Express (PCIe) as the primary storage interconnect.

NVMe is available as a drive interconnect and as a network protocol, thanks to NVMe over Fabrics (NVMe-oF). The network version enables Ethernet and Fibre Channel (FC) networks to transport data at speeds similar to locally attached storage.
The data center journey to NVMe will be multistep, with each step driven by practical need, as well as the speed at which NVMe technology can mature.

### NVMe-based all-flash arrays

The first step to transitioning to NVMe in the data center is using all-flash arrays in which the drives inside the system are NVMe-attached. Externally, however, there are no significant changes to the system. Connectivity to the storage network and the rest of the environment is still the same: traditional Ethernet and/or FC. Even hyper-converged systems that use server-based NVMe drives will still interconnect their nodes via traditional Ethernet. The good news is that inserting an NVMe-based all-flash array into the storage infrastructure is seamless. The bad news is that it doesn't take full advantage of everything NVMe technology has to offer.
What is the payoff of NVMe in the all-flash array if data has to use traditional protocols coming in and leaving the storage system? The reality is that the storage system becomes a choke point for performance, especially in a shared
The NVMe ecosystem by the numbers

The NVMe ecosystem today

- **$65 billion**: Estimated value of NVMe market by 2021
- **>130**: Companies either active in NVMe Express group or offering NVMe products, as of May 2018
- **2.5 inch**: NVMe form factor SSDs gaining market share over plug-in NVMe PCIe SSD cards

Click on image to view full size
There is a network interconnecting the communication among software, CPU and storage. With all data sent to and from this system, the quicker data can traverse the internals of the system, the better overall performance will be.

Until NVMe burst onto the scene, the internal network in most all-flash arrays was SAS. Now, it is quickly becoming NVMe. SAS is a slower connection and has to deal with the SCSI protocol inefficiencies. NVMe is faster in terms of connectivity and more efficient in how much data it can operate on at any given point in time.

**NVMe-based scale-out storage**

NVMe-oF is still in its early days. While it works, it is still fragile, and interoperability leaves a lot to be desired. This instability is why the first step to an NVMe data center is all-flash arrays with NVMe internal storage but external traditional networking. **NVMe-oF does work**; vendors just have to keep tight control over the components to eliminate problems. The need for a controlled network deployment makes NVMe-based scale-out storage an ideal second step for the NVMe transition.
How NVMe-oF works
Scale-out storage creates a cluster of servers (nodes) -- each with their own internal storage. It aggregates the storage from each of the nodes into a single virtual pool of storage. The network that interconnects these nodes is typically traditional Ethernet using IP. The internode communication is intensive, especially with the addition of more nodes. NVMe-oF is a more ideal interconnect thanks to its high performance and low latency. It should enable scaling to more nodes with better overall performance.

End-to-end NVMe

The next step to transition to NVMe technology is end-to-end NVMe, where the storage system and the servers attaching to it are all connected via NVMe. This design should bring in-server storage performance to shared storage. Major switch vendors Cisco and Brocade have added NVMe support to their switches, and there are several NVMe host bus adapter (HBA) cards available.

Unfortunately, there are several roadblocks to end-to-end NVMe:

- **Interoperability issues.** Until any network card can work with any other network card and switch, deployments will go slowly.
- **Infrastructure upgrades slowly.** Switches and HBAs that don't support NVMe need to reach end of life, which could take years.
- **Lack of need in terms of broad implementation.** We are actually at a point where the network (without NVMe) and the storage (with or without) are faster than what most data centers need.
How to get started

For the few data centers taxing their current all-flash arrays and network, moving to NVMe right away might help. But they should also consider a network upgrade at the same time. There are several NVMe all-flash specialists that can either deliver a turnkey end-to-end product or have all the partnerships in place to do so.

Most data centers, however, can't take immediate advantage of NVMe, so this is not a situation where an organization needs to throw out its all-flash array and replace it with an NVMe-based all-flash array. As storage system refreshes occur, it makes sense to look for flash arrays that are at least NVMe in part. It also makes sense to make sure that any investment in network infrastructure -- be it switches or HBAs -- has NVMe compatibility built in.

NVMe is not a case of hurry up and wait; it is more a case of hurry up and plan. Now is the time to make sure future investments -- be it in storage or in infrastructure -- are ready for NVMe technology.
In this e-guide

- The present and likely future of the NVMe protocol  p. 2
- Move to NVMe technology will depend on need and speed  p. 8
- 4 NVMe storage array options to consider  p. 16
- Further reading  p. 22

4 NVMe storage array options to consider

George Crump, President, Storage Switzerland

When selecting your first NVMe storage array, it’s important to understand the four basic implementation options to get the system that best meets your enterprise’s performance needs and budget.

