Technology Insight Paper

Networking The Next Generation of Enterprise Storage: *NVMe Fabrics*

By Russ Fellows

September 2016

Evaluator Group

Enabling you to make the best technology decisions

Executive Summary

Network accessible, shared storage has been a significant factor in the rise of enterprise storage systems that have proliferated over the past twenty years. To many, networked storage is synonymous with SAN or Storage Area Networks particularly for accessing block storage. Originally SAN's were Fibre Channel only, but with the advent of iSCSI and other Ethernet based storage protocols, the definition of a SAN now includes other networking types.

The benefits of shared, networked storage are well understood by IT architects and administrators, often with new storage purchases requiring support for shared network access via Fibre Channel or Ethernet based SAN technologies. With new storage technologies emerging, IT architects are once again exploring how to network the next generation of storage products, without loosing the features they rely upon to provide the always on infrastructure required by enterprise and cloud environments. New, high-speed storage technologies require a network that can support both high data transfer rates and high I/O rates with minimal overhead.

New storage technologies will require an enterprise class network capable of providing shared access to storage, with all the resiliency, security and manageability features that enterprise IT professionals have come to expect. Emerging storage protocols known as NVMe enable next generation non-volatile storage products to deliver performance and economic benefits required by future cloud and enterprise applications.

These next generation storage systems will require new storage networks to support their full potential. In this paper we explore the important characteristics of NVMe storage products, and the storage network options required to support these developing products and technologies.

The Need for a New Storage Interface

Over the past twenty years, the SCSI interface has been the most dominant protocol for accessing storage. Although other protocols exist, most have lost out to SCSI in enterprise environments. At the system level, other interfaces such as Serial ATA are also used. However, as new storage media technologies have emerged such as NAND Flash and others, old storage protocols are becoming a bottleneck, limiting the access speed of the non-volatile storage.

Protocols designed for rotating media were not designed to handle the low latencies possible with memory based storage media, nor were they designed to handle the massive parallelism inherent in today's highly virtualized application environments. Although SCSI and ATA (SAS and SATA) have been adopted to accommodate NAND flash, they inherently add significant overhead. Designed for rotating disk media, SCSI and SATA protocols were architected to operate with 100 microseconds of latency, which was insignificant compared with the latency of rotating media.

The NVMe Interface

A new storage interface is emerging, known as Non Volatile Memory express, or NVMe. It takes its name from the fact that data is stored using memory based technologies such as NAND flash for long term storage. Flash and other storage technologies including 3D-Xpoint and emerging memory technologies provide non-volatile memory at significantly lower prices than current volatile memory such as Dynamic RAM (DRAM). As these technologies mature, a range of performance and cost points that were previously unacceptable for non-volatile storage now make economic sense for the data center.

Today, NAND flash media is capable of delivering data with 100 - 200 microseconds of latency; however, the protocol overhead associated with SATA or SCSI can significantly reduce the effective speed of NAND particularly for latency sensitive applications requiring high I/O and low delay. Next generation memory technologies have access latencies of 10 microseconds or less, which is 10X faster than NAND flash. Clearly, storage access protocols designed 30 years ago for rotating media are no longer appropriate for these new non-volatile memory technologies.

The NVMe interface is designed with following key attributes:

- Support for up to 64K I/O queues with minimal command overhead
- Each I/O queue supports 64K I/O operations
- Each I/O queue is designed for simultaneous multi-threaded processing
- NVMe protocol enables hardware automated queues
- NVMe commands and structures are transferred end-to-end
- The NVMe protocol may be transported across multiple network fabric types

Due to these characteristics, NVMe has significantly higher efficiency and lower latency than SCSI, which when coupled with NVMe's ability to manage more I/O operations further improves performance in highly virtualized environments. As shown in Figures 1 and 2 below, the NVMe protocol is capable of operating with non-volatile memory technology at latencies of 20 microseconds and below. This is data from Intel testing showing data points for access times and I/O rates using NVMe compared to existing SCSI and SATA interfaces.

