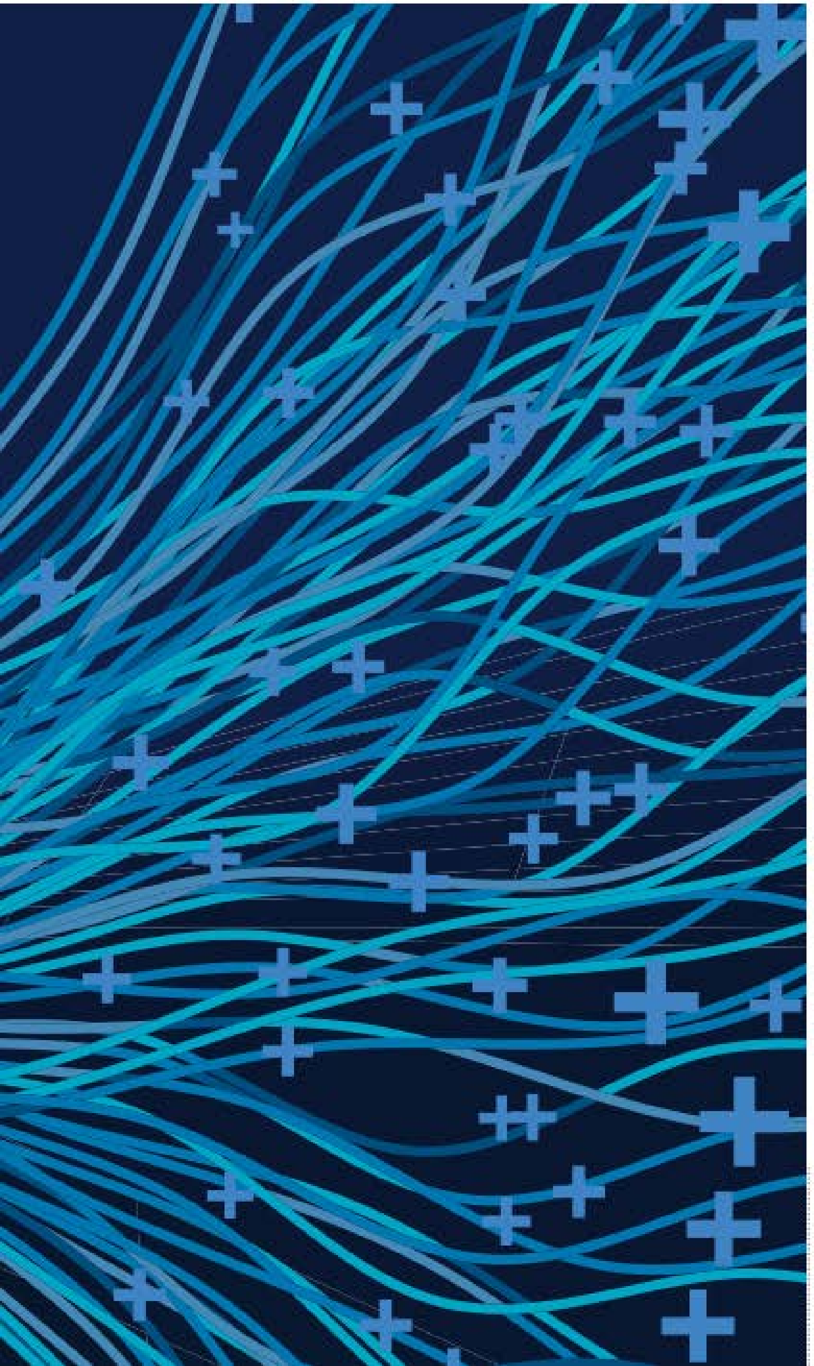


*E-Guide*

# Networked Flash Storage Goes to the Next Level

# In this E-Guide:



This e-guide explains where solid state storage is going. Learn how flash arrays have become the new normal in the data center. And get up to speed on how an increasing variety of networked NVME options push the boundaries of storage performance. You'll also get insights into their implementation challenges and how to solve them.

