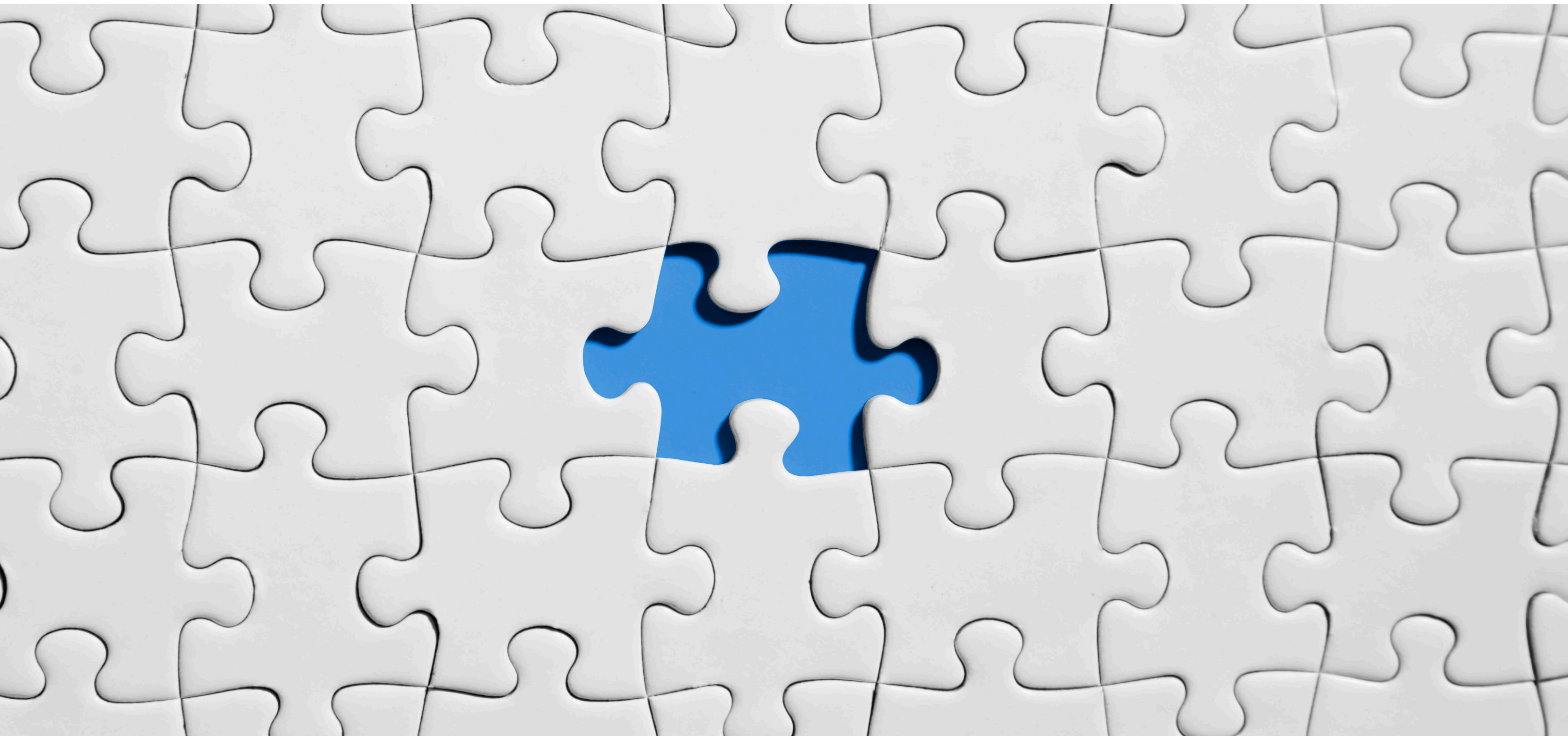
Dell Technologies

NVME THE FUTURE



Abhijith M

Sales Engineer Analyst

Dell Technologies | PSS

Abhijith.M@dell.com

Naga Koduru

Senior Sales Engineer Analyst

Dell Technologies | PSS

Naga.koduru@dell.com

Knowledge Sharing Article

© 2020 Dell Inc. or its subsidiaries.