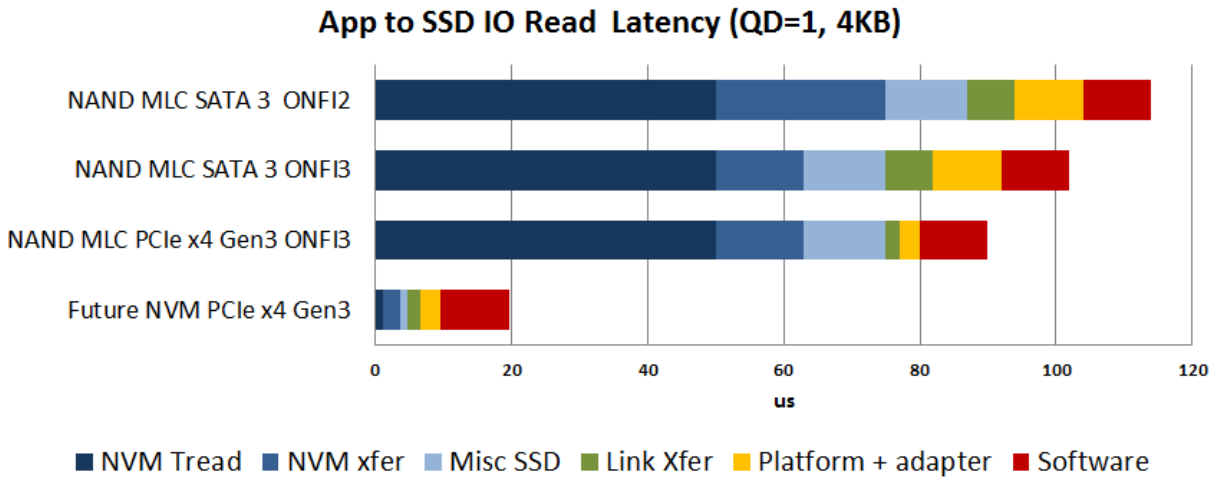

Figure 1: NVMe vs. SATA Transfer Times for NAND media (Source: Intel)

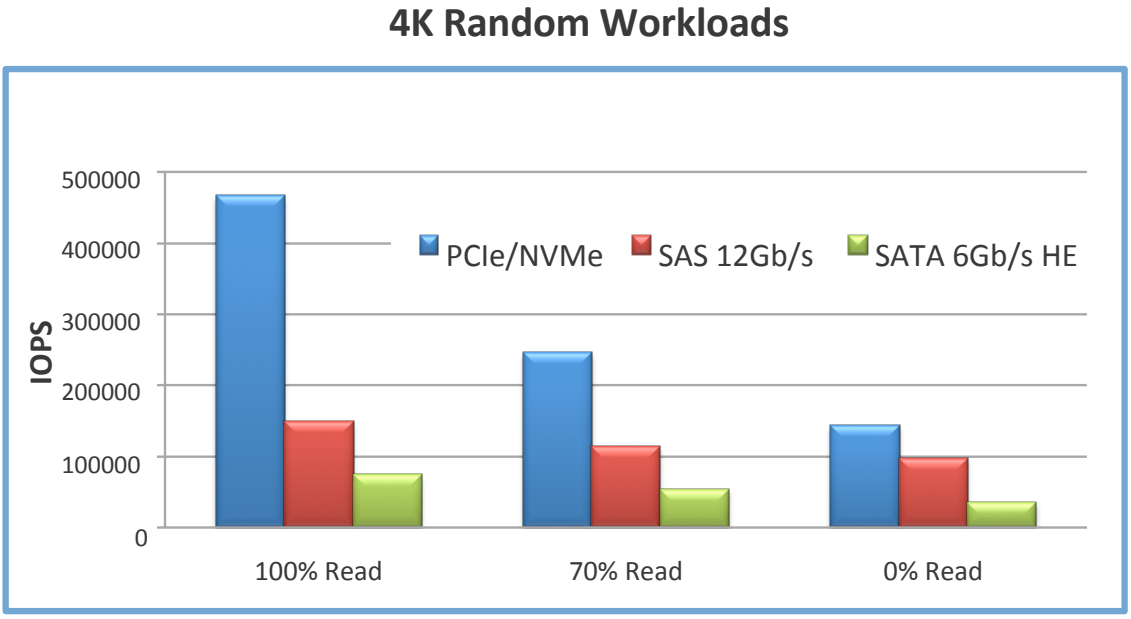

Figure 2: NVMe vs. SAS & SATA I/O Rate for NAND media (Source: Intel)

Evaluator Group comments: The reduced overhead and latency associated with NVMe compared to SCSI and SATA are critical factors for enabling the high I/O rates possible with next generation non-volatile memory technologies used for storage.

