MARVELL

White Paper

FC-NVMe (NVMe over Fibre Channel)

QLogic® Enhanced 16GFC / 32GFC / 64GFC HBAs
Concurrent FCP (FC-SCSI) and FC-NVMe (NVMe/FC)

May 2023

Background and Summary

Back in 1956, the world's first hard disk drive (HDD) shipped, setting a path for subsequent generations of drives with faster spinning media and increasing SAS speeds. Then in the early 1990s, various manufacturers introduced storage devices known today as flash-based or dynamic random access memory (DRAM) based solid state disks (SSDs). The SSDs had no moving (mechanical) components, which allowed them to deliver lower latency and significantly faster access times. HDDs and SSDs have evolved, along with new and faster bus architectures such as PCI Express (PCIe) which have helped to further improve access speeds and reduce latency in conjunction with the Non Volatile Memory Express (NVMe) standard and the ensuing products.

Fibre Channel (FC) is a high-speed network technology primarily used to connect enterprise servers to HDD- or SSD-based data storage. 16GFC and 32GFC are the dominant speeds today (64GFC HBAs are being introduced and the industry has a strong roadmap to 128GFC and beyond). Fibre Channel is standardized in the T11 Technical Committee of the International Committee for Information Technology Standards (INCITS) and has remained the dominant protocol to access shared storage for many decades. Fibre Channel Protocol (FCP) is a transport protocol that predominantly transports SCSI commands over Fibre Channel networks. With the advent of NVMe, FC has transformed to natively transport NVMe and this technological capability is called FC-NVMe.

