Quantum®

WHITE PAPER

NVMe, RDMA, AND OTHER EMERGING TECHNOLOGIES

A Technical Primer

CONTENTS

INTRODUCTION

Over the past decade, enterprise-grade non-volatile "flash" storage has gone mainstream. While not exactly inexpensive, enterprise flash prices have finally dropped to the point that talk of the "all-flash data center" no longer generates quite as much eye-rolling as it once did. Spinning disk capacity is still relatively low-cost, but the technical barriers to greater HDD density and performance are mounting. The vast majority of data centers will house a mix of flash and spinning disk for some time, as organizations deploy both technologies according to their requirements.

By now everyone with a personal computer is familiar with the performance gains available simply by switching from HDD storage to SSD. What many don't realize is that they aren't getting everything they paid for with that upgrade. Flash storage is so fast that the bottleneck in storage systems is no longer the storage devices themselves. The bottleneck has moved upstream, and it's now the storage interface holding things back.

The solution to this mismatch is NVMe, Non-Volatile Memory express. NVMe was designed to unlock the performance of all types of non-volatile memory, from flash SSDs to the latest persistent memory technology. This paper will describe what NVMe is, how it enables greater performance in both standalone and networked implementations, and some of the key use cases that can benefit from this speed.

The enterprise non-volatile storage landscape is broad and confusing. Like any hot sector, both startups and established players alike promise to solve all your problems. There is a whole new dictionary worth of acronyms to navigate, and the new tech can be expensive, upping the stakes. Cutting-edge persistent memory devices such as DRAM and 3D-Xpoint are the priciest. In the middle are enterprise NVMe SSDs and SAS SSDs. Device price is only part of the story, however. TCO matters. Factors such as performance density, infrastructure costs, and the extent to which time equals money all impact whether NVMe storage is suitable for a specific application.

In the broadest sense, nearly anywhere that speed of production of a digital product translates into higher revenue or a competitive advantage is a likely candidate for NVMe storage.

To understand how we got here and why NVMe is so important, it's helpful to quickly look back a few decades.

A BRIEF HISTORY OF MODERN DATA STORAGE

In the beginning (or at least quite a while ago) was the hard disk, the HDD. Arguably the first cheap bulk storage medium for real-time operations, the PC revolution made it ubiquitous, with accompanying steady declines in cost. Form factors shrank, density per platter and number of platters increased. What didn't change were the essential physics of the device. With only a single head stack with a single actuator, read and write operations can only happen one at a time, in serial. As a result, ATA and SAS, the communication protocols used to communicate with HDDs, have only a single queue.

Also in the beginning (or at least quite a while ago) was non-volatile storage. Modern NAND flash was invented in the 1980s. Its widespread use in consumer electronic devices drove development and ultimately price declines. Flash has the advantage that it's not mechanical, and not limited to processing commands in serial. Low latency and highly parallel accessibility are two of flash's greatest strengths.

HDDs dominated for so long because of their good performance, relatively high capacity and low cost per TB compared with flash. Over time, however, the price delta has been decreasing, and the highest capacity flash SSDs are now much larger than the largest HDDs. Flash is most often used in applications where the performance is worth a higher price, but sometimes it has lower TCO than HDD due to its extremely high capacity and performance density, which translates to lower power, cooling, and physical space consumption.

Figure 1 – Storing Data Since...

ENTER NVMe

Flash-based storage is wonderful, and you can speed up just about any old disk array by swapping out HDDs for SSDs. But to take full advantage of flash's low latency and high parallelism, you must do more than just swap devices—some new connectivity technology is needed. On the parallelism front, SAS's single queue architecture doesn't have any, so that's a mismatch with flash. With HDDs, there is a lot of latency due to the mechanics. If the software upstream is faster than the hardware, no problem. But now with flash, the latency of the storage device itself is so low that the software and protocols become the slowest links in the chain. Finally, having to go through a hardware storage controller to get to the PCIe bus adds more latency and another bottleneck to both HDDs and SSDs.