Considerations for Next Generation Storage Networks

There are several considerations for storage networking selection in enterprises that may be broadly categorized as either technical or business requirements.

Technical Requirements for NVMe Storage Network

Primary selection considerations include:

- High bandwidth, with speeds greater than 10 Gb/s per connection to servers
- Low latency, with overhead of less than 1 microsecond
- Ability to easily aggregate multiple network links between hosts and switches
- Ability to easily aggregate multiple network links between switches (aka. ISL links)
- High security through use of an air-gap storage network separate from general LAN

Business Requirements for NVMe Storage Network

Selection considerations for business requirements of an NVMe storage network include:

- Backward compatible with existing storage network infrastructure
- Multiple vendor support to ensure continued enhancements and price competition
- Large established base of trained IT professionals to manage the environment
- Management tools and technologies exist to enable efficient management

A Storage Network for NVMe

Recently, the NVMe organization released the first NVM Express over Fabric specification, known as NVMeoF[1]. The options for NVMe access to storage include PCIe, Fibre Channel, and RDMA protocols. Typically, PCIe is used for system connectivity, with shared remote connectivity to NVMe provided by a network or fabric connection. The NVM Express organization has created a specification for NVMe transported over RDMA fabrics, known as NVMeoF. Separately, the Fibre Channel standards group is finalizing the FC specifications to support NVMe over FC fabrics and released the 1.0 spec in June 2016.

Each of the technologies shown in Figure 3 has vendor and technical groups that have shown preliminary support for transporting NVMe over their fabric. Widespread standardization and interoperability is expected to occur over the next several years as NVMe storage options increase.

[1] **NVMeoF Specification: www.nvmexpress.org/wp-content/uploads/NVMe_over_Fabrics_1_0_Gold_20160605.pdf**

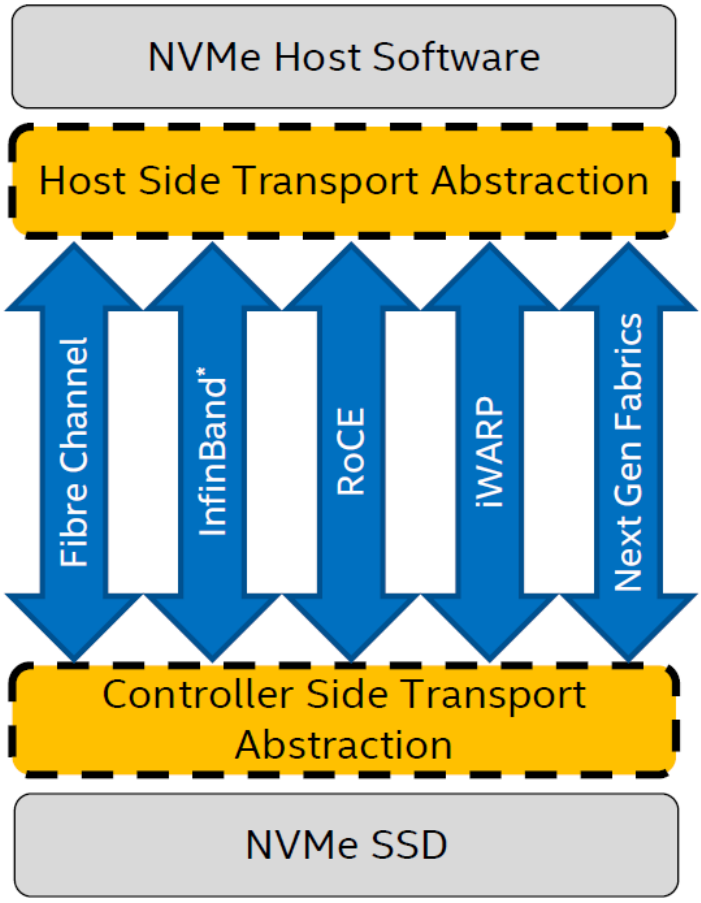

Figure 3: NVMe Fabric Options (Source: NVM Express Organization)

Note: The diagram above from the NVM Express organization includes "Next Gen. Fabrics" as an NVMe transport option. However, this does not refer to a specific technology, but rather the option of additional transport technologies being supported in the future.

