NVMe over Fabrics

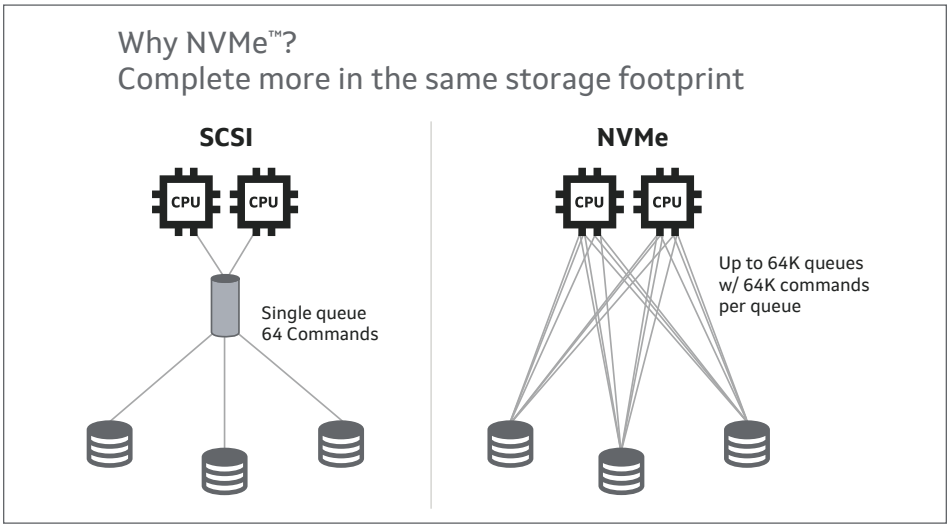
Use case driven strategies for choosing between FC-NVMe, NVMe/TCP, NVMe/RoCE

Non-Volatile Memory Express™, (NVMe) is not new. NVMe is a host controller interface, captive to the host system designed specifically to talk to memory devices and NVMe drives that are commonplace now in servers. This includes solid state or memory-class storage which deliver better performance for server workloads and applications when compared to mechanical hard disk drives. Because NVMe is a host controller interface, connectivity for NVMe storage has been limited to within the physical server or storage array.

For connecting servers to shared storage, storage networks today utilize either Fibre Channel or iSCSI storage network connectivity. In both cases, SCSI commands are transmitted between devices. SCSI has been used for decades to allow computers to talk to devices like hard drives, tape drives, scanners and even printers in the early days.

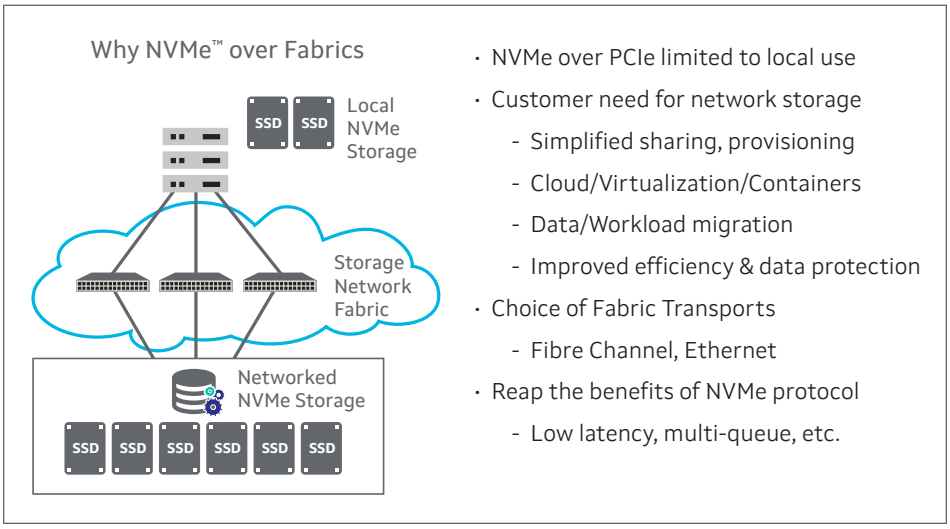
SCSI is a single queue command language, which means all devices must communicate using the same queue. On the other hand, NVMe is a language that was designed to communicate only with memory devices, so it was designed with multi-queue capability. This makes NVMe ideal for today's storage applications as it supports 64K command queues with 64K commands per queue, compared to SCSI's single queue and 64 command structure. Users can expect workloads to run faster with NVMe because many storage queries can be done in parallel.