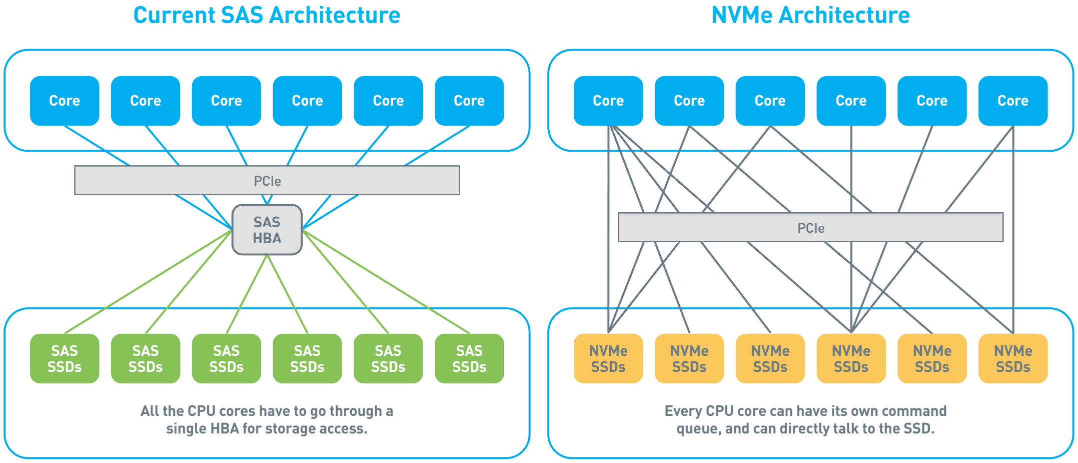
The answer to these problems is the interface specification called NVMe, promoted by the NVM Express organization (https://nvmexpress.org). NVMe was designed specifically to take advantage of the speed, low latency, and in-device parallelism of non-volatile memory devices.

	NVMe	SAS	SATA
Start year	2009	2005	2000
Throughput	~8 Gbps/lane	12 Gbps	6 Gbps
Latency	<20 µ sec	<500 µ sec	<500 µ sec
Read IOPs	~1.0 mil	~200K	~50K
Command Set	NVMe	SCSI	ATA
Number of Cmds	13	~200	
#Queues	65,535	1	1
#Cmd-Queue-depth	65,536	254	32
Form-factors	2.5 in, U.2, M.2, Rules	2.5 in, 3.5 in	2.5 in, 3.5 in

Figure 2

Figure 2 summarizes many of the important differences between NVMe and the traditional interfaces, but a few deserve emphasis. One is the massive single-device parallelism potential of NVMe. Whereas SAS had a single queue with up to 254 entries, a single NVMe drive can have up to 65,535 queues, each with 65,536 entries. It's up to device manufacturers to decide how many to implement, but the specification is built with a lot of headroom.

Practically speaking, having many queues means that multiple CPU cores may each access many NVMe drives concurrently, with a queue for each core/drive pair, as illustrated on the right side of figure 3.


Current SAS Architecture

NVMe Architecture

All the CPU cores have to go through a single HBA for storage access.

Every CPU core can have its own command queue, and can directly talk to the SSD.

Figure 3 – NVMe Before & After

Another advantage illustrated in figure 3 is the fact that NVMe devices are directly connected to the PCIe bus instead of an intermediary like a SAS controller or HBA. Shared memory is used to transfer data from the application to the storage, bypassing the CPU.

Also critical is the fact that the NVMe interface is just much faster in every way, with reduced latency and higher throughput and IOPs capability.

SHARED STORAGE AND NETWORKING NVMe

Everything described thus far applies to storage devices connected within a single server. But that's boring and relatively easy. Quantum is all about high-performance shared storage for demanding workflows like 8K video playout and autonomous vehicle research. Sharing requires a network.

To share the goodness of NVMe, instead of using shared memory, messages are transmitted between hosts and storage over a network. That network can be Fibre Channel (FC), Infiniband (IB), or Ethernet, but Ethernet is what almost everyone wants these days. For one thing, Ethernet adapters and switches are inexpensive. More important than equipment cost is convenience. New deployments don't require learning the dark arts of FC or IB networking—most organizations already have Ethernet expertise in-house. Even if you have an FC or IB storage network deployed today, there can be benefits to converging on a single network infrastructure. This isn't to say that FC and IB connectivity isn't important. Dedicated storage networks based on those technologies will remain viable for some time. Any vendor who claims otherwise probably only supports Ethernet.

