NVME OVER FABRICS: NEW CLASS OF STORAGE

Ravi Kumar Sriramulu  
Manager, Sales Engineer Analyst  
Dell EMC  
ravi.sriramulu@emc.com

Chethan Nagaraj  
Sales Engineer Analyst  
Dell EMC  
Chethan.Nagaraj@dell.com

Narendra Nekkanti  
Senior Sales Engineer Analyst  
Dell EMC  
Narendra.Nekkanti@dell.com
Table of Contents

Introduction ........................................................................................................................................ 3
How fast should the storage perform? ............................................................................................... 3
Current offerings and its limitations ................................................................................................. 4
Overcoming limitations using NVMe .................................................................................................. 4
Architecture of NVMeoF .................................................................................................................. 6
Implementing NVMe ....................................................................................................................... 6
  NVMe over Fabrics using Remote Direct Memory Access (RDMA) .................................................. 7
  NVMe over Fabrics using Fabric Channel (FC-NVMe) .................................................................... 7
Technical Characteristics .................................................................................................................. 7
Business Use cases .......................................................................................................................... 8
  Use Case 1: Faster Flash .................................................................................................................. 8
  Use Case 2: Storage Class Memory as a Storage array Cache ....................................................... 8
  Use Case 3: In-memory databases for Data Management .............................................................. 9
Dell EMC Offering transition to NVMe over Fibre Channel ............................................................. 9
Conclusion ........................................................................................................................................ 10

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell EMC’s views, processes or methodologies.
Introduction

Flash memory – also known as flash storage – is a type of non-volatile memory, used in enterprise servers, storage and networking technology. Dell EMC has a fleet of dedicated All Flash Storage systems which are built entirely using Flash drives, offering very high performance levels catering to customer expectations for performance, efficiency and reduced operational cost.

Continuing the trend of high performance computing are NVMe devices which are non-volatile memory attached directly to the host system via PCIe bus. It is mainly designed to capitalize on low latency and internal parallelism of flash storage devices, where the traditional All Flash storage systems struggle in spite of having hundreds of flash drives.

NVMe has a limitation of being local to the host over its PCIe bus which makes it difficult to extend its potential to different use cases. This limitation is alleviated by new technology called ‘NVMe over Fabrics’.

This Knowledge Sharing article discusses how ‘NVMe over Fabrics’ unleashes capabilities of NVMe, by enabling the use of alternate transports to PCIe and acts as an alternative to SCSI standards. This will facilitate in extending NVMe capabilities to a wide variety of business use cases, beyond rack-scale architectures to datacentre-wide fabrics.

How fast should the storage perform?

Really fast! Or may be faster than that. Though this sounds a bit absurd, this is how the world has been functioning. You cannot stop with what you have. There is always room to improve and make things more efficient. We have come a long way from punch cards in 1800s, magnetic tapes in 1900s, evolution of room-sized magnetic disks from 1950s which could store 5Mb of data to current days 2.5” form factor and finally to flash-based devices in the 2000s. As the technology continues to transform, storage devices evolved to what they are today – solid state drives, which can hold up to 30 TB capacity.

Evolution in terms of density of storage media is one thing; an equally important aspect of storage media is the performance. Punch cards could easily take up days to hours to complete an operation and today’s SSDs are measured in micro seconds. Increase in performance is the combination of both the technology and architecture of the storage media itself coupled with the software/protocol stack that transfers data to and from this media.

A large amount of data is being generated at a very high volume by human users and machines. Technological advancement has enabled us to analyse this data to achieve meaningful results that can solve real world problems. This has led to the rise of many buzzwords– machine learning, artificial intelligence, personal assistant, predictive analysis and many more. The main aim of data analytics is to consume the data and analyse it, as quick as possible, on per human user or per machine level, and output the results in real time. These results can then be used to make solutions for real world problems, again, in real time.

Time is the key. Though you can distribute a certain analytics job among thousands of servers in a server farm, there still is a challenge to load and destage enormous amounts of data into a compute platform in very less time. This can only be achieved by a highly efficient and low-latency storage infrastructure that at the same time delivers very high performance.
Current offerings and its limitations

Storage industry’s answer to high performance and low-latency application workloads today is All-Flash Arrays (AFA), storage arrays filled completely with solid state drives. Due to advancements in flash storage media coupled with demanding workloads, every company has developed and positioned All Flash Arrays in their portfolio. There are many start-ups which started with purely All Flash Arrays. Dell EMC has half dozen AFAs in its portfolio and these are some of the best of breed systems, each targeted to a specific workload type. These have served very well for the past few years, and continue to serve their target market. However, they have begun to age as the demands keep increasing to unrealistic performance levels in today’s context.