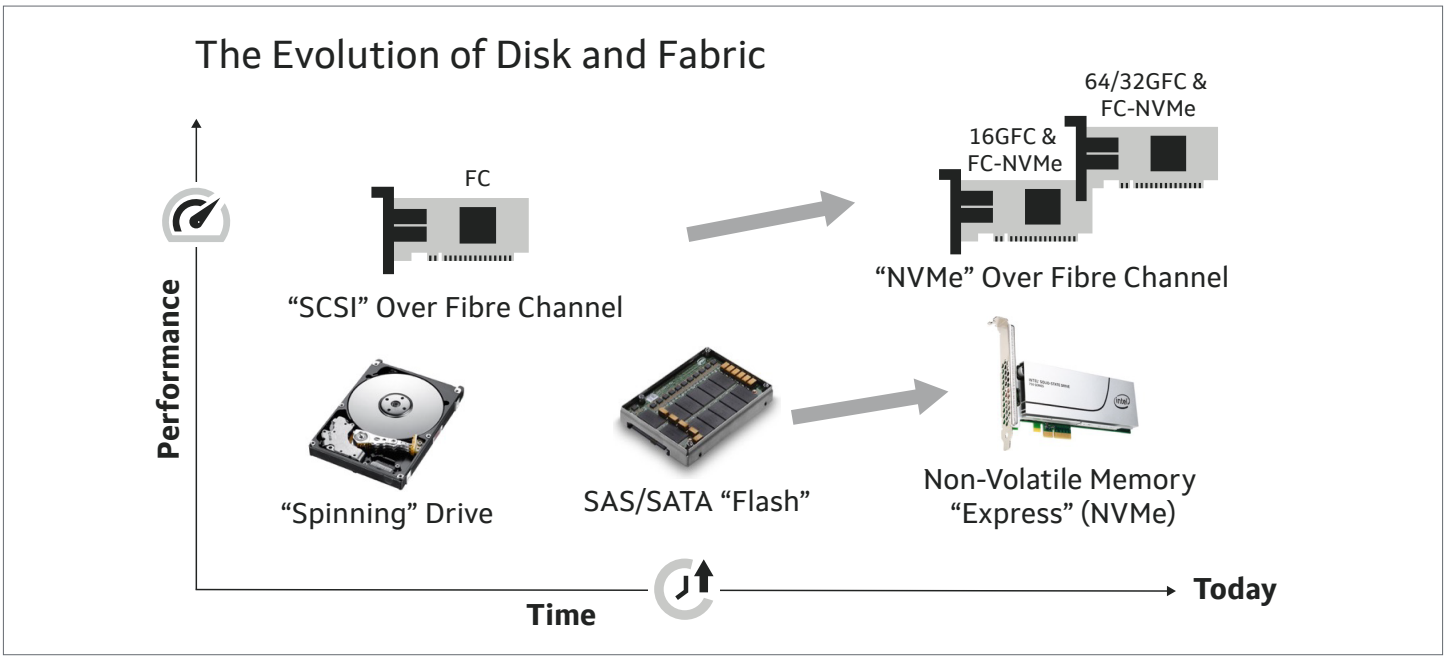

Figure 1: Evolution of Disk and Fibre Channel Fabric

SCSI and NVMe Differences

While the SCSI/AHCI interface comes with the benefit of wide software compatibility, it cannot deliver optimal performance when used with SSDs connected via the PCIe bus. As a logical interface, SCSI was developed when the purpose was to connect the CPU/memory subsystem with a much slower storage subsystem based on rotating magnetic media. As a result, SCSI introduces certain inefficiencies when used with SSD devices, which behave much more like DRAM than like spinning media.

The NVMe device interface has been designed from the ground up, capitalizing on the low latency and parallelism of PCIe SSDs, and complementing the parallelism of contemporary CPUs, platforms and applications. At a high level, the basic advantages of NVMe over AHCI relate to its ability to exploit parallelism in host hardware and software, manifested by the differences in command queue depths, efficiency of interrupt processing, and the number of un-cacheable register accesses, resulting in significant performance improvements across a variety of dimensions.

Table 1 below summarizes high-level differences between the NVMe and SCSI logical device interfaces. Figure 3 shows how the Linux storage stack is simplified when using NVMe. All of these purpose-built features bring out the most efficient access method for interacting with NVMe devices.

Features	Legacy Interface (AHCI)	NVMe
Maximum command queues	1	65536
Maximum queue depth	32 commands per queue	65536 commands per queue
Un-cacheable register accesses (2000 cycles each)	4 per Command	6Register Accesses are cacheable
MSI-X	A Single Interrupt	Register Accesses are cacheable
Interrupt steering	No steering	Interrupt steering
Efficiency for 4KB commands	Command parameters require two serialized host DRAM fetches	Gets command parameters in on 64-byte fetch
Parallelism and multiple threads	Requires synchronization lock to issue a command	Does not require synchronization

Table 1: Efficiency and Parallelism Related Feature Comparison Between SCSI and NVMe

NVMe Deep Dive

NVMe is a standardized high-performance host controller interface for PCIe storage, such as PCIe SSDs. The interface is defined in a scalable fashion such that it can support the needs of enterprise and client applications in a flexible way. NVMe has been developed by an industry consortium, the NVM Express Workgroup. Version 1.0 of the interface specification was released on March 1, 2011 and continued to evolve into version 1.4 which released 2019. Today, over 100 companies, including Marvell, participate in defining the interface. In July, 2021 a restructured NVMe 2.0 specifications was released to enable faster and simpler development of NVMe technology, supporting the seamless deployment of flash-based solutions in many emerging market segments. The NVMe 2.0 specifications include evolutionary new features like Zoned Namespaces (ZNS), Key Value (KV), Rotational Media and Endurance Group Management.

The NVMe interface is:

- Architected from the ground up for this and next generation non-volatile memory to address enterprise and client system needs
- Developed by an open industry consortium
- Architected for on-motherboard PCIe connectivity
- Designed to capitalize on multi-channel memory access with scalable port width and scalable link speed

NVMe is designed from the ground up to deliver high bandwidth and low latency storage access for current and future NVM technologies. The NVM Express standards include:

- NVM Express (NVMe) Specification – The register interface and command set for PCI Express technology attached storage with industry standard software available for numerous operating systems. NVMe is widely considered the de facto industry standard for PCIe SSDs.
- NVMe Management Interface (NVMe-MI) Specification – The command set and architecture for out of band management of NVM Express storage (e.g., discovering, monitoring, and updating NVMe devices using a BMC).
- NVMe over Fabrics (NVMe-oF) Specification – The extension to NVM Express that enables tunneling the NVM Express command set over additional transports beyond PCIe architecture. NVMe over Fabrics technology (e.g. FC-NVMe) extends the benefits of efficient storage architecture at scale in the world's largest data centers by allowing the same protocol to extend over various networked interfaces.