The difference between FC, IB, and Ethernet is that FC and IB were designed from the beginning to efficiently handle storage traffic. They deliver data "reliably" (without dropping any) and in order. Ethernet was designed as a general-purpose network and is "unreliable" by design. It will deliver your data, but pieces may take different paths and arrive out of order or be dropped in transit. Upper-layer protocols at the sender and receiver must account for the fundamental unreliability of Ethernet, buffering and re-ordering data and detecting and re-sending dropped packets. All of this takes time, another way of saying it introduces latency—and latency is the enemy of storage networking.

To make a screaming fast, low-latency storage network with Ethernet requires deploying additional technology beyond what most organizations already have in place. Data Center Bridging (DCB) features such as Priority Flow Control (PFC) or Differentiated Services (DifServ) may be needed to create a "lossless Ethernet", and Remote Direct Memory Access (RDMA) is an absolute requirement. A thorough treatment of DCB is beyond the scope of this paper, but RDMA will be explored in some detail, because it's the secret to successful use of NVMe devices across an Ethernet network.

RDMA IS THE SECRET SAUCE

The single biggest contributor to performance and reducer of latency for Ethernet-networked NVMe is RDMA. iSCSI Extensions for RDMA (iSER) and NVMe over Fabrics (NVMeoF) are built on top of RDMA, and it's been used in IB forever. The RDMA Consortium (https://rdmaconsortium.org) was formed back in 2002 specifically to create and promote technology for RDMA over TCP/IP networks. To understand why this old concept has become the new savior, we need to review how a regular TCP/IP Ethernet network is built.

Traditional network protocol stacks like TCP/IP are many layers deep and implemented in the operating system kernel. There are buffers at each step of the way. The CPU is involved in copying data to and from those buffers, and to and from applications sending and receiving data. Computationally expensive context switches are required, generating more latency and CPU utilization. Pile storage protocols on top, and things get even worse.