NVMe is also much more efficient than SCSI with only 13 required commands compared to hundreds for SCSI. This, and the fact NVMe is designed to talk to memory-based storage, greatly reduces latency compared to SCSI implementations. Combining all these advantages, this means more jobs can be performed concurrently, allowing more work to get done in a shorter amount of time.



With the costs of NVMe storage declining, system architects and administrators want to take advantage of low-latency NVMe connectivity beyond the captive server. They want NVMe commands to flow from server to server across the network for software-defined storage or HCI applications, and to also flow from servers to next-generation NVMe Flash Storage Arrays, across a low-latency storage fabric. They want to be able to simplify sharing, provision storage, support cloud-native applications, support server and storage virtualization, as well as support container environments. To do this with NVMe storage, a new kind of storage fabric is required.

The NVMe over Fabric (NVMe-oF) standard was developed to address this challenge. NVMe-oF enables NVMe commands to be encapsulated and sent across storage networks, including FC SANs and Ethernet networks. This new storage fabric is being rolled out in a variety of flavors using different transports. The Fibre Channel standard was enhanced to support NVMe over Fibre Channel or FC-NVMe. Likewise, Ethernet is leveraged for transporting NVMe commands in the network using low-latency RDMA enabled network adapters or RNICs. The third, and latest iteration to be introduced, is running NVMe over standard TCP/IP Ethernet networks.



Compare and Contrast Storage Protocols for NVMe over Fabrics

Fibre Channel is the mainstay transport protocol used in the data center for connecting servers to shared storage arrays. As next generation All Flash arrays are developed, using NVMe drives, it is only natural to extend Fibre Channel to be used for transporting NVMe storage commands, just like it is used for transmitting SCSI commands. This is called FC-NVMe or sometimes NVMe/FC.

With FC-NVMe, host and target adapters encapsulate NVMe commands and responses into Fibre Channel frames. Those frames are then transmitted across the FC SAN using FC switches and host/target adapters. FC-NVMe drivers were developed as well as native support for FC-NVMe drivers in key operating systems like Linux®, Microsoft Windows® Server and VMware® ESXi.

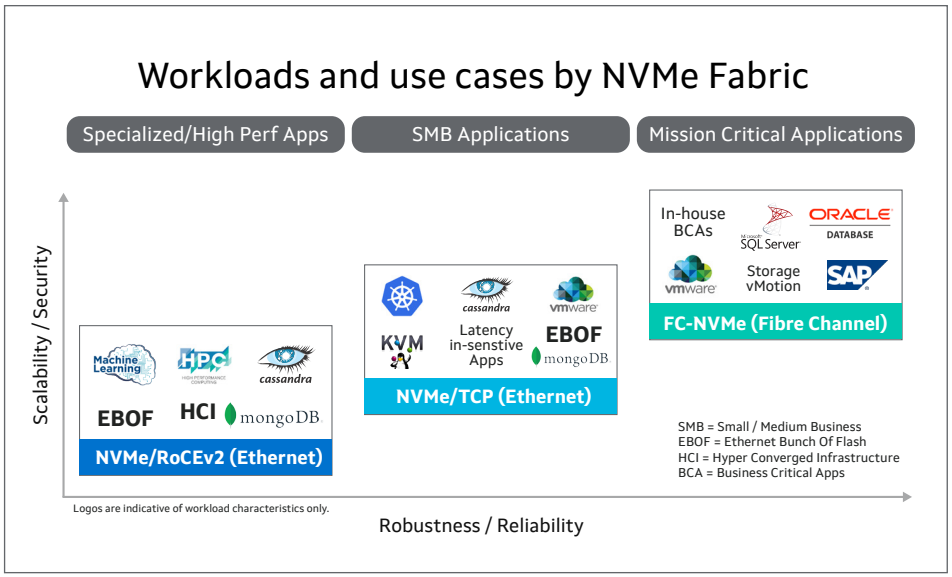
FC-NVMe is supported by most newer QLogic FC HBAs including the Enhanced 16GFC, Enhanced 32GFC and 64GFC adapters. It should be noted that there are no changes required to the Fibre Channel switch infrastructure, as FC switches are agnostic to what is within the payload of the FC frames (SCSI or NVMe). They are simply responsible for transmitting the FC frame appropriately throughout the SAN.