# NVMe-oF Products are Changing the Face of Data Storage

*Kurt Marko, Consultant*

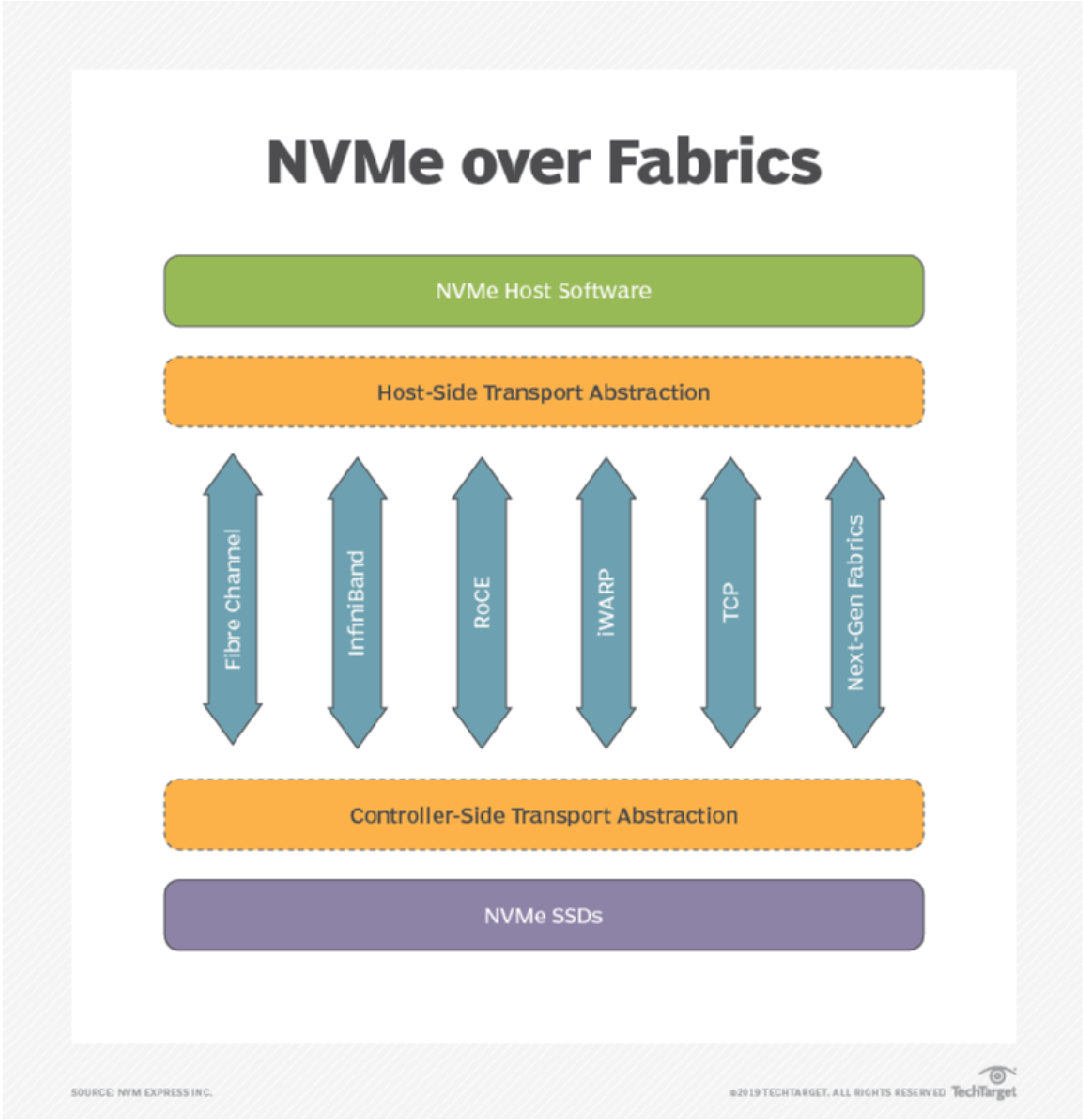
The NVMe interface is rapidly becoming the preferred interconnect for flash disks and all-flash arrays because of its superior performance and lower system overhead. Indeed, even as IDC recently warned of a sales slowdown for external storage in the EMEA region, it noted that the AFA market was red hot -- up 20% annually to 35% market share. Storage buyers are clearly ready to spend precious capital on better performing NVMe devices rather than SSDs using legacy storage interfaces.