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Abstract

This Knowledge Sharing article introduces NVMe and a brief background of protocols currently used. It also provides an overview of approaches on how NVMe is used in the SAN environment to increase performance and reduce bottlenecks that we currently see. We outline the benefits of using NVMe in the data center and the advantages of using NVMe over other protocols.

Table of Contents

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

Executive Summary

This article provides an overview of Non-Volatile Memory Express (NVMe) technology. It examines how protocols used in SAN for data transfer have changed over time and provides insight on how NVMe is developed from the ground-up to gain benefits that legacy storage protocols did not offer.

1. Introduction

NVMe is key to unlocking the next level of performance in Storage arrays. NVMe is a high performance protocol designed for modern drives from the ground-up to overcome limitations imposed by SAS and SATA.

SAS and SATA were designed with hard disk drives (HDD's) in mind. When next gen media (i.e. SSD's) arrived, these protocols were still used (and still are). However, this created a bottleneck on the storage side as these Flash drives were not used effectively when using SAS/SATA protocols. NVMe was built so that these bottlenecks would be overcome. The NVMe architecture takes advantage of the parallelism of modern CPUs and solid-state drives (SSD's). NVMe maximizes the power of Flash drives and in turn, removes the bottleneck from the Storage. It also opened the door to the next media disruption with Storage Class Memory (SCM).

2. Background

Data has become the most valuable asset to all businesses. The rapid increase in data generated worldwide has created a need to re-strategize how data is captured, stored, accessed and transformed into meaningful information.

As processor and memory speeds increase every day, storage becomes a bottleneck. To speed applications, getting data to the applications for processing is critical.

SAS and SATA were designed some time ago. These protocols were very efficient with HDDs since the CPU cycles consumed in getting the data to application was only a small proportion of the overall I/O operation. Most of the time was consumed by HDD's responding to re-positioning of the head to read/write data to the drives. However, the situation reversed once Flash Drives or SSD's arrived. Now the drives were much faster in magnitude compared to the HDDs that are used. This is explained in greater detailin section 3.2., "Why NVMe".

3. NVMe Overview

3.1 Introduction to NVMe

NVMe stands for Non-Volatile Memory express. Non-volatile memory is the type of memory that doesn't lose the data when powered off or during a power loss. Flash and SSD's are a type of non-volatile memory. NVMe is a way to access this data.

NVMe is a storage protocol that enables host software communication with non-volatile memory across a PCIe bus. It is a communication interface and driver that defines a command set and feature set for PCIe SSD's. It is the industry standard for PCIe SSD's in all form factors.

NVMe was designed for SSD's. It communicates between the system CPU and storage interface using the high speed PCIe sockets.

3.2 Why NVMe

Before we discuss why NVMe came into picture, let's understand how previous protocols fetched data from the drives.

Imagine a series of conveyor belts that have blocks of data on them and are continuously rotating. Consider these our HDDs (Spinning drives). We have a robot with an arm that we program to get our desired data block. The robot is also on a track and can move only sideways on the track as depicted in Figure 1(a).

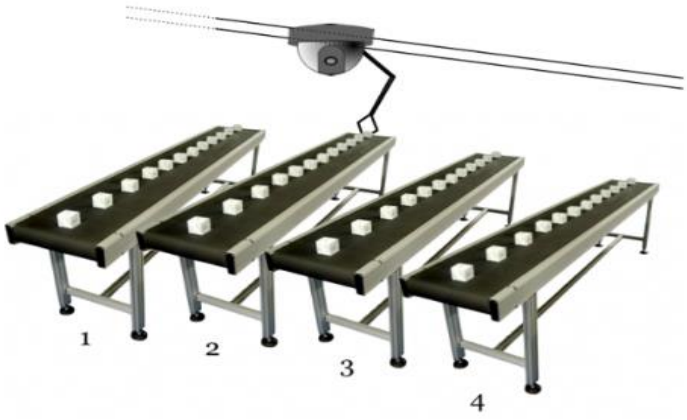
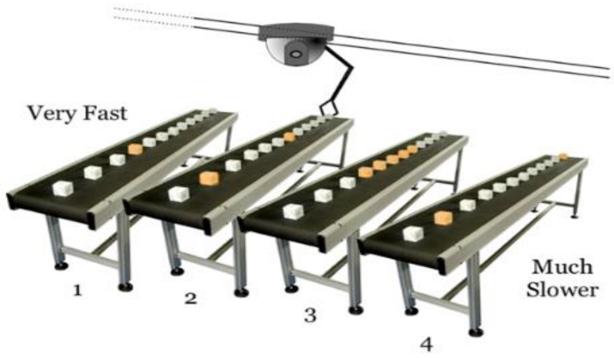