As a result of the simplicity, parallelism and efficiency of NVMe, it delivers significant performance gains over Serial Attached SCSI (SAS).

NVMe over Fabrics

Most of the NVMe in use today is held captive in the system/server in which it is installed, but this is expected to change rapidly in the next few years with the introduction of NVMe based External Storage Arrays. While there are a few storage vendors offering NVMe arrays on the market today, the vast majority of enterprise data center and mid-market customers are still using traditional storage area networks, running SCSI protocol over either Fibre Channel or Ethernet Storage Area Networks (SAN).

NVMe-oF will offer users a choice of transport protocols. Today, there are three standard protocols that will likely make significant headway into the marketplace. These include:

- NVMe over Fibre Channel (FC-NVMe or NVMe/FC)
- NVMe over RoCE RDMA (NVMe/RoCE)
- NVMe over TCP (NVMe/TCP)

NVMe over Fabrics is an end to end protocol and its successful deployment centers around an ecosystem that comprises of three keyparts. First, users will need an NVMe-capable storage network infrastructure in place. Second, all of the major operating system (O/S) vendors will need to provide support for NVMe-oF. Third, customers will need disk array systems that feature native NVMe. FC-NVMe is by far the most advanced in terms of meeting these requirements.

Approach	Description	Implementation Notes
NVMe over Fibre Channel Fabric **(FC-NVMe)**	Encapsulate NVMe commands within FC frames and route through FC fabric	No FC switch changes required if 16GFC or above. FW/driver changes required for host HBA and target storage. Secure, scalable connectivity for shared storage.
NVMe over Ethernet Fabric with RoCE **(NVMe/RoCE)**	Encapsulate NVMe with RDMA (and DCB) in Ethernet to route through the network	Leverages lossless Ethernet fabric which may require Ethernet network upgrade.
NVMe over TCP/IP **(NVMe/TCP)**	NVMe over standard TCP/IP (without RDMA or DCB)	Ratified November 2018. Runs on existing Ethernet infrastructure with no changes.

Fibre Channel (FC-NVMe)

Fibre Channel, more specifically “Fibre Channel Protocol (FCP)” has been the dominant protocol used to connect servers with remote shared storage comprising of HDDs and SSDs. FCP transports SCSI commands encapsulated into the Fibre Channel frame and is one of most reliable and trusted networks in the data center for accessing SCSI-based storage. While FCP can be used to access remote shared NVMe-based storage, such a mechanism requires the interpretation and translation of the SCSI commands encapsulated and transported by FCP into NVMe commands that can be processed by the NVMe storage array. This translation and interpretation can impose performance penalties when accessing NVMe storage and in turn negates the benefits of efficiency and simplicity of NVMe.

NVMe-oF solutions, such as RoCEv2, require every switch in the path to have complex configuration settings for enablement of lossless traffic and congestion handling. Additionally, server NICs might need to be updated to RDMA enabled NICs and along with this, server administrators’ expertise on RDMA/OFED is needed. If customers have to choose an Ethernet based fabric to scale out NVMe, NVMe/TCP provides a simpler approach, but has higher latency and consumes more CPU cycles than Fibre Channel due to lack of offloads (Figure 2).