The benefit here is that FC-NVMe gets to take advantage of all the inherent features of Fibre Channel fabrics including high bandwidth, low latency, security, and manageability. The only difference between today's Fibre Channel SAN and FC-NVMe is that in addition to transmitting SCSI commands between servers and storage, NVMe commands can also be transmitted. This results in a great solution for customers who already have FC SANs deployed to enable them to connect their servers to NVMe storage arrays.

For IT architects that want to leverage NVMe fabrics, but don't have Fibre Channel SANs, the choice between Fibre Channel and Ethernet should be based on the mission and business critical nature of the application being hosted on NVMe storage. Typically Fibre Channel SANs are chosen for connecting enterprise class applications like databases, ERP systems and highly virtualized servers. If not, they will likely opt to use Ethernet as the transport. NVMe over RDMA fabrics is an Ethernet solution that was developed around the same time as FC-NVMe. NVMe over RDMA is available today, with RDMA over Converged Ethernet (RoCE) being the RDMA protocol of choice due to its low-latency characteristics. This type of fabric is called NVMe over RoCE v2 or NVMe/RoCE and requires RDMA-enabled NICs, referred to as RNICs, that support the RoCE RDMA protocol.

To be deployed in the data center, NVMe/RoCE requires a few potential changes to the Ethernet network fabric to work. First, the server needs to be configured with RNICs capable of supporting RoCE RDMA. Because RDMA is an IO offload technology, IO processing is done more efficiently in the adapter and requires minimal CPU resources. This reduces the time it takes to process each IO command, reducing latency. Second, to deliver ultra-low latency over Ethernet, a lossless Ethernet network is required. This can be complex to deploy as it requires advanced networking features like Data Center Bridging (DCB), Priority Flow Control (PFC) and Explicit Congestion Notification (ECN) to be enabled. In addition, this type of Ethernet network does not scale well, so a best practice is to limit the configuration to no more than two hops in the network. Thus, adoption of this protocol will likely be limited to small scale applications like DAS extensions or AI that require a server or small cluster to access a large volume of storage with high performance/low latency connectivity. Note there are some proprietary implementations of NVMe/RoCE that do not require the lossless Ethernet network, but these offerings lock the user into a single vendor for both RNIC and switching infrastructure.

To address support for using standard NIC cards and to deliver high scalability, the NVMe group developed a third approach for NVMe fabric connectivity called NVMe over TCP (NVMe/TCP). TCP/IP is the networking protocol that enables us to connect to the internet. It runs on standard NIC adapters (and RNICs too), is highly scalable and used by virtually every server from the edge to the cloud. The pervasiveness provides seamless deployment of NVMe/TCP using software initiators in the host and target storage devices.



With NVMe/TCP, software initiators are utilized to provide the command and control for encapsulating the NVMe commands into the TCP packets that are sent across the network. Standard TCP/IP congestion control is used to route the packets from the source to the destination. That means virtually any modern NIC with bandwidth of 10GbE or more can support NVMe/TCP. No RDMA is required, but because software initiators are used, the CPU and operating system (OS) stack are involved in the transmission and encapsulation of every single NVMe/TCP command, which can greatly reduce the overall performance and efficiency of the system. See Chart Figure 4 which compares the performance of various NVMe over Fabrics in a VMware ESXi deployment and note how stark the CPU Utilization difference is between FC-NVMe which is fully offloaded and NVMe/TCP which utilized server CPU cycles to process I/O.