Figure 1(a)

Now, suppose we need to get a data block from the conveyor belts (drives). The robot has to move sideways to the appropriate belt and then wait for the correct data block to arrive in position so that it can be picked up. Also, let's say each conveyor belt moves at a different pace as shown in Figure 1(b).



The head has to move to the right place and wait for the right block to come around

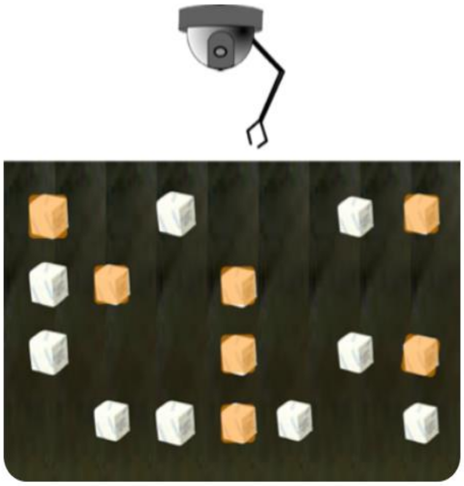
Figure 1(b)

The above example depicts how Spinning drives work, i.e. the spindle moves to place, waits for the disk to rotate and then fetches the data. In the case of our storage, the instructions sent to the robot head to fetch the data block is SCSI commands.

The older protocols were built with the limitations that these conveyor belts posed. An important point to note here is that the robot has just one arm and could fetch only one block at a time, i.e. it was one command at a time; the rest must be queued and all the commands are maintained in the same single queue.

When Flash drives arrived, the concept of conveyor belts no longer applied since we don’t have spinning parts inside them. Though access of data blocks was faster with SSD’s, the advantages of Flash drives could not be leveraged completely because we were still following protocols that were built for HDD’s.

We did not have to wait for the data block to come to the appropriate place to fetch it. Basically, the robot can see all the blocks and the blocks do not change their positions. However, the robot still responds to SCSI commands, i.e. one at a time though multiple data blocks can be picked up.



Ah, no need to wait for the blocks to come to the right place – they're right there!

Figure 1(c)

This is where NVMe comes in. It was built from the ground up for SSD's. Where SCSI had just one queue for commands, NVMe is designed to have 64,000 queues and each queue can have 64,000 commands simultaneously. An analogy for the example we used to depict how HDDs work, in the case of NVMe, the robot has 64,000 arms, each having the ability to handle 64,000 commands.

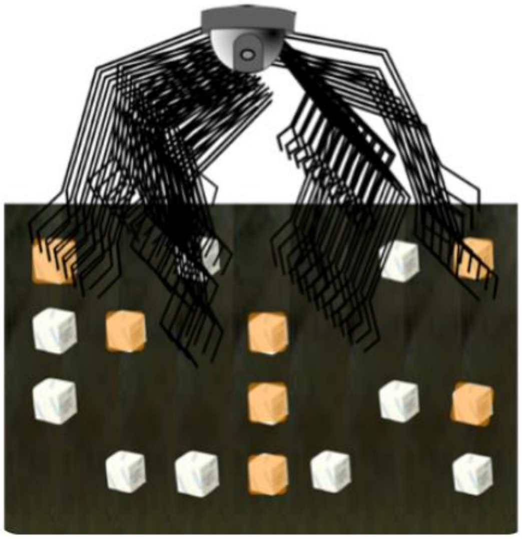


Figure 1(d)

With the introduction of NVMe, we can completely use the advantages of SSD's.

3.3 NVMe Over Fabrics

NVMe over Fabrics, also known as NVMe-oF and non-volatile memory express over fabrics, is a protocol specification designed to connect hosts to storage across a network fabric using the NVMe protocol.

The protocol is designed to enable data transfers between a host computer and a target solid-state storage device or system over a network, accomplished through a NVMe message-based command. Data transfers can be transferred through methods such as Ethernet, Fibre Channel (FC) or InfiniBand.