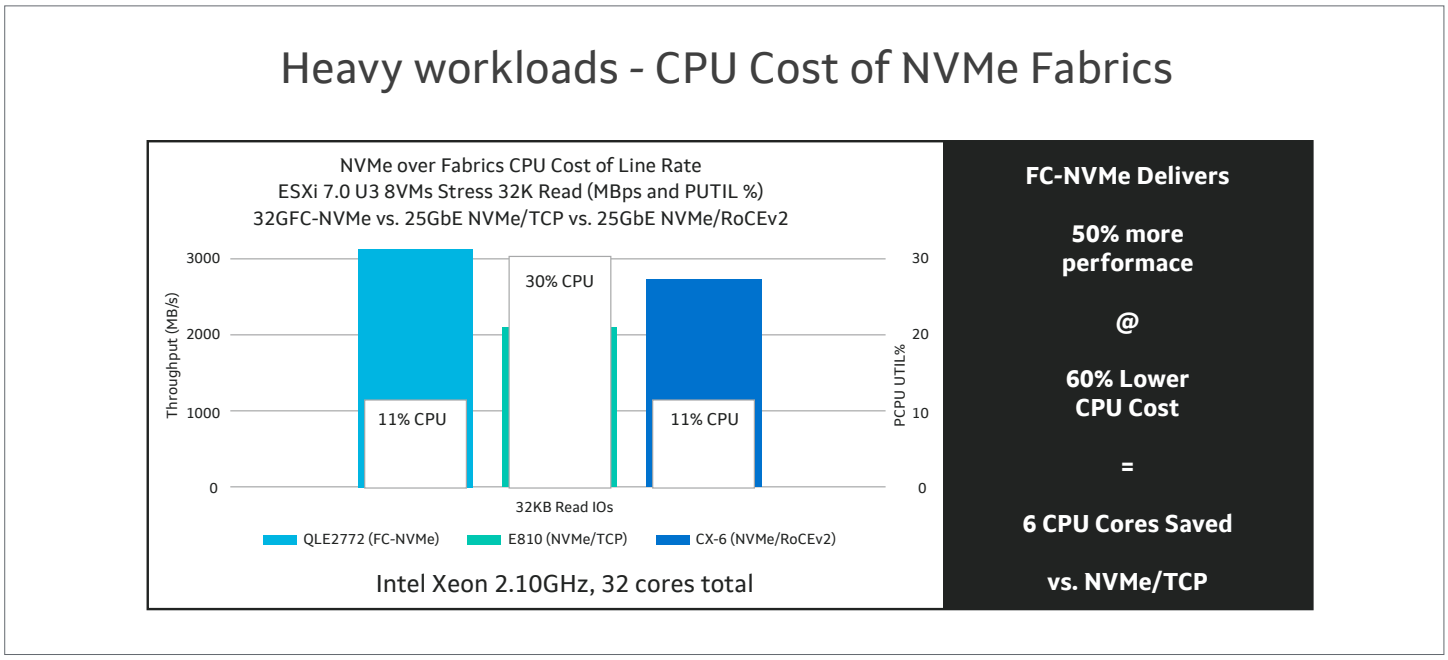

Figure 2: Heavy workloads - CPU Cost of NVMe Fabrics

FC-NVMe extends the simplicity, efficiency and end-to-end NVMe model where NVMe commands and structures are transferred end-to-end, requiring no translations. Fibre Channel's inherent multi-queue capability, parallelism, deep queues, and battle-hardened reliability make it an ideal transport for NVMe across the fabric. FC-NVMe implementations will be backward compatible with Fibre Channel Protocol (FCP) transporting SCSI so a single FC-NVMe adapter will support both SCSI-based HDDs and SSDs, as well as NVMe-based PCIe SSDs. No changes to the FC switching infrastructure are required to support FC-NVMe.

Test results based on a configuration from the same initiator server to two different target systems, running the similar test commands adjusting for the different target devices as they are presented. One target system had NVMe SSD capable devices populated, while the other target system had SSD devices presented across its SAS interface.