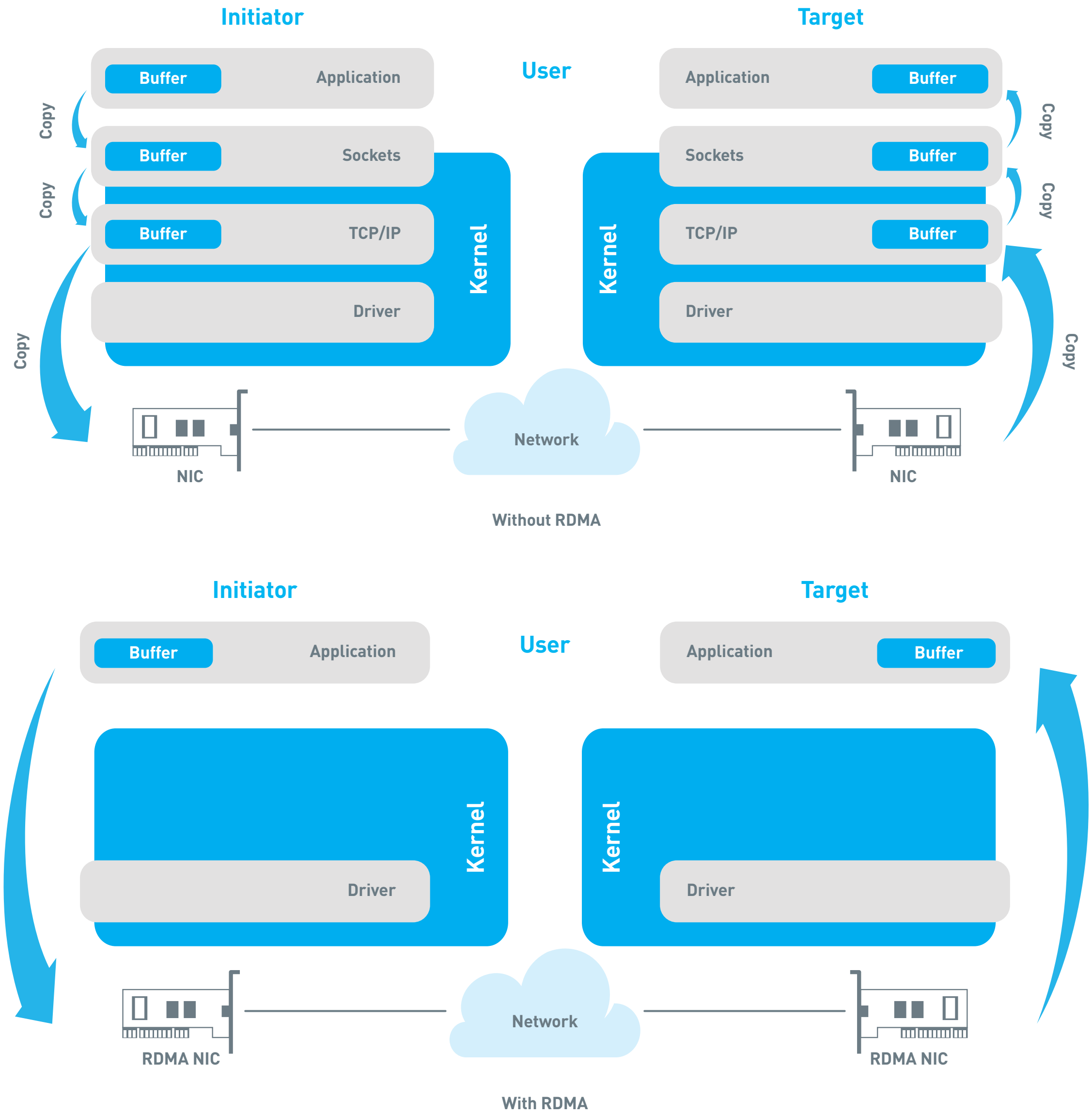


Figure 4 – RDMA Before & After

In contrast, RDMA is a gigantic shortcut for data. It's like taking the expressway to get across town instead of waiting at a bunch of red lights in the city center. With RDMA, data is moved directly between user-space applications. Nearly all the details are offloaded to the network interface hardware.

The CPU is involved in setting up and supervising connections, but data movement is completely offloaded to the NIC. This frees valuable CPU cycles for other tasks. All application I/O happens in user space, bypassing the operating system kernel, which eliminates buffering and context switching. Network protocols formerly handled in the kernel are offloaded to the NIC hardware.

The effect of this offloading and streamlining is an order-of-magnitude reduction in additional latency imposed by the network, down to ~10 μs or less. This is what makes RDMA the hero behind Ethernet-networked NVMe.

ETHERNET RDMA OPTIONS

Because RDMA is so important, it would be nice if implementing an RDMA-capable network were straightforward. Unfortunately, there are a lot of different ways to get to the goal. What follows is a brief overview of the existing options for RDMA-capable Ethernet, offered as a starting point for further research. As always, the best choice for any specific deployment depends on individual circumstances and requirements.

With new technology, initial products frequently use proprietary implementations. Once the landscape starts maturing, standards emerge that show the way forward more clearly. As standards-based solutions multiply, the proprietary options fade away or are relegated to niche use cases where they have some unique advantage. This story is playing out with Ethernet-networked NVMe storage now. Some proprietary and semi-proprietary solutions exist, but they are being quickly swamped by standards-based products. The two relevant standards for building an RDMA-capable Ethernet for networking NVMe are iSCSI extensions for RDMA (iSER) and NVMe over Fabrics (NVMeoF).

iSER was conceived to extend iSCSI to use RDMA and moved into IETF standards in 2007 with RFC 5046. iSER relies on one of three network transports—TCP with RDMA services (iWARP), RDMA over converged Ethernet (RoCE), or IB. Don't confuse iSER with old-school iSCSI. The two are incompatible, and iSER can provide vastly higher performance thanks to RDMA. Under the covers, however, it still involves the SCSI protocol. In theory this means there is greater latency and potentially lower performance than a pure NVMe implementation like NVMeoF, but whether this matters in practice depends on the design of the storage system and how it is used.

NVMeoF is newer. Version 1.0 of the specification was published in 2016 and has continued to evolve since then. NVMeoF is designed to extend the efficiencies of NVMe over potentially any network fabric, today including not just Ethernet but also IB and FC. RoCE and iWARP are the RDMA Ethernet transports supported, and there is a new non-RDMA Ethernet implementation based on traditional TCP sockets known as NVMe/TCP. FC and IB are the non-Ethernet alternatives.

To boil this down, running NVMe with RDMA over Ethernet, whether via iSER or NVMeoF, requires an Ethernet network (both NICs and switches) that support RoCE or iWARP. The two are not compatible, meaning a RoCE client cannot access a storage resource presented via iWARP and vice versa.