NVM Express Inc. is the nonprofit organization that published version 1.0 of the NVMe specification on March 1, 2011. On June 5, 2016, the same organization published version 1.0 of the NVMe-oF specification. NVMe version 1.3 was then released in May 2017. This update added features to enhance security, resource sharing and SSD endurance.

The NVM Express organization estimated that 90% of the NVMe-oF protocol is the same as the NVMe protocol, which is designed for local use over a computer's Peripheral Component Interconnect Express (PCIe) bus.

Vendors are working on developing a mature enterprise ecosystem that supports end-to-end NVMe over Fabrics, including the server operating system, server hypervisor, network adapter cards, storage OS and storage drives. In addition, SAN switch vendors – not limited to Cisco Systems Inc. and Mellanox Technologies – are trying to position 32 gigabits per second (Gbps) FC as the logical fabric for NVMe flash.

Since the initial development of NVMe-oF there have been multiple implementations of the protocol, such as NVMe-oF using remote direct memory access (RDMA), FC or Transmission Control Protocol/Internet Protocol (TCP/IP).

3.4 Uses of NVMe over Fabrics

Although still a relatively young technology, NVMe-oF has been widely incorporated into network architectures. Using NVMe-oF can help provide a state-of-the-art storage protocol that can take full advantage of today's SSDs. The protocol can also help bridge the gaps between direct-attached storage (DAS) and SANs, enabling organizations to support workloads that require high throughputs and low latencies.

Initial deployments of NVMe were DAS in servers, with NVMe flash cards replacing traditional SSDs as the storage media. This arrangement offers promising high-performance gains when compared with existing all-flash storage, but it also has its drawbacks. NVMe requires the addition of third-party software tools to optimize write endurance and data services. Bottlenecks persist in NVMe arrays at the level of the storage controller.

3.5 Benefits of NVMe over Fabrics

Benefits of NVMe-based storage drives include:

- low latency
- additional parallel requests
- increased overall performance
- reduction of the length of the OS storage stacks on the server side
- improvements pertaining to storage array performance
- faster end solution with a move from Serial-Attached SCSI (SAS)/Serial Advanced Technology Attachment (SATA) drives to NVMe SSDs
- variety of implementation types for different scenarios

3.6 Technical characteristics of NVMe over Fabrics

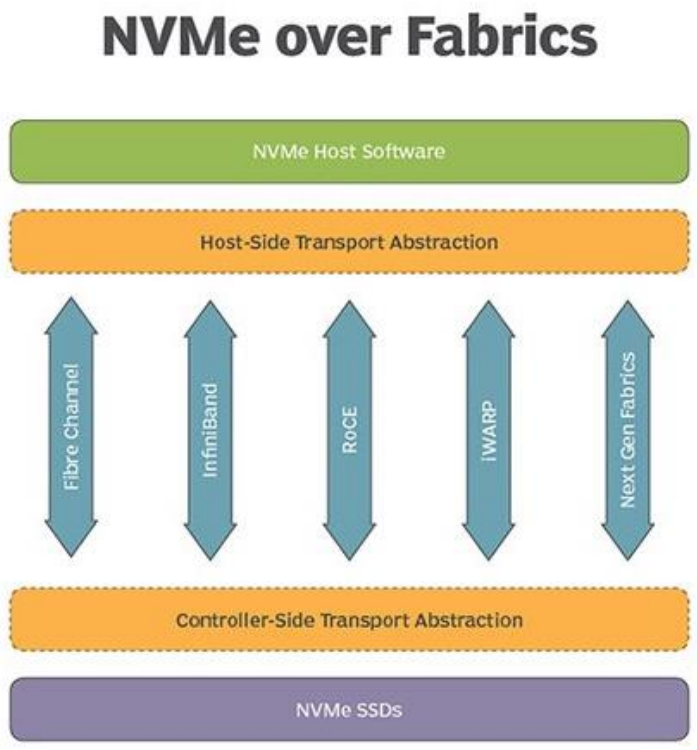
Some of the technical characteristics of NVMe-oF are:

- high speed
- low latency over networks
- credit-based flow control
- ability to scale out up to thousands of other devices
- multipath support of the fabric to enable multiple paths between the NVMe host initiator and storage target simultaneously
- multihost support of the fabric to enable sending and receiving commands from multiple hosts and storage subsystems simultaneously

3.7 NVMe over Fabrics vs. NVMe: Key differences

NVMe is an alternative to the Small Computer System Interface (SCSI) standard for connecting and transferring data between a host and a peripheral target storage device or system. NVMe is designed for use with faster media, such as SSDs and post-flash memory-based technologies. The NVMe standard accelerates access times by several orders of magnitude compared to SCSI and SATA protocols developed for rotating media.