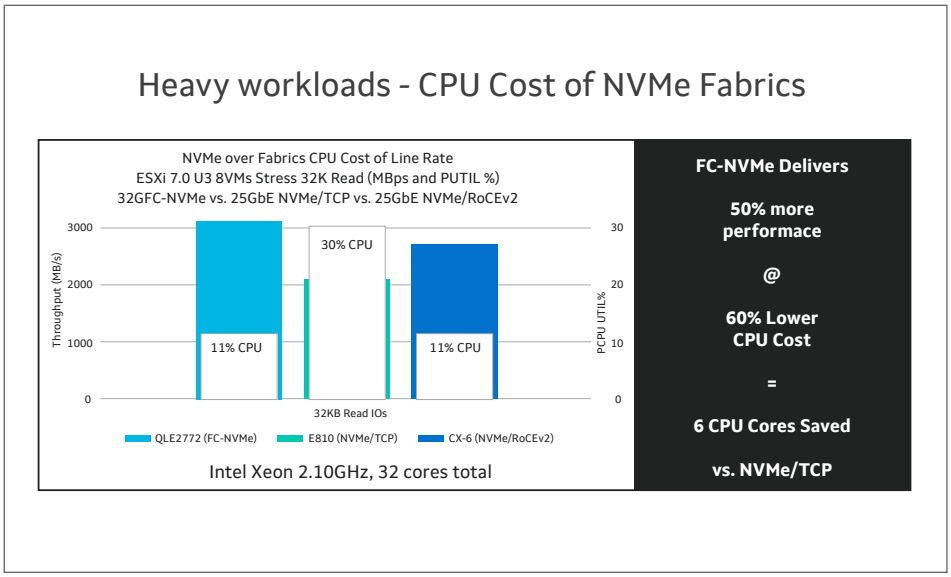


Figure 4: Heavy workloads - CPU Cost of NVMe Fabrics

Below is a table that compares the three NVMe -oF approaches.

Table 1: Characteristics of NVMe over Fabric Networks

	FC-NVMe	NVMe/RoCE	NVMe/TCP
Bandwidth	16-32GFC	10-100GbE	10-100GbE
IOPS	~1M	~1.5M	~1.5M
Network Type	Fibre Channel SAN	Lossless Ethernet	Ethernet Network
Scalability	Yes, dedicated fabric	Limited – 1 to 2 hops	Yes
Adapter Latency	Low	Very Low	High
OS Support	Linux, VMware, Windows	Linux, VMware	Linux, VMware
Security	Yes, dedicated FC network	No	No

Looking ahead, we see a bright future for NVMe over Fabric. Those using Fibre Channel today can almost seamlessly transition to FC-NVMe due to its high performance and security and reliability. Those using Ethernet can in the leverage the simplicity of NVMe/TCP albeit with reduced CPU efficiency and higher latency, but overall lower cost.

Marvell offers QLogic® 269x/27xx/28xx series Fibre Channel HBAs support FC-NVMe today with both Brocade® and Cisco® Fibre Channel switches. For more information on FC-NVMe visit https://www.marvell.com/products/fibre-channel-adapters-and-controllers/fc-nvme.html

MARVELL

To deliver the data infrastructure technology that connects the world, we're building solutions on the most powerful foundation: our partnerships with our customers. Trusted by the world's leading technology companies for 25 years, we move, store, process and secure the world's data with semiconductor solutions designed for our customers' current needs and future ambitions. Through a process of deep collaboration and transparency, we're ultimately changing the way tomorrow's enterprise, cloud, automotive, and carrier architectures transform—for the better.

Copyright © 2020-2023 Marvell. All rights reserved. Marvell and the Marvell logo are trademarks of Marvell or its affiliates. Please visit www.marvell.com for a complete list of Marvell trademarks. Other names and brands may be claimed as the property of others.