> ***Evaluator Group comments: The primary technologies being pursued for NVMe Fabrics are Fibre Channel, InfiniBand and RoCE. Currently iWARP has limited development and there are no standards activities to develop other "Next Generation Fabrics". A product using a PCIe fabric for shared NVMe access exists, but uses a proprietary design and interface.***

As shown in Figure 3, there are multiple options for transporting NVMe over a network or fabric. Each fabric requires specific changes or enhancements to transport NVMe. Standards underway for NVMe over fabrics include:

- NVMe Memory Transports (non-fabric access via PCIe per NVMe specifications)
- Message Transports (Fibre Channel with new ULP's to support NVMe)
- Message / Memory Transport (RDMA protocols includes InfiniBand, RoCE and iWARP)

NVMe over FC

The Fibre Chanel standard is a link level network transport that was designed to support multiple protocols. Due to the flexible design, Fibre Channel only needs new upper layer protocols (ULP's) as defined by the FC standards committee to support NVMe over FC.

Support for new ULP's needed by NVMe do not require changes to Fibre Channel's switching logic, enabling backward compatibility with existing Fibre Channel switches. However, in order to place the NVMe upper layer protocols onto the Fibre Chanel network, new OS and hypervisor device drivers are required along with new Gen 6 HBA's.

Requirements for NVMe over FC fabric

- Fibre Channel Gen 5 and Gen 6 switches supported, full compatibility with SCSI & NVMe over FC
- Generation 6 HBA's with new devices drivers required to support NVMe over fabrics, concurrently along with SCSI

NVMe over RDMA

The class of transports using Remote Direct Memory Access (RDMA) includes InfiniBand, RoCE and iWARP protocols. All of these three options share technology including the commands and in some cases the link or layer-2 transport technology.

The InfiniBand protocol (IB) was standardized in 2001 and utilizes its own transport technology, necessitating host connection adapters (HCA's) along with IB switches for deployment. Currently there are multiple bandwidth options available including 40 Gb/s, 56 Gb/s and 100 Gb/s.

The RoCE protocol has two options, of which the RoCEv2 specification has greater applicability for enterprise environments by supporting routable connections. The RoCEv2 protocol embeds InfiniBand verbs within an Ethernet frame, with the resulting implementation using IB verbs running on an Ethernet transport, along with IP at layer 3 to enable routing.

The iWARP protocol has existed since 2007 and utilizes RDMA techniques for data movement across a network utilizing IP at layer 3 and Ethernet at layer 2. This protocol can utilize Ethernet transport, but requires a hardware NIC to efficiently implement the protocol along with custom device drivers.

Requirements for NVMe over RDMA

- iWARP requires iWARP specific rNIC's and device drivers
- InfiniBand requires both IB HCA's and IB switches
- RoCE requires DCB Ethernet switches, along with driver support in NICs

Storage Network Requirements

In Table 1 below, we compare each of the NVMe Fabric alternatives, evaluating how well they meet the requirements outlined previously.

Consideration	Fibre Channel	RoCE	InfiniBand
100 Gb/s Host Connectivity	Yes - 128 Gb/s	Yes - 100 GB/s	Yes - 128+ Gb/s
Switch ISL Links > 100 Gb/s	Yes - 256 Gb/s	Yes - switch dependent	Yes - to 290 Gb/s
Link Aggregation from switch - to - switch	Built-in since FC Gen 1	Good, with proprietary DCB extensions	Difficult
Link Aggregation from host - to - switch	Built-in since FC Gen 1	Poor, no standard between hosts and switch	Difficult
Air Gap Security	Yes, since FC Gen 1	Possible, requires separate networks	Yes, IB typically is a separate net
Compatibility with Existing SAN	Very high adoption of FC SAN's	Moderate, some adoption of Ethernet SAN's	Low, very limited adoption of IB SAN's
Multiple Vendor Ecosystem	Yes	Yes	No, Single vendor

Table 1: Comparing NVMe over Fabric Standards Options

Evaluator Group comments: Comparing the NVMeoF options that exist currently, it is clear that the two most widely deployed transports, Fibre Channel and Ethernet are well positioned. Although other options remain viable, RDMA over Ethernet such as RoCEv2 and Fibre Channel appear to be the best options for enabling next generation NVMe Fabrics while providing compatibility with existing storage infrastructure already in place.