It’s not necessarily a problem with the storage array architecture or flash storage media but with the way the data is handled, i.e the protocol stack that is being used to transport the data – mainly SATA and AHCI associated with it. AHCI was developed to optimize and improve the performance and reduce the latencies of hard disk drives. While AHCI does help SSDs in some aspects, its main aim was to reduce the rotational latency of spinning drives. Several tests have concluded that the performance of AHCI-SATA combination applied to SSDs can deliver peak throughput of 550-650 Mbps. This is much less than what new gen SSDs can offer. Fortunately, SSDs can move away from AHCI and use newer technologies like PCIe and NVMe which can unlock their true potential.

Overcoming limitations using NVMe

Limitations of SATA and AHCI set the stage for PCIe and NVMe as the standard for flash-based storage media. Initially, SATA was the protocol that was first replaced with PCIe for flash media, because of the obvious complexity inherent to SATA. The figure below denotes how PCIe compares to SATA protocol stack.

PCIe was designed to provide CPU with memory-like access to storage media. Along with this it has much smaller software stack and proves to be a very efficient protocol for flash based devices. Although, this is the right way to use flash media, PCIe still made use of SCSI command stack and AHCI optimizations to address and transfer data. As discussed in the previous section, these were designed for spinning storage media and do not offer any benefits to flash-based storage. In fact, they turn out to be the bottleneck in improving flash storage efficiency.
NVMe was designed from the ground up for the modern flash-based storage media. It features a streamlined memory interface, new queue design and new command set. The table below shows the difference between AHCI and NVMe.

<table>
<thead>
<tr>
<th></th>
<th>NVMe</th>
<th>AHCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency</td>
<td>2.8 µs</td>
<td>6.0 µs</td>
</tr>
<tr>
<td>Maximum Queue Depth</td>
<td>Up to 64K queues with 64K commands each</td>
<td>Up to 1 queue with 32 commands each</td>
</tr>
<tr>
<td>Multicore Support</td>
<td>Yes</td>
<td>Limited</td>
</tr>
<tr>
<td>4KB Efficiency</td>
<td>One 64B fetch</td>
<td>Two serialized host DRAM fetches required</td>
</tr>
<tr>
<td>Uncacheable Register Reads</td>
<td>0 per command</td>
<td>4 per command, 8000 cycles or 2.5 µs</td>
</tr>
</tbody>
</table>

NVMe reduces latency by 50% compared to AHCI, while using fewer cycles, which puts less overhead on compute.

NVMe over Fabrics takes the efficiencies of NVMe to build a low latency, high performance distributed storage architecture. Begun in 2014, NVMeoF is being developed by a consortium of around 80 companies. The goal is to extend NVMe features over extended distances through Ethernet, Fibre Channel and InfiniBand.
Architecture of NVMeoF

We offer a brief overview of architecture since a complete storage stack is beyond the scope of this article. More technical details can be found at www.nvmexpress.org.

NVMe over Fabrics overcomes the distance limitations of NVMe by extending the protocol to greater distances, allowing an NVMe host device to connect to a remote NVMe storage driver or subsystem. The goal of NVMeOF was to add no more than 10 microseconds of latency between NVMe host and a remote NVMe target storage device connected through network fabric versus the ultra-low latency of an NVMe storage device connected over a local server’s PCIe bus.

NVMe is defined over fabrics as a “common architecture that supports wide range of storage and networking fabrics for NVMe block storage protocol over a storage networking fabric. This includes enabling front-end interfaces into storage systems, scaling out to large numbers of NVMe devices and also extending the distance within a datacenter over which NVMe devices and NVMe subsystems would be possible to access”.

NVMe subsystem port has a 16-bit port identifier. It consists of one or more physical fabric interfaces acting together as a single interface between the NVMe subsystem and the fabrics. Link aggregation can be used to group the physical ports and constitute a single NVM subsystem port.

The main reason that NVMe outperforms traditional architecture is because of the way its protocol is designed with high performance and parallelism in mind. NVMe supports 64K separate IO queues and each queue can support 64K commands simultaneously. The best known traditional protocol, SATA/AHCI, supports just one queue with 32 commands. There is no comparison between the two protocols intended for the same purpose of transporting data.

Implementing NVMe

The most obvious option would be to use servers with NVMe devices as NVMe footprint expands in the data center. There are vendors who are already bringing NVMe-capable servers to the market with physical connector and BIOS support.

Hypervisor platform already supports NVMe as do the modern operating systems. VMware’s vSAN has been supporting NVMe for some time. Another option to implement NVMe is by back-end storage connectivity through the storage appliance.

Implementing NVMe would yield faster delivery, low-latency connectivity for flash devices with significant improvement in an array’s performance subject to efficient storage operating system code. HP announced NVMe support for 3PAR arrays, NetApp introduced NVMe as read cache in Flash Cache and Pure Storage implements it in its Flash Array/X Platform.