NVMe has begun displacing SSDs in servers. However, with organizations using clusters of VMs and container servers for most workloads, it hampers workload portability if storage is locked down to a particular system. Fortunately, NVMe-oF has finally made it through the specifications process and NVMe-oF products are starting to provide a viable option for NVMe-based shared storage.

A fundamental feature of NVMe-oF is binding support, namely the ability to operate over various underlying transport fabrics including Fibre Channel, InfiniBand and Ethernet. Protocol bindings serve as the connection between NVMe and the network transport, but due to the technical differences of various network protocols, binding also places restrictions on NVMe capabilities and defines how NVMe is managed using the underlying fabric.

We gave advice on evaluating NVMe-oF options in this article. Although the specification for NVMe-oF over TCP lagged the other bindings, it was finally ratified in late 2018 and ushered in an era in which NVMe-oF hosts and controllers can communicate over any standard IP network. While the TCP specification focuses on software implementations using the TCP stacks on host OSes and AFAs, it doesn't preclude hardware-accelerated implementations.

As a young technology, the ecosystem of NVMe-oF products is rapidly evolving. However, it has now reached a degree of maturity, making NVMe-oF devices suitable for production workloads. The following is a look at some of NVMe-oF product categories, devices, software and supported systems.


NVMe over Fabrics
NVMe Host Software
Host-Side Transport Abstraction
Fibre Channel
InfiniBand
RoCE
iWARP
TCP
Next-Gen Fabrics
Controller-Side Transport Abstraction
NVMe SSDs
SOURCE: NVM EXPRESS INC.
©2019 TECHTARGET. ALL RIGHTS RESERVED
TechTarget

# The state of NVMe integration testing

Each week, the tech news wires are full of NVMe product announcements, many of which won't be released for months. The hype makes it hard to keep track of the market, much less figure out which products are compatible with one another.

Fortunately for technology buyers, the University of New Hampshire InterOperability Laboratory (UNH-IOL) has long served as a clearinghouse for reliable information on network devices and their support of various network standards. UNH-IOL has been testing NVMe-oF products for two years and its integrators list is an excellent resource for finding standards-compliant products.

The list includes several product types, such as:

- NVMe host platforms
- NVMe drives
- NVMe switches
- NVMe management interfaces
- NVMe-oF hardware and software targets: Fibre Channel, TCP and remote direct access memory (RDMA) over Converged Ethernet, or RoCE
- NVMe-oF initiators
- NVMe-oF switches

## Up to code

Some of the products that have completed UNH-IOL NVMe-oF compliance testing.

| | |
|---|---|
| TCP switch | Hewlett Packard Enterprise (HPE) M-series SN2100M |
| TCP target | Lightbits LightOS<br>Solarflare XtremeScale NIC<br>Toshiba Memory KumoScale software |
| TCP target and initiator | SuperMicro X10DRW-i motherboard with SANBlaze software |
| RoCE switch | HPE M-series, HPE SN2100M<br>Mellanox Technologies SN2410, SN2010 |
| RoCE target | Toshiba Memory KumoScale software<br>Western Digital OpenFlex F3000<br>Broadcom SST110, PS1100R<br>E8 Storage software<br>Exten Technologies Exten HyperDynamic software<br>Mellanox Technologies BlueField Platform |
| RoCE target and initiator | SuperMicro X10DRW-i motherboard with SANBlaze software<br>Marvell FastLinQ |
| RoCE initiator | Mellanox Technologies ConnectX-5, ConnectX-6 |

SOURCE: UNIVERSITY OF NEW HAMPSHIRE INTEROPERABILITY LABORATORY

# NVMe-oF arrays

Compliance testing is a tedious process that's dependent upon vendors submitting products they believe will pass. Hence, it necessarily lags product introductions, particularly in a market as dynamic as NVMe.

