Compliments of :

HITACHI
Inspire the Next

BROCADE
A Broadcom Company

NVMe over Fibre Channel

for dummies
A Wiley Brand

Boost performance with super-low latency

Maintain mission-critical storage SLAs

Reduce risk with concurrent SCSI/NVMe

AJ Casamento
Marcus Thordal

3rd Brocade Special Edition

Brocade and Hitachi - Alone, Exceptional. Together, Unrivaled.

Driven by one of the closest, longest-standing alliances in the industry – Hitachi and Brocade have been collaborating for more than 20 years, focusing on equipping today's customers with the technology to become the dynamic enterprise of tomorrow.

Hitachi Vantara and Brocade set the gold standard in future-proof storage networking performance, availability, resilience, simplicity, and security for the evolving organization. Integrating automation architecture, lowest power consumption, and the most advanced SAN monitoring and diagnostics in the industry.

About Brocade

Brocade, a Broadcom Inc. Company, is the proven leader in Fibre Channel storage networks that serve as the foundation for virtualized, all-flash data centers. Brocade Fibre Channel solutions deliver innovative, high-performance networks that are highly resilient and easier to deploy, manage, and scale for the most demanding environments. Brocade Fibre Channel storage networking solutions are the most trusted, widely deployed network infrastructure for enterprise storage.

www.broadcom.com

About Hitachi Vantara

Hitachi Vantara, a wholly-owned subsidiary of Hitachi, Ltd., guides its customers to what's next by solving their digital challenges. Hitachi Vantara applies its unmatched industrial and digital capabilities to its data and applications to benefit both business and society. More than 80 percent of the Fortune 100 trust Hitachi Vantara to help them develop new revenue streams, unlock competitive advantages, lower costs, enhance customer experiences, and deliver value.

www.hitachivantara.com

NVMe over Fibre Channel

3rd Brocade Special Edition

by AJ Casamento and Marcus Thordal

for dummies®
A Wiley Brand

NVMe over Fibre Channel For Dummies®, 3rd Brocade Special Edition

Published by

John Wiley & Sons, Inc.

111 River St.

Hoboken, NJ 07030-5774

www.wiley.com

ISBN: 978-1-394-15984-0 (pbk); ISBN: 978-1-394-15985-7 (ebk). Some blank pages in the print version may not be included in the ePDF version.

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Manager and Development Editor: Carrie Burchfield-Leighton

First Edition Co-author: Curt Beckmann

1st and 2nd Editions Project Editor: Martin V. Minner

Sr. Managing Editor: Rev Mengle

Acquisitions Editor: Ashley Coffey

Business Development Representative: Matt Cox

Brocade Contributors: Marc Angelinovich, Howard Johnson, Dave Peterson, Juan Tarrio

Table of Contents

Introduction

Unless you've been holed up in Siberia herding reindeer since the launch of *Sputnik 1*, you probably know that your kid's hand-me-down iPhone could whup the *Apollo 13* in both computing and storage capacity. And you likely are aware that the intense pace of innovation is continuing. Today's network speeds are not only millions of times faster but also carry millions of times more data per second. Processing speed has grown exponentially. As for storage, the contents of the entire Library of Congress can be squeezed into an affordable disk array. Times have indeed changed.

About This Book

NVMe over Fibre Channel For Dummies, 3rd Brocade Special Edition, focuses on a relatively small but important aspect of information technology. Non-Volatile Memory Express (NVMe) over Fibre Channel is a technology that touches computer memory, storage, and networking. If you're a hardened computer geek, you've probably heard of it. If not, this won't be the last time. Like most of the IT world, NVMe over Fibre Channel enjoys a rich history and has evolved at breakneck speed, building on the capabilities of preceding technologies while avoiding the shortcomings of competitive ones. For the inexperienced, this book introduces a technology that's been evolving into the next big thing in networked storage.

Simply put, NVMe over Fibre Channel has it all. It has the ultra-low latency needed for working memory applications, with the reliability that's critical to enterprise storage. Because Fibre Channel (FC), as all network geeks know, is a premium datacenter network standard, NVMe over Fibre Channel is able to leverage fabric-based zoning and name services. Best of all, NVMe over Fibre Channel plays well with established FC upper-layer protocols, enabling a low-risk transition from Small Computer System Interface (SCSI) to NVMe without the need to invest in experimental infrastructure.

Your knowledge of network and storage technology is likely well beyond that of your Aunt Mary, who calls every weekend asking for help with her USB drive. If not, this book offers plenty of reminders and sidebars to guide you through the more difficult parts and to help decipher the endless acronyms lurking around every corner of the IT world.

Icons Used in This Book

A number of helpful icons are scattered throughout this book. These reinforce and further explain important concepts and keep you out of trouble with your boss.


TIP

Pay special attention to the Tip icons. They contain small bits of information that make your job a lot easier and prevent you from ordering late-night pizza delivery because you're stuck in the server room reconfiguring a storage array.


REMEMBER

If you're someone who spends their days reading thick hardware manuals, you're bound to forget things along the way. If so, the Remember icon may just be your new best friend. It points out considerations that you may otherwise trip over.


WARNING

Computer hardware and networking equipment is expensive. Replacing it because you made the wrong decision is even more so. Heed the Warning icons if you want to avoid costly mistakes in your IT strategy.