Evaluator Group Assessment

Questions for Deploying New Storage Technologies

What most technology professionals really want to know with respect to NVMe fabrics are answers to the following questions:

1. When will NVMe storage technologies be widely available?
2. Which NVMe storage fabric option will best meet our requirements?
3. Which options allow flexibility to support current and future storage networking needs?

Recommendations

The industry excitement around NVMe is unmistakable. There is significant vendor activity occurring around multiple aspects of NVMe storage. New companies and products are emerging that leverage NVMe, including NVMe fabric access using proprietary technologies, such as PCIe and other others.

Companies that are early adopters often forgo standards in order to obtain products sooner than their competitors. In order for a technology to become mainstream, enterprise IT relies on standards that provides the interoperability and price competition necessary to deliver economic benefits.

NVMe is an emerging standard, providing advantages with existing NAND Flash and will become even more important for upcoming non-volatile storage media, enabling further performance and efficiency gains. Moreover, NVMe access to storage will emerge over the next several years. Deployments will depend upon individual organizations performance requirements, and where they are within their storage transition cycles.

The choice of an NVMe storage fabric will depend upon several factors. The key factors outlined previously in Table 1 highlight some of the more significant considerations. The correct choice for an individual company will depend upon their existing SAN infrastructure and their willingness to accept risk if the choice is to deploy a new SAN fabric technology. The optimal choice will provide flexibility to support current storage systems, while enabling a transition to new NVMe accessed storage systems in the future.

New storage network purchases should support NVMe fabrics, which include DCB Ethernet, Gen 6 Fibre Channel along with InfiniBand equipment. Although next generation fabrics, such as PCIe fabrics may become an option for NVMe, there are currently only a few proprietary products using PCIe fabrics with very limited deployments. Currently, all implementations utilize custom proprietary technology, making PCIe fabrics a higher risk than other alternatives.

Final Thoughts

It is possible that multiple NVMe fabrics will become viable choices for enterprise IT during the next decade. Existing IB, DCB Ethernet and FC networks have proven value and deliver benefits over other options. Even if one interface emerges as a better option, it is likely that both Fibre Channel and DCB Ethernet will continue with significant deployments for many years beyond 2020.

Despite some vendors' claims, neither Fibre Channel or Ethernet for Storage are dead, nor do enterprises need to rip and replace existing infrastructures using these technologies. For environments using Fibre Channel, their current Generation 5 and new Generation 6 switches will support both SCSI over Fibre Channel, and new NVMe over Fibre Channel.

Enterprise IT users and architects should be aware of the coming wave of NVMe storage devices and subsequent NVMe accessed storage systems. In order to leverage the benefits of NVMe storage systems, enterprises will need to have NVMe capable storage networking infrastructure in place.

In order to ensure that the transition to NVMe enabled storage has minimal impact on the storage networks, IT professionals should begin planning for the transition to NVMe fabric capable network technologies. New storage networking purchases should include Generation 6 FC equipment or DCB capable Ethernet. Ideally, storage fabric deployments should provide a future path for NVMe storage while still meeting the enterprise storage requirements for current applications.

In summary, high-speed NVMe access over a dedicated storage fabric will become an important cornerstone of next generation datacenters and cloud infrastructures. By planning for these changes now, IT professionals can ensure they are not misled by claims of the need to immediately transition to new, storage fabrics or claims of an existing technology's demise.

About Evaluator Group

Evaluator Group Inc. is a technology research and advisory company covering Information Management, Storage and Systems. Executives and IT Managers use us daily to make informed decisions to architect and purchase systems supporting their digital data. We get beyond the technology landscape by defining requirements and knowing the products in-depth along with the intricacies that dictate long-term successful strategies. ***www.evaluatorgroup.com*** *@evaluator_group*