The following is a sample of some of the storage systems supporting NVMe-oF:

- Apeiron Data has a proprietary alternative to NVMe-oF it calls NVMe over Ethernet (NoE). It delivers NoE via a 2U appliance supporting as much as 24 2.5-inch drives and 32 40-GbE ports. The design is reminiscent of the ATA-over-Ethernet systems built by Coraid, but with an NVMe performance boost.
- E8 Storage offers 1U and 2U appliances with as many as eight 100-GbE ports and 68 to 136 TB of usable capacity with drivers for the following Linux systems: RHEL 7.1 and above or CentOS 6.7 and above; Ubuntu 14 and above; SUSE Linux Enterprise 12 and above; and Debian 8.6 and above. E8 also offers a software product for servers with NVMe drives and compatible network interface cards (NICs) that provides similar capabilities to its hardware by aggregating capacity into volumes available over an NVMe fabric.
- Excelero's NVMe-based storage software platform pools NVMe capacity on standard servers, regardless of the local or network file system used. It creates distributed block volumes over any network fabric and protocol. For example, a capacity-optimized platform with 24 NVMe drives provides almost 1 million IOPS at less than 200 microseconds latency and can scale to millions of IOPS in a single-server rack.
- Pavilion Data has a 4U array providing 14 TB to 1 petabyte of capacity, using 2.5-inch NVMe drives with up to 20 million IOPS and 20-microsecond latency. It can scale from one to 40 100 GbE network ports and two to 20 active controllers. The array connects via RDMA-capable Ethernet, but also supports NVMe-over-TCP for

NVMe-oF Products are Changing the Face of Data Storage

5 NVMe-oF Implementation Questions Answered

Flash Arrays are the New Normal

clients that don't support RDMA. Pavilion has drivers for RHEL, CentOS 7.4 or greater and Ubuntu 16 or greater.


- Pure Storage Flash Array//X with Purity software offers from 55 TB (3U) to 3 PB (6U). The Purity DirectFlash software supports RoCE and provides a range of storage services like data compression, deduplication, high availability, snapshots, data encryption and Windows file services.
- Vast Data offers both software and hardware products using NVMe-oF. They provide a scale-out, transactional file system that can scale to server clusters of up to 10,000 systems. Vast has developed what it calls a Disaggregated, Shared-Everything architecture with three distinctive elements:

1. Separate control and data plane hardware with x86 servers running the storage software and data spread across NVMe-oF disk enclosures.

NVMe-oF Products are Changing the Face of Data Storage

5 NVMe-oF Implementation Questions Answered

Flash Arrays are the New Normal

2. A high-speed Ethernet or InfiniBand network fabric connecting compute and storage nodes. Compute nodes are typically housed in 2U quad-server enclosures with four 100-GbE NICs, while the storage nodes are also a 2U chassis with 44 x 15.36 TB quad-level cell flash disks and 12 x 1.5 TB U.2 XPoint devices.

3. Distributing metadata across the storage nodes on dedicated fast Intel Optane 3D XPoint memory in the NVMe-oF enclosures. Such a distributed metadata design enables the compute nodes to be stateless and use non-redundant, lower-cost systems.

## OS and application support

As the UNH-IOL test data indicates, most major Linux OSes, including RHEL, CoreOS, SUSE and Ubuntu, provide NVMe support, as do all recent versions of Windows Server (2012, 2016) and the Microsoft Windows 10 client.

NVMe-oF is a critical foundation for composable infrastructure that enables physical hardware components -- servers, storage capacity and network interfaces -- to be logically carved up into virtual instances that are allocated to particular VMs, container clusters and application servers. Aside from being the foundation for next-generation cloud data centers, NVMe-oF products are particularly attractive for analytics, AI training and high-performance computing applications with high levels of storage IO that can take advantage of RDMA to significantly reduce IO overhead on compute servers.

# 5 NVMe-oF Implementation Questions Answered

*Stacey Peterson, Senior Site Editor*

NVMe over fabrics is making its way into the enterprise, enabling organizations to extend NVMe's high-performance, low-latency advantages across their infrastructure.