TECHNICAL STUFF

For you dedicated hardware professionals with framed photographs of Jack Kilby and Robert Metcalfe hanging in your cubicles, keep an eye out for the Technical Stuff icons; they're chock-full of additional details on esoteric subjects.

Beyond the Book

This book can help you discover more about NVMe over Fibre Channel, but if you want resources beyond what this book offers, we have some insight for you:

- www.broadcom.com/solutions/data-center/fc-nvme: Visiting this URL takes you to Brocade's main NVMe over Fibre Channel page where you can find this book as well as videos and other papers about NVMe.
- docs.broadcom.com/docs/FOS-90-Traffic-Optimizer-OT: Read through this white paper about a new Gen 7 Fibre Channel feature that also works with NVMe. It's the ability to separate traffic based on similar traffic characteristics like speed or protocol. This guarantees performance and eliminates congestion caused by a speed/performance mismatch.
- docs.broadcom.com/doc/FOS-90-Fabric-Notifications-OT: This paper discusses Fabric Notifications at much greater length. For an introduction to the topic, Check out Chapter 6 of this book.
- www.broadcom.com/solutions/data-center/storage-fabrics-technology: If you want to learn more about Fibre Channel in general or about an individual product, this landing page takes you to the right place.

IN THIS CHAPTER

- Figuring out the ABCs of NVMe over Fibre Channel
- Debating storage versus classic memory
- Speeding access to flash
- Seeing how NVMe relates to SCSI
- Awaiting future benefits of NVMe over Fibre Channel
- Strengthening your fabric

Chapter 1
Exploring NVMe over Fibre Channel

This chapter offers a brief introduction to Non-Volatile Memory Express (NVMe) over Fibre Channel. We introduce (or reintroduce, for you veterans) a bunch of acronyms, discuss the merits of solid-state storage and see how it compares to memory, and break down the component parts of this exciting, relatively new technology. We want you to understand why it may just be the best thing for your organization since pocket protectors.

Wading through the Alphabet Soup of NVMe over Fibre Channel

NVMe over Fibre Channel is a full-featured, high-performance technology for NVMe-based fabric-attached enterprise storage, but it's a no-compromise solution for NVMe working memory use cases as well. (In this chapter, we talk about how those use cases differ.) NVMe over Fibre Channel is relatively new, even though its component parts aren't. Fibre Channel (FC) has been

the leading enterprise storage networking technology since the mid-1990s. Speeds of 16G (called Gen 5) are widespread. Gen 6 FC became available in 2016, delivering twice the speed of Gen 5 and a staggering 8 times the bandwidth on 32G FC links, and it's selling like hotcakes. Gen 7 switches and host bus adapters (HBAs) are now also available with support for 64G and many advanced new features. FC is primarily used to carry the Small Computer System Interface (SCSI) protocol, which is the historic leader for direct-attached PC or server storage. SCSI on FC is called, boldly enough, FC Protocol (FCP).

NVMe refers to several related things:

- An open collection of standards for accessing and managing nonvolatile memory (NVM), especially high-performance, solid-state memories such as flash or 3D XPoint
- That collection's primary specification/revision (NVMe 1.4b), which provides a common, high-performance interface for accessing NVM directly over PCI Express (PCIe), as shown in Figure 1-1
- The nonprofit corporation NVM Express (`www.nvmexpress.org`), which works to develop and promote the standard and is supported by a wide range of technology companies such as Intel, Dell/EMC, Samsung, Micron, and others

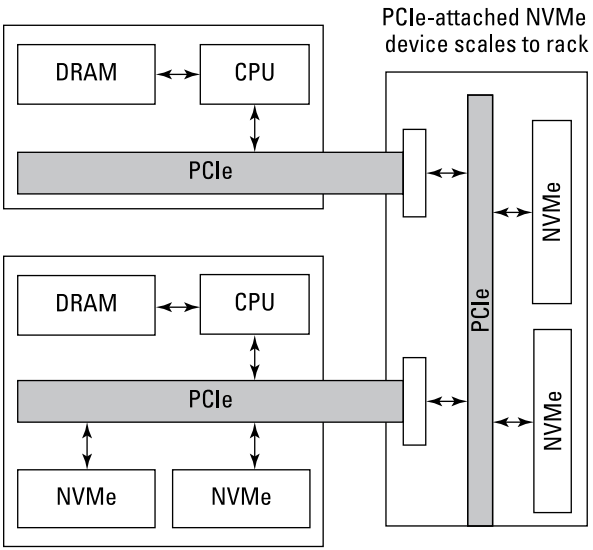


FIGURE 1-1: NVMe connects to a server PCIe bus internally or externally.

PCI-based NVMe has low latency, but it has important limitations relative to fabric-based media. The benefits of fabric connectivity include greater capacity, better capacity utilization, shared access, and advanced services (like deduplication, snapshots, replication, and backups). Using a fabric also eliminates single points of failure and simplifies management. To bring all these benefits to the NVMe ecosystem, NVM Express developed NVMe over Fabrics (NVMe-oF), which defines how NVMe commands can be transported across different fabrics in a consistent, fabric-independent way. Figure 1-2 shows that process, which makes life a lot easier for software developers.