Another advantage is when NVMe storage arrays come along, customers don’t have to upgrade to NVMe all at once as SCSI and NVMe can coexist on the same infrastructure.
Two popular methods of implementation:

**NVMe over Fabrics using Remote Direct Memory Access (RDMA)**

Traditional network protocols are not designed for ultra-low latencies so using them with NVMe defeats the purpose of designing such an efficient protocol. Network protocols that implement Remote Direct Memory Access features are best suited for implementation of NVMe. RDMA protocols are some of the quickest network protocols that induce very low latency on top of NVMe through hardware offloads, and keep the overall latency to very low levels.

The above objective could be achieved by allowing flash devices to communicate overs RDMA fabrics including InfiniBand and RDMA over Converged Ethernet (RoCE).

**NVMe over Fabrics using Fabric Channel (FC-NVMe)**

Fibre channel is the most trusted and widely used purpose-built storage network. Setting up the Fibre Channel as the transport protocol for NVMe over fabrics is the first and most important step in realizing the performance potential of flash in storage networks.

Decades of use has proven that Fibre Channel has the performance, scalability and reliability in handling increasing demand and evolving storage applications. More adoption can be achieved quickly by combined NVMe and FC as most organizations have FC infrastructure in their datacenters.

**Technical Characteristics**

**Reliable, credit-based flow control and delivery mechanisms**

This method of flow control helps the network or fabric be self-throttling, providing guaranteed delivery at the hardware level without any drop in frames due to network congestion. Credit-based flow control is native to Fibre Channel, InfiniBand and PCI Express transports.

**An NVMe-optimized client**

This enhances the clients to send and receive native NVMe command directly to Fabrics without any intermediate layer, i.e transport layer such as SCSI.

**A low-latency Fabric**

Optimised fabric for low latency imposing no more than 10 microseconds end-to-end, including the switches.

**Multi-Host support**

NVMe fabric should be able to support sending and receiving commands from multiple hosts to multiple storage subsystems at the same time.

**Fabric scaling**

NVMe fabric should be able to scale out to tens of thousands of devices.
Multi-path support

The fabric should be able to support multiple paths simultaneously between any NVMe host initiator and any storage target.

Business Use cases

Even though NVMe over fabrics is in its infancy, many users are being exposed to huge amount of fast and efficient NVMe storage that NVMe arrays can deliver.

For example, there is great potential benefit of having a pool of fast storage available for latency-sensitive applications such as online transaction processing (OLTP), data warehousing and analytics-intensive platforms. These would all benefit by having an enormous pool of fast reliable, low-latency storage.

Below are three real-time business cases that will help in understanding the benefits of leveraging NVMe over Fabrics.

Use Case 1: Faster Flash

- Achieving lower latency with NVMe over fabrics compared with legacy SCSI
- Very low CPU utilization
- Scalable to 100’s of drives thus making higher capacity available to applications
- 25GE/32GFC 100GE/128GFC bandwidth to support 32G (PCIe) and faster NVMe drives

Use Case 2: Storage Class Memory as a Storage array Cache
- Improved performance: 4 times greater performance
  - Leverage 32G NVMe
  - Benefit from low latency media
- Caching/fast storage removes PCIe latency
- Benefits from improved performance, higher bandwidth and lower latency

**Use Case 3: In-memory databases for Data Management**

![](image)

- New Storage Tier: In-Memory Databases
- Eliminates datacentre siloes
- Eliminates stranded storage
- Enables
  - Snapshots
  - Data Tiering (HANA)
  - Data Availability
  - Workload Migration

**Dell EMC Offering transition to NVMe over Fibre Channel**

The protocol stack layer has been exposed as the next latency bottleneck after all-flash arrays appeared on the scene. The ecosystem for this technology is expected to be built in 2017 and 2018 as HBA drivers, operating systems, enterprise features (MPIO), and matured storage array options become more available.

Customers are now able to future-proof their Dell EMC 14G PowerEdge servers by purchasing Gen6 FC16 and FC32 HBAs. Dell EMC HBAs will be time-to-market with NVMe over FC enablement for the different operating systems via driver and firmware updates. Customer need the correct HBAs, a compatible operating system, and acquire an NVMe over FC storage array once they become available.

The availability of NVMe over Fibre Channel-enabled adapters enables Dell EMC to provide enterprises with both investment protection and the flexibility to move to NVMe storage at the pace of their business.
Conclusion

NVMe over fabrics will be the next big thing in storage technology. NVMe over fabrics is still in its early days but would be interesting to watch how this category develops over the coming years. Performance benchmarking tests have been conducted using NVMe and NVMeOF. Comparing the results to performance of traditional architecture – SATA, SAS, AHCI, etc. – makes them look like they were designed in the punch card and magnetic tape era.

Success of NVMeOF lies in the success of the applications that use NVMeOF, Looking at current disruptive market trends, it appears certain that this new technology will replace everything that exists today.