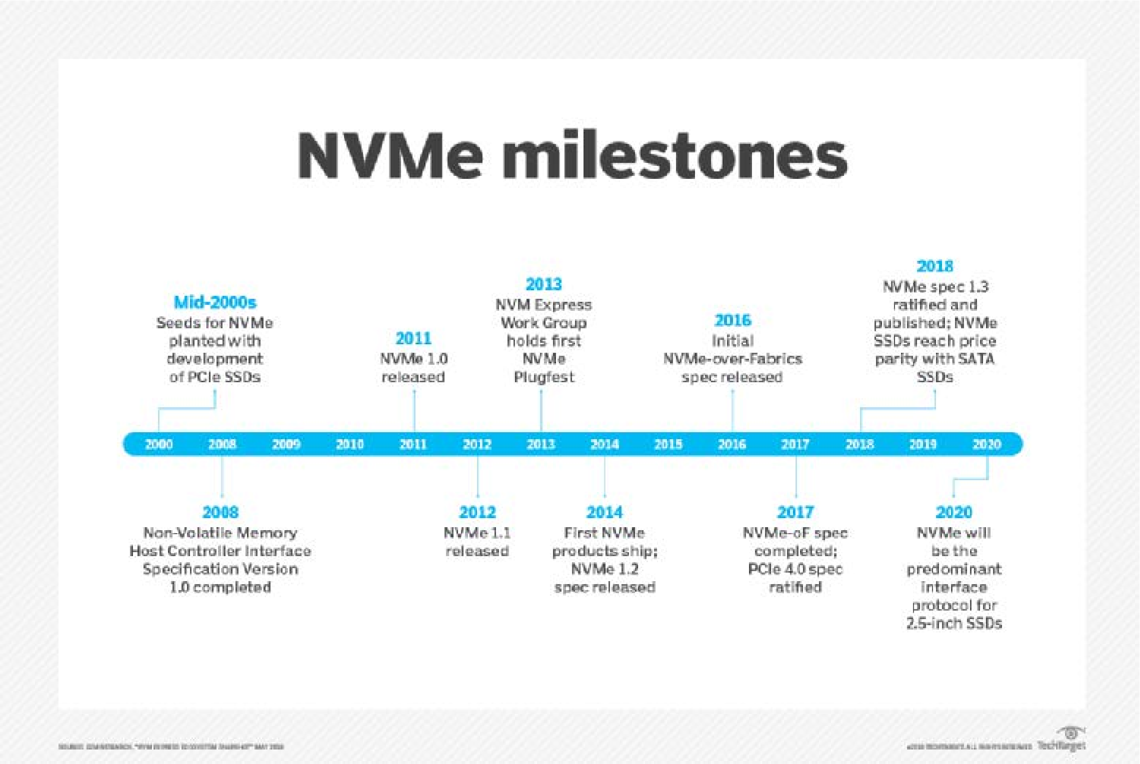
An NVMe-oF implementation maps commands and responses between an NVMe host and shared memory, facilitating communication. It extends the distances over which NVMe devices can be accessed and enables scaling to multiple devices.

For most organizations, the first step into NVMe technology is an NVMe all-flash array using traditional network connections to the storage network. More than 60% of all-flash arrays will use NVMe storage media by 2022, according to G2M Research predictions. From there, a move to NVMe-oF may follow, depending on performance needs. G2M also estimated that more than 30% of AFAs will use NVMe-oF to connect to hosts or expansion shelves in the next three years.

NVMe-oF isn't without its complications and challenges, however. You must gather a lot of information and answer a number of questions before making the move. What follows is a look at five important NVMe-oF implementation questions to answer before deploying this new and critical technology.

# What are the NVMe-oF implementation options?

When the initial NVMe-oF spec came out in 2016, it provided support for Fibre Channel (FC) and remote direct memory access (RDMA) fabrics, with InfiniBand, RDMA over Converged Ethernet and Internet Wide Area RDMA Protocol included under the RDMA umbrella. The NVMe-oF 1.1 spec, released in November, 2018, added TCP as a fabric option.



With FC and RDMA, data can be transferred from a host to an array and back without being copied to memory buffers and I/O cards as must be done with a standard TCP/IP stack. This approach reduces the CPU load and data transfer latency, speeding up storage performance. However, FC and RDMA are more complex than TCP and can require special equipment and configurations when implementing NVMe-oF.

Enterprises can use TCP NVMe-oF transport binding over standard Ethernet networks, as well as the internet, making it simpler to implement. It doesn't require any reconfiguration changes or special equipment, eliminating many of the complications involved in using NVMe over FC or RDMA.

## What factors must you consider when implementing NVMe-oF?

NVMe-oF is designed to use an organization's existing storage fabric, but that assumes the existing storage infrastructure is compatible with NVMe-oF. Some vendors don't strictly adhere to the NVMe-oF standard and, instead, use a proprietary approach that can lead to compatibility issues.

Another area to consider with NVMe-oF implementation is ensuring the storage infrastructure has adequate throughput. If the connectivity between the system's initiator and the target doesn't keep up with the physical NVMe storage device, system latency can increase. It's also important to check that storage system software is up to date. Old device drivers and OS kernels can cause issues.