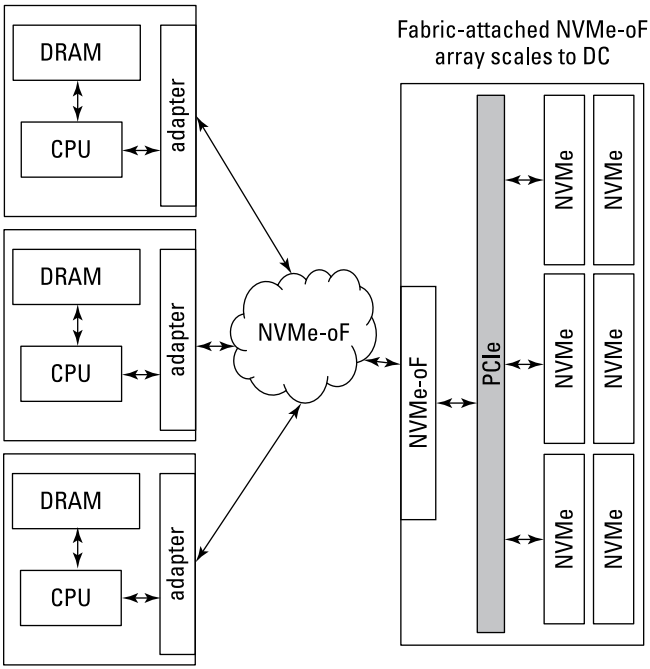


FIGURE 1-2: Using NVMe-oF to increase the scale of NVMe.

The NVMe-oF 1.0 specification, released in 2016, describes two fabric categories:

- **FC fabrics:** NVM Express chose the T11 standards body for FC, which handles all FC standards, to define NVMe over Fibre Channel. In mapping NVMe onto FC, the T11 committee members followed in the footsteps of SCSI, making it

straightforward to carry both SCSI and NVMe traffic on the same infrastructure. The T11 committee finished its work in October 2017. The FC-NVMe-2 specification introduces some enhancements to make NVMe over Fibre Channel even faster and more reliable.

- **Remote Direct Memory Access (RDMA) fabrics:** RDMA is an established protocol that has run for years on InfiniBand, RoCE (pronounced "rocky"), and iWARP (we warned you about acronyms). Building on RDMA allowed NVM Express to target three existing fabric transports in one effort.

Note: In early 2017 (after NVMe-oF 1.0), a group at NVM Express advanced an effort to map NVMe over Transmission Control Protocol (TCP) (without RDMA) so NVMe-oF can run in existing datacenters that lack RDMA support.

REMEMBER

FC is capable of transporting multiple higher-level protocols concurrently, such as FCP and NVMe over Fibre Channel, as well as FICON, a mainframe storage protocol (see Chapter 3 for more info on FICON). That bears repeating: NVMe over Fibre Channel can coexist on your FC storage area network (SAN) and HBAs right along with your existing FCP or FICON traffic.

NVMe over Fibre Channel offers robust interoperability, darned fast performance, and extremely scalable architecture. Whether you're faced with a legacy storage network in need of an upgrade or a spanking new memory-centric implementation, NVMe over Fibre Channel offers a best-of-both-worlds solution while allowing a smooth transition for traditional users.

Picking Sides: Is It Storage, or Is It Memory?

Some may detect a hint of tension built into the phrase NVMe over Fibre Channel. That's because FC is a storage-oriented technology while the term *NVM* is obviously memory-oriented. By contrast, the other NVMe fabrics (InfiniBand, RoCE, and iWARP — we cover these more in Chapter 4) are memory-oriented (they support RDMA). Indeed, recent conferences covering flash and other persistent memory technologies have gushed over the arrival of the "storage/memory convergence."

Mapping the dichotomy

For decades, memory and storage have represented a dichotomy. Both could hold information, but memory was built into the server, while storage has largely been separate, holding data independent of a server or application. To some degree, that dichotomy has been self-reinforcing:

- Classic enterprise storage is relatively slow compared to dynamic random-access memory (DRAM), with extensive error checking and read/writes that are often sequential. Memory is designed to be fast but transient.
- Hard disk drives (HDD) and solid-state drives (SSD) scale far beyond normal memories. They're cheaper per bit and can persist their data when powered down, which is important for archiving.
- Enterprise storage also supports a range of service-level agreements (SLAs), which are contracts between storage customers and their storage providers, with cool features like redundant array of independent disks (RAID), replication, deduplication, and compression. Try that with memory.

Table 1-1 offers a comparison of memory and storage characteristics.

TABLE 1-1 Memory and Storage Characteristics

Feature	"Ideal Memory" Priority	Flash Memory Is Like . . .	NVMe Protocol Is Aimed at . . .	"Ideal Storage" Priority
Read bandwidth	Very high	Memory	Memory	Medium
Write bandwidth	Very high	Storage	Memory	Medium
Read latency	Very high	Memory	50/50	Medium
Write latency	Very high	Storage	50/50	Medium
Read granularity	High	Memory	Storage	Low

(continued)

TABLE 1-1 *(continued)*

Feature	"Ideal Memory" Priority	Flash Memory Is Like . . .	NVMe Protocol Is Aimed at . . .	"Ideal Storage" Priority
Write granularity	High	Storage	Storage	Low
Scale	GB to TB	GB to PB	Storage	TB to EB
Random access	Very high	Memory	Memory	Low
Persistence	Low	Storage	Storage	Very high
Rewritable	High	Storage	Both	Low to medium
Reliability	High	Memory	Storage	Very high
Density	Medium	Storage	Storage	High

So, memory and storage still look different and that's likely to continue. To paraphrase Mark Twain, reports of the convergence of memory and storage may be somewhat exaggerated. Perhaps we can say that there's something of a trend toward convergence instead of an imminent event. We can also say that there's an emerging convergence of protocols for shared memory and shared storage.