WHAT'S NVME STORAGE GOOD FOR?

As discussed earlier, one of the biggest strengths of NVMe-attached storage is low latency. This makes it particularly suitable for high-performance, high-resolution streaming media where dropped frames are unacceptable. Playout of 4K and 8K, UHD and HDR video workloads is the poster child, but editing and graphics work, and ingest or delivery of high numbers of lower-resolution streams are also appropriate fits.

Other types of workloads can benefit from the high-IOPS characteristics of NVMe storage, including AI/ML applications, IOT, and high-performance analytics, especially where real-time requirements exist.

Though the price of NVMe storage may seem high, the initial acquisition price is often offset by environmental factors. Building an HDD-based system with very high performance relies on aggregating together many spindles. This can result in a dramatic mismatch between capacity and performance, where one is forced to purchase far more capacity than needed just to get enough spindles to meet the performance requirement. In addition to the rack space penalty, all those spinning hard drives need to be powered and cooled. It's possible for an NVMe-based storage appliance to occupy just a few rack units and provide performance equal to HDD-based systems that occupy many full racks. This is what is referred to as performance density—the number of IOPs, or 8K video streams, or other units of work that can be accomplished per rack unit. NVMe storage has extremely high performance density when compared to HDD-based solutions, and even those based on SAS-attached SSDs.

For applications where only some of the data needs ultra-fast access, some of the time, NVMe storage can even be used as a "tier zero" in a hybrid storage appliance to extend its benefit and reduce cost.

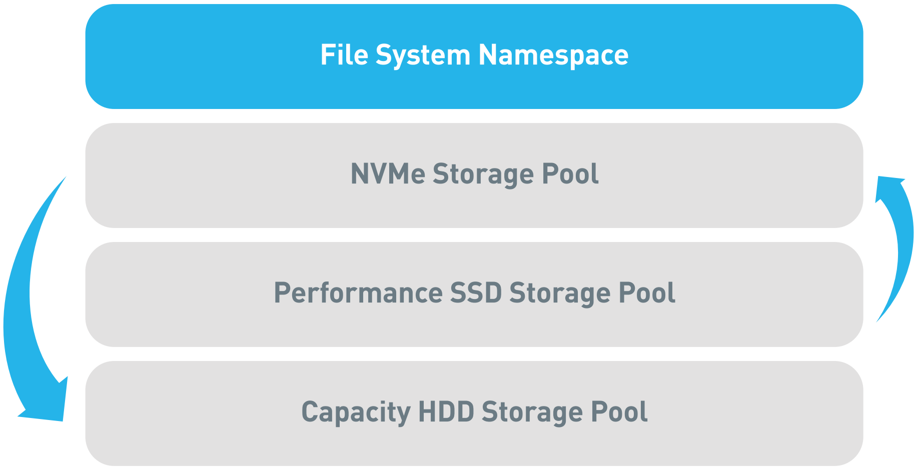


Figure 5 – Tiered Storage System Incorporating NVMe

CONCLUSION

The present and future of high-performance storage is NVMe-connected non-volatile memory technology. NVMe, leveraging RDMA, enables low-latency, high-bandwidth sharing of this new storage, minimizing or eliminating many traditional networked storage bottlenecks. When deployed intelligently, NVMe storage is affordable and accessible. Combining NVMe drives with other types of storage in a hybrid system can drive performance up while keeping costs under control. Quantum's expertise in designing shared storage systems for video and other high-bandwidth applications now extends to include NVMe-based products targeted at the media & entertainment, video surveillance, autonomous vehicle research, and similar markets. Visit www.quantum.com for more information.

ABOUT QUANTUM

Quantum technology and services help customers capture, create, and share digital content—and preserve and protect it for decades at the lowest cost. Quantum's platforms provide the fastest performance for high-resolution video, images, and industrial IoT, with solutions built for every stage of the data lifecycle, from high-performance ingest to real-time collaboration and analysis and low-cost archiving. Every day the world's leading entertainment companies, sports franchises, research scientists, government agencies, enterprises, and cloud providers are making the world happier, safer, and smarter on Quantum. See how at **www.quantum.com**.

www.quantum.com • 800-677-6268