## Where does NVMe-oF leave rack-scale PCIe?

Before NVMe-oF appeared, it looked like data centers were headed toward rack-scale, switched PCIe fabric. DSSD, a company EMC acquired and then dismantled, developed technology that used the PCIe switch to share a block of flash across hosts in a rack. At the time, DSSD's 100 microsecond (µs) latency compared favorably to all-flash array vendors' 1 millisecond latency.

But then NVMe-oF came on the scene, matching DSSD's 100 µs latency while letting customers use their existing standard Ethernet and FC networks, a development that shelved rack-scale PCIe for the time being. There's room for that to change, though, given the Gen-Z Consortium's focus on developing a rack-scale infrastructure that shares GPU and memory along with storage and I/O cards.

## How does NVMe-oF help with scale-out storage?

Tools to scale-out a storage architecture use extra servers or nodes to increase its performance and capacity. In the past, scale-out tools tended to be complex to implement. But NVMe-oF is changing all that by skirting the limitations and complexities of traditional storage architecture.

NVMe-oF's advantage is in creating a more direct path between the host and the storage media. As a result, data isn't transmitted through a centralized controller. This more direct I/O path reduces the length of the path and enables a single host to talk to many drives and vice versa. This approach reduces storage system latency and increases scale-out capability.

# Do NVMe and NVMe-oF solve all storage performance issues?

NVMe technology greatly enhanced performance between the server or storage controller CPU and attached flash SSDs over SAS and SATA SSDs. And NVMe-oF has reduced latency and improved the performance of shared storage. However, NVMe and NVMe-oF have also exposed another performance challenge: the CPU chokepoint in the server or the storage controller.

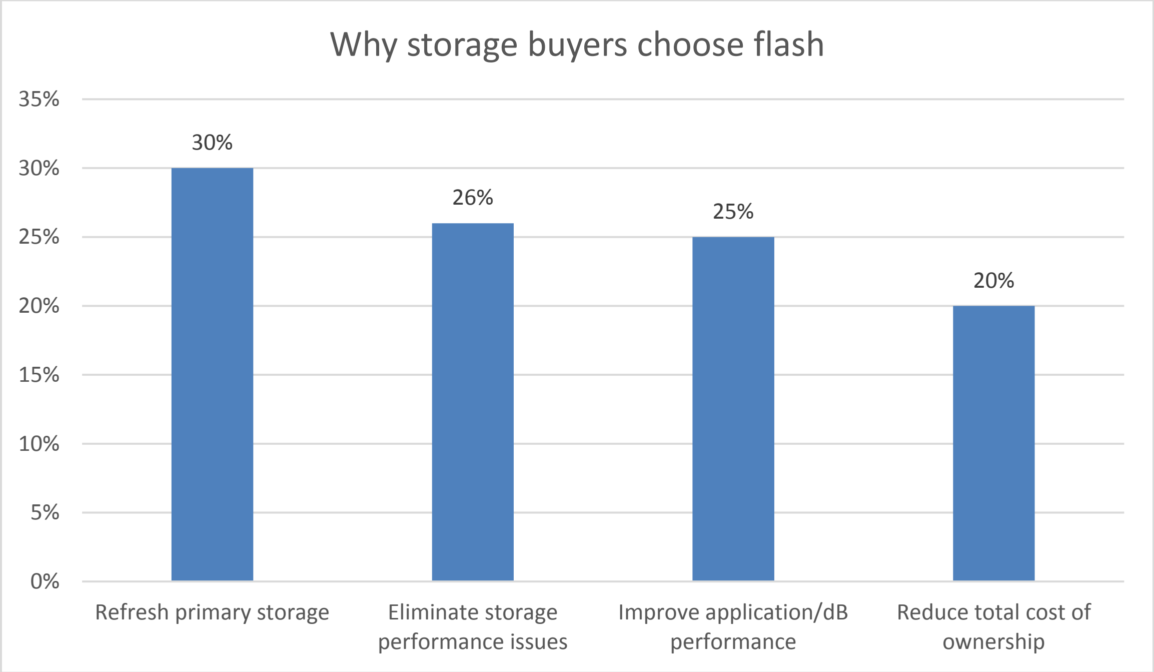
The latest processors support increasing numbers of SSDs. And as more drives are used, the supporting hardware gets increasingly complicated, with more CPUs, drives, drive drawers, switches, adapters, transceivers and cables for IT to cope with and manage. More hardware also results in diminishing marginal returns on overall performance.