On errors

While understandable, it's somewhat ironic how memory errors are largely tolerated (or at least not eliminated) in computing, while storage errors are not. This is true at a variety of tiers. Laptops (user grade compute) typically have no error correction on DRAM but have cyclic redundancy check (CRC) error detection built into drives. Servers (enterprise-grade compute) have error correction code (ECC) DRAM, which corrects single-bit errors but simply aborts or shuts down on dual-bit errors. By contrast, enterprise-grade storage includes redundancy in some form, such as RAID or erasure coding.

Instead of fixing every memory error, the industry approach is to abort and restart a computation using the same stored data. That's because storage has a higher level of guarantee, or SLA. This terminology is often used even within organizations between consumers of storage and their IT departments.

OTHER DATA ABOUT DATA

For decades, magnetic recording technologies in the form of disk and tape were the dominant storage technologies, while silicon (mostly DRAM) has been the dominant memory technology. The main technology driver behind the NVMe protocol has been the steady improvement in density and performance of flash memory. Flash has displaced disk-based storage over a number of years; flash has been the default storage in laptops for the past ten years and since 2015 has been transitioning to NVMe attached SSDs.

WARNING

Solving the problems of working memory wouldn't fully solve the rare but meaningful untrustworthiness of computation. Viruses, loss of network connectivity, and power failures are just a few of the events that often disrupt inflight computation. This is why overinvesting in working memory rarely makes sense. By contrast, long-term storage must be recoverable because no "do-over" mechanism exists. The moral? Don't cheap out on storage. We return to this point in upcoming chapters.

Accelerating Access to Flash

Flash is disrupting the storage world in a big way, but it's hardly a first for the industry. The storage market is such a poster child for disruption that way back in 1995, Harvard's Clayton Christensen used it as a prime example in his classic Harvard Business School paper on disruption. The early disruptions were more about size and cost, while flash is more about performance.

Flash has always offered much faster read capability (especially for random access) than spinning disk drives. But flash's early density was much lower than disk drives. In addition, writing to flash is trickier than writing to DRAM or magnetic storage media. Flash has relatively low write endurance — in the neighborhood of one million erase/write cycles for each flash block. Repeated writes to a block of flash can also degrade the reliability of adjacent blocks. So, despite its speed, flash's early shortcomings limited it to niche uses.

Over time, the density of flash increased dramatically, and effective software algorithms were developed to mask its write challenges. The combination of speed, density, and tolerable write endurance have brought flash to the point where it's the technology of choice for production data and has displaced spinning disks in the datacenter.

Indeed, the killer speed benefit of flash, as well as other solid-state memory technologies, has highlighted that the old tried-and-true storage protocols had a weakness: performance.

Understanding How NVMe Relates to SCSI

The basis of most of today's storage-oriented protocols, including FCP, is the SCSI standard established in the 1980s. SCSI was originally built around hard disk drives but has been extended a number of times to include other storage devices while maintaining backward compatibility. SCSI currently supports well over 100 commands. Besides hauling a lot of baggage, SCSI also lacks deep command queues.

SCSI's numerous extensions and extended support for legacy applications have resulted in a protocol stack that's sluggish in comparison to the NVMe stack, which has dramatically enhanced queuing and has been simplified and optimized for both semiconductor memory and today's operating systems. Figure 1-3 shows this comparison.

REMEMBER

For those who may, at some point, be migrating SCSI-based storage assets into an NVMe environment, here are some important things to understand:

- **Mapping legacy SCSI to NVMe:** The NVMe community has recognized the importance of the storage market and the prominence of SCSI within that market. That's why the NVMe standards groups invested time and energy to ensure that NVMe could implement the functionality needed by legacy storage-dependent applications.

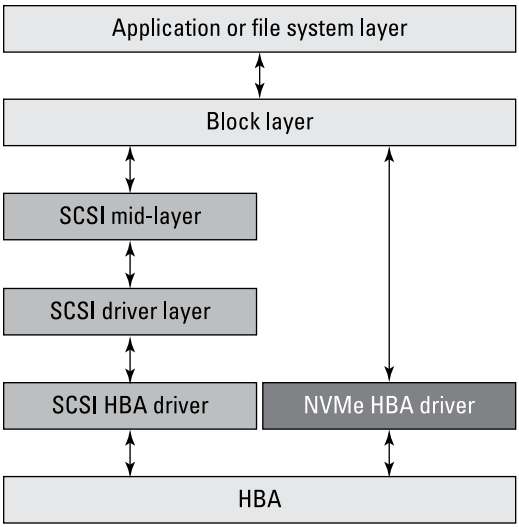


FIGURE 1-3: Comparing the SCSI and NVMe software stacks.