1. SAS replacements

NVMe extracts better performance from flash media by using Peripheral Component Interconnect Express, as well as increased command counts and queue depths. Replacing SAS-connected drives with NVMe-connected drives is the most common implementation method. Creating an NVMe-based system is straightforward because most all-flash array (AFA) software runs on top of a Linux kernel and Linux natively supports NVMe. So, moving to NVMe is seamless for the software.

Switching to NVMe can impact compute, though. Vendors choosing the SAS replacement method must increase the CPU power of their systems to benefit from the NVMe investment. The more expensive NVMe drives and more
powerful CPUs put these NVMe all-flash arrays at a higher price point compared to SAS-based systems.

Merely replacing SAS flash with NVMe flash limits the performance improvement to interactions within the system. External connectivity is typically still Fibre Channel (FC) or traditional Ethernet, so once data leaves the NVMe storage array, SCSI and its latency are reintroduced. Nevertheless, many organizations see a performance improvement, especially those where the SAS-based AFA is flooded with I/O streams.

2. Hybrid integration

Hybrid arrays mix flash and hard disk drives. These systems can keep costs down and performance relatively the same as an AFA, assuming a large enough flash storage area and accesses from the hard disk tier are kept to a minimum. The problem with a hybrid system is that the performance delta between flash and hard disk can be too large, and when a flash miss occurs, users may notice the performance drop.
These hybrid systems integrate NVMe flash and high-density SAS flash. They keep costs down by limiting the size of the NVMe tier. It needs only to be big enough to store the most active data segments. Because of the smaller size, there's also less need for increased CPU power. And the use of two flash technologies means there's almost no noticeable performance impact.

While SAS flash is fast, it's not as fast as NVMe. Many organizations may find they must continue to increase the size of the NVMe flash tier to keep pace with workloads like high-end databases and big data analytics processing.
3. Scale-out systems

Scale-out systems can also benefit from NVMe. Today, the interconnections between nodes made over the traditional IP protocol add latency. NVMe-oF enables internode communications at internal storage speed. It's as if the cluster nodes were internally connected to each other. The reduction in latency should enable scale-out systems to scale further without impacting storage I/O latency.

4. End-to-end NVMe

End-to-end NVMe connectivity is the next step. It enables hosts and bare-metal applications to communicate with storage at speeds and latencies similar to DAS.

End-to-end NVMe requires more than just installing a new NVMe storage array. Organizations looking to adopt this approach must also upgrade their network. They don't, however, need to replace their network because all FC switches and most storage class Ethernet networks support NVMe and traditional SCSI-based protocols simultaneously. The same is true of network cards and host bus adapters.
Most of the early vendors to ship end-to-end NVMe systems are startups. These vendors invest in making sure their storage systems don't bottleneck the NVMe data flow through the use of field-programmable gate arrays and even application-specific integrated circuits to offload storage software processing. They target AI and machine learning workloads, which justify the massive I/O potential of these systems.

Picking the best NVMe storage array

The key question is: How much performance does your organization need? All NVMe options promise to improve the performance of flash arrays by reducing latency. The problem is that the performance improvement may be more than many organizations will ever need, and that performance comes at a price.

Storage infrastructures are reaching the point where they can deliver more performance than an organization requires. Buying the fastest possible system that fits within the IT budget may no longer be a sound strategy.
In addition to understanding the NVMe storage array options, you also must predict what your maximum I/O requirements will be over the next five years and select the array that best meets that need. You may find that a traditional SAS-based system provides the required performance and saves your organization a significant amount of money.
In this e-guide

- The present and likely future of the NVMe protocol  p. 2
- Move to NVMe technology will depend on need and speed  p. 8
- 4 NVMe storage array options to consider  p. 16
- Further reading  p. 22

About SearchStorage

At SearchStorage.com, it's all in the name.

Well, that's not entirely true, but our name does aptly describe our goal: to provide you with relevant, independent advice, tips, and news that help you address your specific storage needs.

In 1999, we launched SearchStorage.com to create a space in which storage experts could share objective content to help IT pros, like yourself, solve any and all of today's storage challenges. It's clear that storage challenges are only becoming more complex. So, whether you have a specific question around implementing parallel I/O technology or you're purchasing a new NAS system, you'll find detailed articles on SearchStorage.com to master all your latest storage projects. Welcome!

For further reading, visit us at https://searchstorage.techtarget.com/

Images: Fotalia

©2019 TechTarget. No part of this publication may be transmitted or reproduced in any form or by any means without written permission from the publisher.