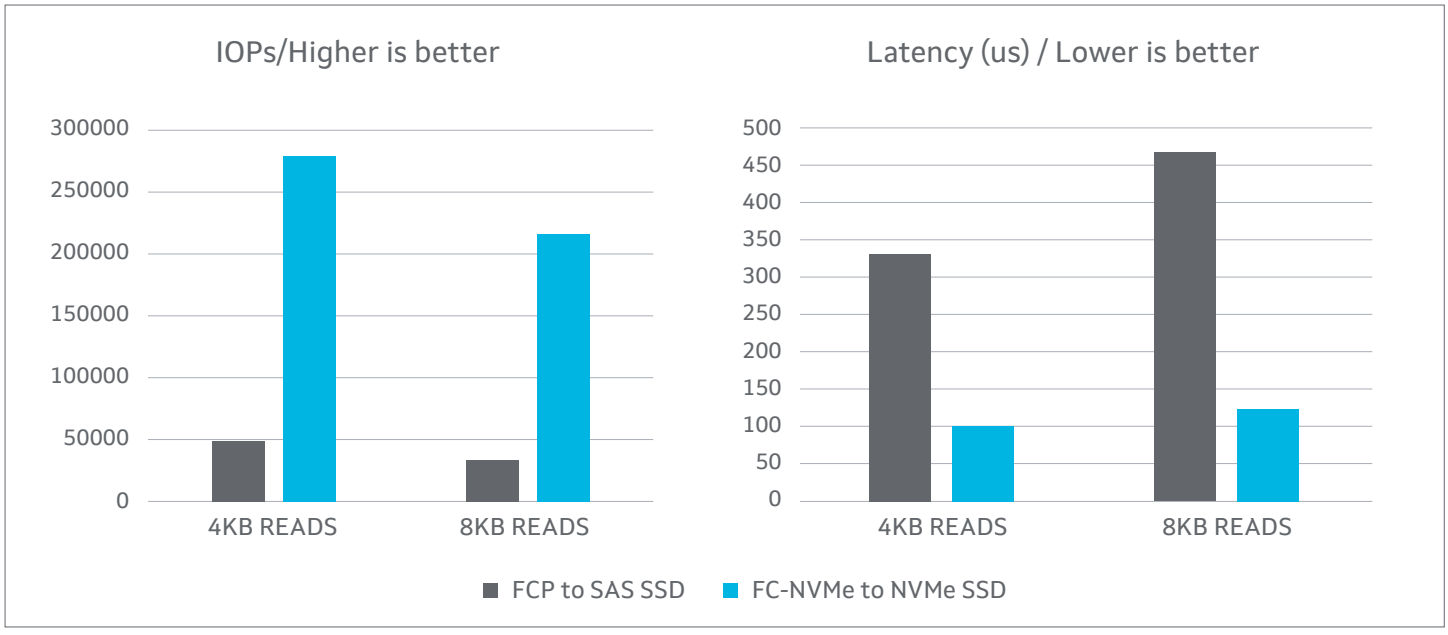

Figure 3. FC-NVMe Advantages: FC-NVMe to NVMe vs FCP to SAS

Marvell and FC-NVMe

Marvell is a global leader and technology innovator in high-performance server and storage networking connectivity products and leads the Fibre Channel adapter market having shipped over 2+ million ports to customers worldwide. As current and future data-intensive workloads transition to utilizing low latency NVMe flash-based storage to meet ever increasing user demands, Marvell has combined the lossless, highly deterministic nature of Fibre Channel with NVMe. FC-NVMe targets the performance, application response time, and scalability needed for today's and next generation data centers, while leveraging current Fibre Channel infrastructures. Marvell is continually pioneering this effort with industry leaders, which over time, will improve and yield significant operational benefits to data center operators and IT managers.