- **LUNs and namespace IDs:** A logical unit number (LUN) is the SCSI mechanism for identifying different volumes within a single storage target. In other words, each volume is a LUN. NVMe uses the term *namespace ID* (NSID) in a similar way. Namespace is a curious term, considering that each one is treated as a set of logical block addresses (LBAs), not as a set of names.
- **Enhanced command queuing:** FC-NVMe exposes the enhanced queuing capabilities of NVMe, allowing thousands of parallel requests across a single connection. With today's multithreaded servers and virtual machines running dozens to hundreds of applications, the benefits of parallelism are massive.
- **FCP:** FCP is, like namespace, an odd moniker for what is really SCSI over FC. FCP isn't about basic FC; it's about the way SCSI features were implemented on top of FC.
- **Leveraging FCP:** Without going into the gory details, you should know that NVMe over Fibre Channel uses a new FC-NVMe frame type for non-input/output (I/O) communications while reusing the FCP frame type for I/O operations. So, if you capture all the frames running across an NVMe-over-FC interface, you'll see FCP in the mix.

- **Other protocols:** The name *FCP for SCSI-on-FC* may be partly responsible for a perception that FC is limited to SCSI, but it's important to remember that SCSI isn't the only popular protocol used with FC. The mainframe storage protocol FICON runs over FC, and NVMe is another protocol that now runs on top of FC (which is the whole point of this book).

WARNING

SCSI's sluggishness is a characteristic of the protocol stack, not the FC transport. Some NVMe advocates have pointed to the implementation of SCSI over FC and incorrectly blamed its relatively poky performance on FC. This is an incorrect assertion: FC-NVMe is faster than SCSI over FC.

Anticipating Future Benefits of NVMe over Fibre Channel

One of the key benefits of NVMe over Fibre Channel is its scalability. Built from the ground up with nonvolatile memory in mind, it also leverages the speed and robustness of FC. (We say more about the inherent benefits of FC in Chapter 2 and new benefits of NVMe over Fibre Channel in Chapter 6.) Leveraging FC as a transport gives users easy access to all the speed and parallelism of NVMe-oF with none of the disruption entailed in building parallel infrastructure.

REMEMBER

As NVMe over Fibre Channel increases its lead on competing technologies, keep in mind the following additional considerations:

- As flash becomes more storage-oriented, SCSI, the dominant storage protocol, has hampered one of the key advantages of flash: its speed. That will change as more storage vendors embrace NVMe over Fibre Channel.
- New NVM media such as Intel/Micron's 3D XPoint (Optane) entered the market (in 2017) with SSD media delivering much faster read and writes (in the sub 10 nano seconds range) as well as 100 times or even 1,000 times write endurance.

Invigorating Your Fabric

As the underlying technology of storage arrays moves from spinning disk to flash, and from flash to even faster technologies, the increasing speed will generate increased pressure to save those precious hundreds of microseconds that NVMe offers over SCSI.

In addition, many applications will have mixed needs, requiring some storage-oriented volumes and some memory-oriented volumes. Even more compelling, there will be times when you want to do both with the same information. That is, you'll want to maintain a master copy of some data asset, enabling all the high reliability features of enterprise storage.

At the same time, other consumers of that data asset may only need high-speed read access to that data. It may make sense to publish (using the dual-protocol concurrency of your FC fabric) your master storage volume to a "working reference memory" on an NVMe over Fibre Channel drive. Such an image could be read-only and would have no need for features that could drive up latency or cost. By using memory-oriented NVMe over Fibre Channel arrays, you may save money as well.

IN THIS CHAPTER

- Seeing where FC belongs in the storage ecosystem
- Establishing performance metrics
- Optimizing performance
- Improving queuing
- Counting on reliability

Chapter 2
Delivering Speed and Reliability with NVMe over Fibre Channel

Ethernet was already the mainstream network of choice when Fibre Channel (FC) first started shipping in 1997. Ethernet was well on its way to overrunning protocols like Token Ring and asynchronous transfer mode (ATM), and many in the networking community doubted whether FC had any future at all. Boy were they wrong. In the face of strong Ethernet headwinds, FC networking managed to grow to the point where nearly all enterprises, financial institutions, and governments relied on it for their mission-critical storage needs. FC's success wasn't magic. FC was and remains different from Ethernet in important ways. Understanding why FC has been so successful in an Ethernet-dominant world is a necessary first step toward deciding whether to adopt this robust networking technology.

Reviewing FC's Place in the Storage Ecosystem

Ethernet remains the primary transport for communication between servers and between clients and servers. However, distinct advantages do exist to using FC in a storage-centric environment. Traditional Ethernet, including most Ethernet deployed today, doesn't do much about network congestion. Instead, it pushes the responsibility of reliable transport to the upper-layer protocols. If you pretend for a moment that your network is like a freeway at rush hour, Ethernet allows unlimited cars (frames) to enter the on-ramps, but then it steers them into the ditch (drops the frames) when the freeway runs out of space. For those cars that don't reach their destinations, no biggie: Transmission Control Protocol (TCP) patiently, hesitantly, tries again, sending as many cars as necessary, even if there's already a car crash up ahead.