The cause of this NVMe-oF implementation performance challenge is storage software not designed for CPU efficiency. One way to address this problem is to rewrite storage software to make it more efficient. Some vendors, such as StorOne, take that approach. Others are trying alternative solutions, such as adding more CPUs, putting dynamic RAM caching in front of NVMe SSDs and using computational storage to put more processors and RAM on NVMe drives.

# Flash Arrays are the New Normal

Over 1,000 storage buyers we surveyed in the first half of 2019 were buying one or another form of flash storage arrays. Their primary reason: to refresh legacy storage.

That demonstrates the long journey flash technology has had from a very expensive, niche solution to the most performance-intensive, transaction-processing loads to, well, an everyday option for new storage. As the chart below shows, the four main drivers of flash array purchases highlight several advantages the technology offers.



NVMe-oF Products are Changing the Face of Data Storage

5 NVMe-oF Implementation Questions Answered

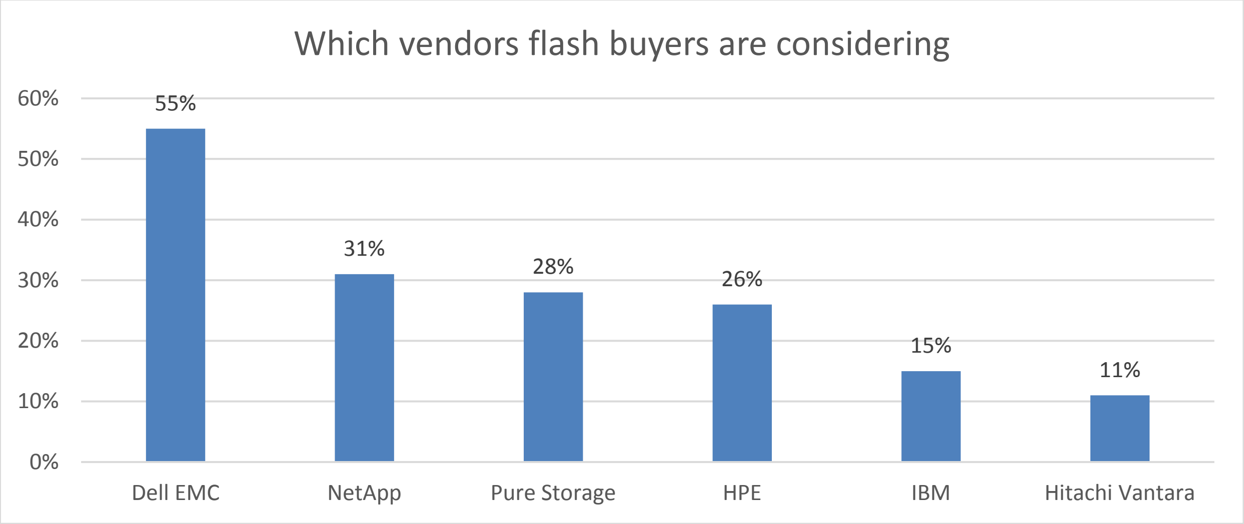
Flash Arrays are the New Normal

These buyers are clearly motivated by the benefits of eliminating spinning disk:

- Higher read and write speeds
- Fewer application bottlenecks
- Less time and money spent on replacing failed hard drives

And while databases are still the most popular workload (54%), storage for virtual servers (47%) and general business applications (38%) are not that far behind. As well, flash continues to be a way to ensure good performance for virtual desktop infrastructure (VDI) (28%).

As with several other forms of enterprise storage, Dell EMC is a vendor most buyers are considering. But as the chart below shows, a closely packed group of NetApp, Pure Storage and HPE is each also getting a look from more than a quarter of buyers.



While Pure Storage is a newer and pure-play flash vendor, the other leading flash array vendors are all familiar names to enterprise storage buyers. That bodes well for the continued penetration of flash technology into on-premises storage as prices continue to decline, technology options multiply, and users work out their plans for combining on-prem and cloud resources.