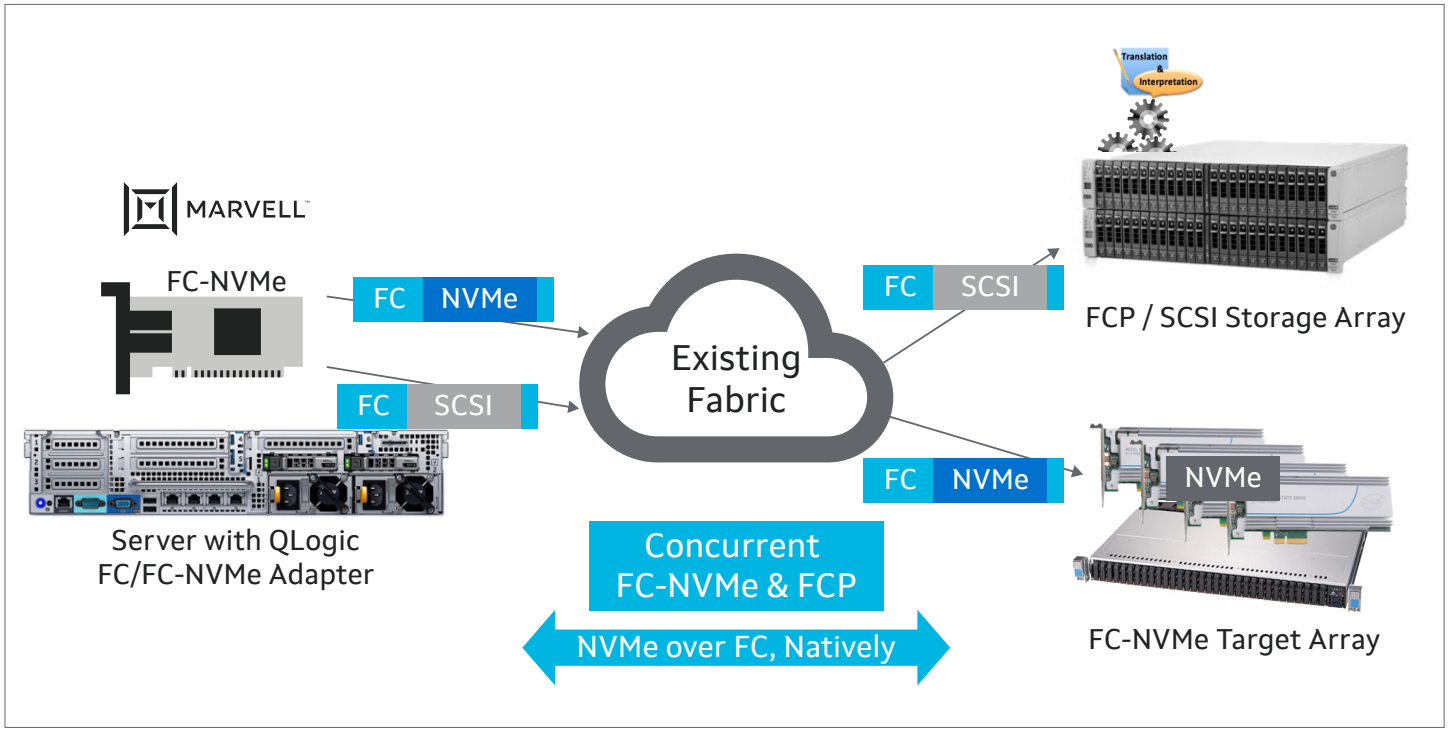

Figure 4. FC-NVMe and Full Backward Compatibility with FCP

The Marvell® QLogic® FC-NVMe technology is aimed at providing a foundation for lower latency and increased performance, while providing improved fabric integration for flash-based storage. Figure 5 depicts how for highly virtualized use cases the Marvell FC HBAs can deliver greater than half a million IOPS at key block sizes used by enterprise applications. These results are based on head to head performance benchmarks conducted in Marvell internal labs.