FC, on the other hand, was designed for well-ordered, reliable transport of data regardless of load. It has a feature similar to entrance ramp lights. No frames are lost, and each is delivered in the proper order. Within Gen 7, FC freeways expand or isolate lanes based on max speed.

Evaluating Performance Metrics in Storage Context

You can't improve what you can't measure. That's why it's important to establish performance metrics before embarking on any improvement project, even for something as trivial as planting a garden — if the seeds aren't deep enough or you used the wrong fertilizer, you may just go hungry.

Storage is no different. Without a clear understanding of how your company's megabuck hardware investment is performing, or whether data is being lost or users are complaining over slow access to corporate data, you may end up with a lot more time on your hands. Maybe you can take up gardening?

FABRIC METRICS

Fabric metrics can often resemble storage metrics and can sometimes cause confusion if you're not familiar with their distinct meanings. Both storage and networking folks should be alert to the subtleties to avoid embarrassing apples-versus-oranges fruit salad episodes.

While storage latency measures a full storage operation from start to finish, fabric latency tells you how much incremental latency a fabric device would add to a connection relative to a direct connection. It's measured as the time between the arrival of the first bits of a frame and the time those same bits are first transmitted ("first in, first out" — the FIFO model). The fabric latency is the same for both read and write operations and for different input/output (I/O) sizes.

Fabric throughput is usually viewed as how much data can be pushed through the fabric when all ports are running at maximum speed. A 64-port, 10G device usually has a throughput of 640 Gbps (double that if input and output are separately counted, which is known as full-duplex). But some low-end devices have internal "oversubscription" and can't forward at full speed on all ports at once, so check the fine print.

No fabric metric corresponds to IOPS. However, the IOPS metric exists because the latency metric doesn't tell the full story when I/O operations overlap. Similarly, when multiple traffic flows overlap in a network and create congestion, no simplistic metric exists to capture behavior.

Of course, there's more to overall performance than the speed of the network. If your hard drives are dogs, why bother with FC? Similarly, if the server is still using Pentium Pro processors, you'd better address your server performance first before attacking any network upgrades. As with everything in life, some things have greater priority. Sure enough, the whole NVMe conversation itself was started by high-speed solid-state drives (SSDs).

This section focuses on storage metrics; they're what the storage consumer ultimately cares most about, and they better enable apples-to-apples comparisons between different Non-Volatile Memory Express (NVMe) over Fabrics options. (For a peek at fabric

metrics, see the sidebar, "Fabric metrics.") The storage community has long used three key metrics for measuring performance:

- Latency
- Throughput
- Input/output operations per second (IOPS)

The relative importance of these performance indicators depends a great deal on the user application. Systems that focus on minimizing response time value latency above the other metrics. Streaming high-definition video requires huge amounts of data, so good throughput is paramount, and the heavy read/write activity seen in databases calls for high numbers of IOPS (pronounced *eye*-ops). Even the experts vary. One calls latency the "king" of storage metrics while another refers to IOPS as the "grandfather."

Storage metrics

Here's a quick introduction to the key storage metrics:

- **Latency,** especially read latency, is the main claim to fame of flash-based systems, and it's the benefit most often touted for NVMe-based data transfers (normally in comparison to serial attached Small Computer System Interface [SCSI]).

  Storage users care about overall operation latency, which is the time from when a read or write operation begins until that operation has fully completed. Storage latency depends on the size and direction of the I/O operation (reads versus writes), whether I/Os are random or sequential, as well as the connection speed, and the relevant values should be included when mentioning a particular latency metric value.
- **Throughput** describes how fast, in megabytes (MB) or gigabytes (GB) per second, a storage device can read or write data. These metrics are most often better for large I/Os. As with latency measurements, throughput measurements should state the I/O direction (read or write) and access type (sequential or random). I/O size is good to include for completeness, but if it isn't mentioned, you can assume the number applies for larger I/Os.
- **IOPS** tells you how many individual read and/or write operations a device can handle per second. Like latency and throughput, IOPS metrics vary by the size, direction, and access type (sequential versus random) of the I/Os.

Typically, a device's IOPS metric is higher for smaller I/Os. That's why quoted IOPS metrics are often for an I/O size such as 4 kilobytes (K). However, many applications that demand high IOPS use larger I/O sizes, such as 64K. Look closely to ensure that your device IOPS metric is aligned with what your application needs.

In a simple case, IOPS are closely related to latency. If you imagine a connection where a 4K read operation takes 1 millisecond (ms), you may expect to be able to perform 1,000 of those I/Os in a second, and indeed that may be the case. But because this simple picture doesn't always hold true (more on this later in the chapter), it's useful to check both latency and IOPS metrics.

REMEMBER

Some applications use small I/O sizes and demand high IOPS and low latency with mostly random-access patterns, such as databases or data warehouses. Other applications perform large sequential operations, either primarily reads (video streaming) or primarily writes (backup to tape drives). IOPS and latency will be the key metrics to optimize for in the former case, while throughput will be the one to look out for in the latter.