NVMe supports 64,000 queues, each with a queue depth of up to 64,000 commands. All input/output (I/O) commands, along with the subsequent responses, operate on the same processor core, parlaying multicore processors into a high level of parallelism. I/O locking is not required, since each application thread gets a dedicated queue.



NVMe-based devices transfer data using a PCIe serial expansion slot, meaning there is no need for a dedicated hardware controller to route network storage traffic. Using NVMe, a host-based PCIe SSD is able to transfer data more efficiently to a storage target or subsystem.

A main distinction between NVMe and NVMe over Fabrics is the transport-mapping mechanism for sending and receiving commands or responses. NVMe-oF uses a message-based model for communication between a host and a target storage device. Local NVMe will map commands and responses to shared memory in the host over the PCIe interface protocol.

3.8 NVMe over Fabrics using Fibre Channel

NVMe over Fabrics using Fibre Channel (FC-NVMe) was developed by the T11 committee of the International Committee for Information Technology Standards (INCITS). FC enables mapping of other protocols on top of it, such as NVMe, SCSI and IBM's proprietary Fibre Connection (FICON), to send data and commands between host and target storage devices.

FC-NVMe and Gen 6 FC can coexist in the same infrastructure, enabling data centers to avoid a forklift upgrade.

Customers use firmware to upgrade existing FC network switches, provided the host bus adapters (HBAs) support 16 Gbps or 32 Gbps FC and NVMe-oF-capable storage targets.

The FC protocol supports access to shared NVMe flash, but there is a performance hit imposed to interpret and translate encapsulated SCSI commands to NVMe commands. The Fibre Channel Industry Association (FCIA) is helping to drive standards for backward-compatible FC-NVMe implementations, enabling a single FC-NVMe adapter to support SCSI-based disks, traditional SSDs and PCIe-connected NVMe flash cards.

3.9 NVMe over Fabrics using TCP/IP

One of the newer developments regarding NVMe-oF includes the development of NVMe-oF using TCP/IP. NVMe-oF can now support TCP transport binding. NVMe over TCP makes it possible to use NVMe-oF across a standard Ethernet network. There is also no need to make configuration changes or implement any special equipment with the use of NVMe-oF TCP/IP. Because the transport binding can be used over any Ethernet network or the internet, the challenges commonly involved in implementing additional equipment and configurations are eliminated.

TCP is a widely accepted standard for establishing and maintaining network communications when exchanging data across a network. TCP will work in conjunction with IP, as both protocols used together facilitate communications across the internet and private networks. The TCP transport binding in NVMe-oF defines how the data between a host and a non-volatile memory subsystem are encapsulated and delivered.

The TCP binding will also define how queues, capsules and data are mapped, which supports TCP communications between NVMe-oF hosts and controllers through IP networks.

NVMe-oF using TCP/IP is a good choice for organizations that wish to utilize their Ethernet infrastructure. This will also enable developers to migrate NVMe technology away from Internet SCSI (iSCSI). As an example, an organization that doesn't want to deal with potential hassles included in implementing NVMe over Fabrics using RDMA can instead take advantage of NVMe-oF using TCP/IP on a Linux kernel.

4. References

- Data Center NVMe for beginners: https://blogs.cisco.com/datacenter/nvme-for-absolute-beginners
- NVMe Express page: https://nvmexpress.org/
- Understanding NVMe and SSDs: https://www.kingston.com/en/community/articledetail/articleid/48543
- NVMe-oF: https://searchstorage.techtarget.com/definition/NVMe-over-Fabrics-Nonvolatile-Memory-Express-over-Fabrics

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.

Copyright © 2020 Dell Inc. or its subsidiaries. All Rights Reserved. Dell Technologies, Dell, EMC, Dell EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.