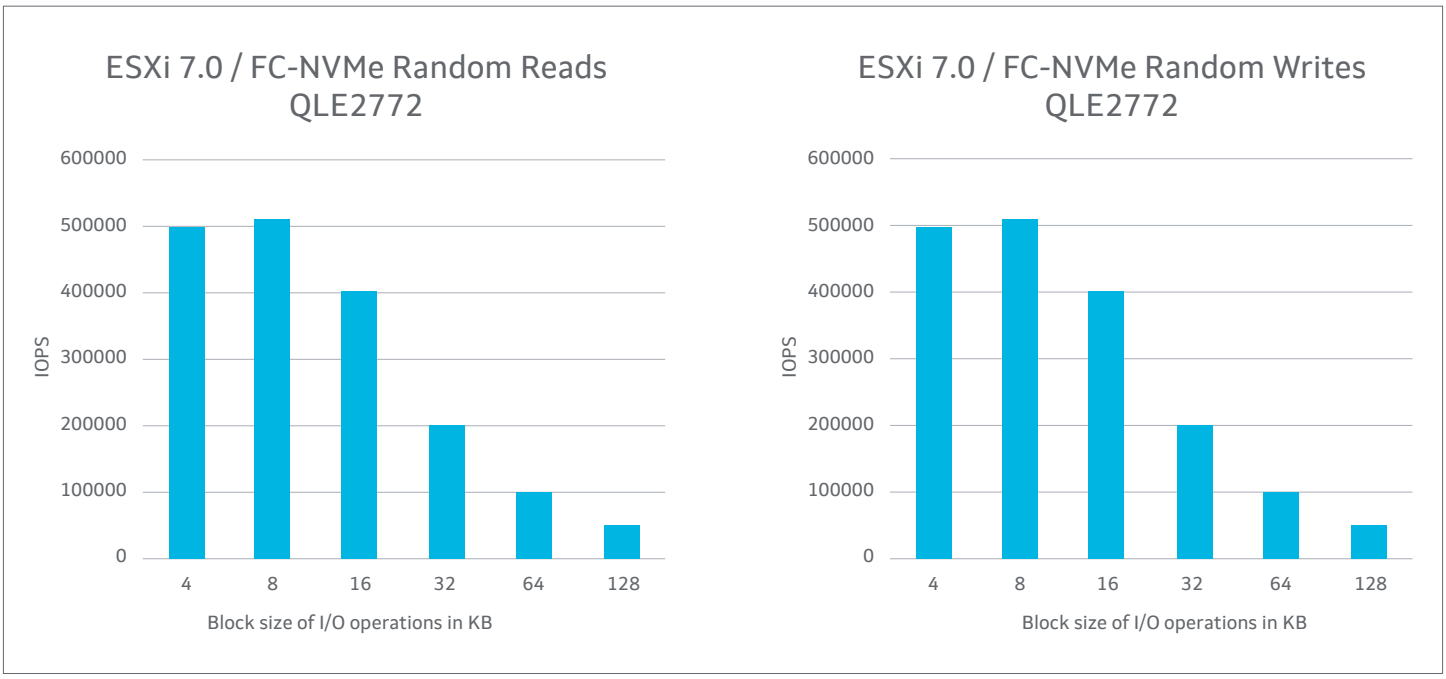

Figure 5: Leading transactional performance with FC-NVMe in vSphere 7.0

Marvell, Brocade, and Cisco are collaborating to drive the standards and the NVM Express over Fabrics Specification 1.0 covering FC-NVMe became a standard released in June 2016 and the T11 Committee ratified the first version of the FC-NVMe standard in August 2017. With leadership from Marvell within the T11 Technical Committee, an update to the standard was published as part of FC-NVMe v2.

FC-NVMe v2

Intended to detect and recover from errors in a manner befitting a low-latency NVMe transport, FC-NVMe v2 (standardized by the T11 committee in August 2020) does not rely on the SCSI or NVMe layer error recovery. Rather, it automatically implements low-level error recovery mechanisms in the Fibre Channel's link layer – solutions that work up to 30x faster than previous methods. These new and enhanced mechanisms include:

- FLUSH: A new FC-NVMe link service that can quickly determine if a sent frame does not quickly reach its destination. Method of Operation: if two seconds pass without the QLogic FC HBA getting a response back regarding a transmitted frame, it sends a FLUSH to the same destination. If the FLUSH gets to the destination, it is determined that the original frame went missing en route, and the stack does not need to wait the typical 60 seconds to detect a missing frame.
- RED: Another new FC-NVMe link service, called Responder Error Detected (RED), essentially does the same lost frame detection but in the other direction. If a receiver knows it was supposed to get something but did not, it quickly sends out a RED rather than waiting on the slower, upper-layer protocols to detect the loss.
- NVMe_SR: Once either FLUSH or RED detects a lost frame, NVMe_SR (NVMe Retransmit) kicks in, and enables the re-transmission of whatever got lost the first time.

Marvell QLogic 2800, 2700 and 2690 Series of HBAs support FC-NVMe v2 (software update may be required), FC-NVMe v2 requires support in storage arrays - please check with your manufacturer.

QLogic HBAs with FC-NVMe

Marvell QLogic 2800, 2700 and 2690 Series HBAs support concurrent FCP and FC-NVMe across multiple operating systems including Microsoft Windows, RHEL, SLES and VMware ESXi 7.0 onwards. Customers can deploy FC-NVMe with confidence across a broad set of heterogeneous environments. For a complete list of FC-NVMe capable storage array compatible with the Marvell QLogic FC HBAs, look for the Interop Matrix on Marvell.com

Adapter Family	Part Numbers	FC Speeds	PCIe Interface	Number of Ports	FC-NVMe
2690 Series Enhanced 16GFC	QLE2690 QLE2692 QLE2694	16/8/4	PCIe 3.0	1, 2, 4	✓
2770 Series Enhanced 32GFC	QLE2770 QLE2772 QLE2774	32/16/8	PCIe 4.0	1, 2, 4	✓
2870 Series 64GFC	QLE2870 QLE2872 QLE2874	64/32/16	PCIe 4.0	1,2,4	✓

Table 2: Complete Marvell QLogic FC-NVMe & FC product line