WARNING

Storage metrics can't tell you everything. In order to allow for apples-to-apples comparisons, performance benchmarking is done in controlled situations, such as a single tester connected to a single device. The upshot is that, in the real world, "your mileage may vary." Although the "controlled situation" aspect of performance benchmarks is legitimate, it also creates natural pressure to tune devices so they excel in the test situation, although they may not always perform as well in familiar real-world environments. Here are two examples:

- **Example 1:** Some devices can have excellent throughput or IOPS numbers for sequential access, but that performance may apply only when you have a small number of storage clients issuing operations. A large number of requesters can overload the device resources, causing the sequential performance benefit to be lost.
- **Example 2:** Some flash arrays keep a pool of erased flash blocks to allow for faster writing. For a write operation, the controller "remaps" the relevant logical block addresses (LBAs) to a block from the pool, writes the new data there, and

marks the old flash block to be erased in the background. If the device runs low on erased blocks, the "garbage collection" process can lurch into the foreground and dramatically slow normal operations until the process finishes.

Because the real world isn't a controlled situation, IT architects who care about consistent high performance demand that their environments include tools that enable rapid and deep investigation into system behavior. Brocade, a Broadcom Inc. company, has long recognized that the expectations of FC customers (unlike the commodity-biased Ethernet market) justify an investment in analytics tools. Those customers who are seeking the performance benefits of NVMe technology are especially likely to find themselves in need of tools for optimizing their environments. (For more information, see the sidebar "Brocade Autonomous SAN.")

BROCADE AUTONOMOUS SAN

Brocade Fabric Vision technology delivers a collection of features that combines comprehensive data collection capabilities with powerful analytics to quickly understand the health and performance on the environment and identify potential impact or trending problems. These features are the foundation for realizing a self-learning, self-optimizing, and self-healing autonomous SAN. These features include several under three categories.

Self-Learning

- **IO Insight:** On supported products, proactively and non-intrusively monitors storage device IO latency and behavior through integrated network sensors, providing deep insight into problems and ensuring service levels
- **Monitoring and Alerting Policy Suite (MAPS):** Provides an easy-to-use solution for policy-based threshold monitoring and alerting

  MAPS proactively monitors the health and performance of any SCSI or NVMe storage infrastructure to ensure application uptime and availability. By leveraging prebuilt rule-based and policy-based templates, MAPS simplifies fabric-wide threshold configuration, monitoring, and alerting.
- **Automatic Flow Learning:** Provides automatic learning of all traffic flows from a specific host to storage across the SAN fabric

With this information, an admin can automatically identify resource contention or congestion that's impacting application performance.

- **Fabric Performance Impact:** Leverages predefined MAPS policies to automatically detect and alert administrators to different congestion severity levels and to identify credit-stalled devices (for example, misbehaving slow-drain devices) or oversubscribed ports that could impact network performance

  This feature pinpoints exactly which devices are causing or are impacted by the congested port, and it quarantines the misbehaving devices.

Self- Optimizing

- **Traffic Optimizer:** Automatically classifies and separates traffic with similar characteristics to optimize performance for most common SAN configurations

  It identifies and isolates traffic flows to prevent negative impact to overall SAN performance.

Self-Healing

- **Fabric congestion notification:** Automatically detects congestion and notifies end devices to automatically mitigate congestion
- **Slow-drain device quarantine:** Automatically quarantines credit-stalled devices to prevent the misbehaving device from impacting the rest of traffic
- **Automatic actions:** Ensures data delivery with automatic failover from physical or congestion issues, such as port decommissioning, port toggling, and port fencing

Souping up device metrics

In this section, you find a discussion of technologies that have traditionally helped storage performance by overcoming some of the limitations of the hardware.

Read caching can help hard disk drives (HDDs) when accessing data sequentially because the disk drive can read an entire disk track and cache the extra data blocks. Read caches don't help SSDs as much (they handle any reads pretty quickly), but sometimes a

read cache can be helpful for SSDs that are otherwise unable to read during write operations.

Write caching can be helpful for flash in two ways:

- **Write speed:** Writing to a dynamic random-access memory (DRAM) is faster than writing to the flash device, which must be erased before it can be written.
- **Write endurance:** An application may write the same block multiple times in a short period. The write cache can wait a bit, then transform multiple cache writes into a single write to flash.

Parallelism is a simple matter of using a large number of underlying devices to deliver higher throughput, and often higher IOPS.

Pipelining describes a system that performs different functions in parallel. For example, read operations may be broken into different functional stages: command pre-processing, physical access of backend devices, error correction, and sending. The functional stages are like different sections of a pipeline, and you can see the read operation "moving through the pipe." With pipelining, the system can be working on different stages of two read operations at the same time. Separate pipelines are normally used for read and write.

Pipelining allows a device's IOPS metric to exceed what you may expect from its latency metric. For example, you may expect a device with a 1 ms latency metric for 256KB reads to have a 1,000 IOPS metric. But when reads are overlapped (later reads are sent before earlier reads have completed), the device may deliver a 1,500 IOPS metric for 256KB reads.

Achieving overlapped reads or writes is tricky for "single threaded" applications, but most performance-sensitive apps are "multi-threaded" by design to generate the overlapping I/Os and take advantage of pipelined device performance. In addition, virtualized servers running a lot of apps in parallel also issue many overlapping I/Os. The NVMe standard includes architectural enhancements that can greatly accelerate overlapping I/Os.


TIP

If you're considering using a device with write caching, ensure that the device's power-fail write-back behavior is aligned with the application needs. If the application requires that all writes are made persistent, the device must guarantee that the cache contents are saved in the event of a power failure.

WARNING

Be aware that architectural performance enhancements can only go so far. Write caches can help with write bursts, but if the long-term requested write throughput exceeds what the hardware behind the cache can swallow, the cache fills up, and the requested writes get throttled to match the underlying device throughput. Pipelines may lose performance benefits when operations to the same underlying device overlap. You'll need to test specific products with your applications.

Achieving High Performance

High performance is relatively straightforward in NVMe over Fibre Channel, just as it is with FC itself. FC vendors have pushed the performance edge from the beginning, with top speeds and features like direct placement of storage payloads to reduce memory copy overhead. Customers contributed as well, by opting for optical connections more often than mainstream networking did. FC provided a much simpler networking stack, with a single network layer, simplified addressing, and topology-agnostic routing that allows for all links to be used in parallel. In addition, FC fabrics are typically implemented as parallel (air-gapped), redundant, active-active fabrics, which offer both reliability and added performance. Similarly, FC's built-in, credit-based flow control offers reliability while also improving performance.

All these optimizations make perfect sense in a datacenter architecture, even if some of these choices would be out of place in a campus or internet context. NVMe over Fibre Channel leverages all the traditional benefits of FC, while providing additional performance benefits inherent in NVMe. The streamlined protocol stack is one benefit; the other major improvement is enhanced queuing.

Making Use of Enhanced Queuing

For decades, storage vendors have competed for leadership on the metrics in this chapter and have long made use of techniques such as caching, parallelism, and pipelining to boost their performance metrics. Another advantage is that flash chips are much smaller than the smallest disk drives. This means using thousands of flash chips in parallel is much easier than using thousands of disk drives.

Meanwhile, just as the storage targets have been moving toward extreme parallelism, so have the storage initiators. Servers are running vastly more threads, cores, and virtual machines. As a result, the number of parallel I/Os that can be generated has risen steeply.

SCSI-based devices offered parallelism, allowing storage initiators to "queue up" multiple commands in parallel. But the SCSI queue depth has been limited for both individual logical unit numbers (LUNs) or volumes and for target ports, which typically support several LUNs. SCSI queue depth limits vary from 8 to 32 commands per LUN, and 512 to 2048 commands per port. Historically, these seemed more than adequate, but as today's environments scale out, SCSI is feeling the squeeze.

The designers of NVMe were aware of the trends, and defined the protocol's queuing depths accordingly. NVMe supports a vast 64k queues with a depth of 64k commands each.

REMEMBER

Enhanced queuing allows for a far greater parallelism. This doesn't mean you'll immediately see 100 times the improvement with NVMe, but it isn't hard to imagine a significant boost of two times or more. The advanced analytics features available with Brocade's FC fabrics enable you to track the queue depth across all the devices and applications in your environment.

Realizing Reliability

Everyone wants reliability: reliable cars, reliable employees, reliable internet. But without reliable data, the examples in this chapter may be difficult to attain.

Of course, everyone in IT knows that users want reliable computing. And yet, there are different ways to deal with errors, and some are more expensive than others. Companies tolerate the occasional crash in their laptops instead of spending the money (and battery life) on the error correction circuitry (ECC) required to minimize bit errors. After all, laptops have many exposures, so users seek to protect them in other ways, such as background backup software.

REMEMBER

Many companies' servers have ECC to fix single-bit errors, but they crash on double-bit errors. Servers, like laptops, are also vulnerable to viruses and power failures, so companies don't spend more for stellar ECC that fixes double-bit errors. They live with the occasional server crash because they can rerun the app to get the results. That's only true because their enterprise storage guarantees availability of golden copies of the critical data assets. How would life be different if you couldn't count on those assets?

Redundant networks and multipath I/O

When *Apollo 13* was launched, it was chock-full of redundant systems designed to make the spacecraft function properly in the event of an isolated failure. Your data assets may not be quite as precious as the lives of astronauts, but you do have critical assets, and you've had enough experience to know where redundancy makes sense. For enterprise storage customers, it's a matter of economics; the right amount of redundancy enables you to deliver "five 9s" reliability, keeping your customers happy.

TECHNICAL STUFF

Literally, "five 9s" means your computing infrastructure works 99.999 percent of the time (which results in the maximum downtime in a year being no more than 5 minutes and 15.36 seconds annually, or slightly longer in a leap year).

WARNING

Cutting corners can save you money, but remember that disappointing your customers can cost you more money in the long run when they toddle down the road to your competitor.

Enterprise storage vendors understand all of that and have invested significant amounts of research and development (R&D) to make robust systems, both in their storage targets and in their storage networks, that keep operating in the presence of the inevitable rare glitch.

The vendors have also worked hard to provide multipath I/Os so you have no single point of failure, and you get the added benefit of tiny or even zero service upgrade windows. Customers have effectively forced the storage vendors to do this by voting with their wallets. Vendors must be prepared to deliver 24/7 enterprise support for their own products as well as other products that they sell.

The enterprise storage vendors understood their market, even in the mid-1990s when Ethernet was beating up on other protocols. Despite the traction of Ethernet, the vendors chose FC — for many good reasons.

Features of a lossless network

Whether you lose a dog or your car keys, losing something you value is a terrible thing, and so you take care of those things. Other things, like a paper soda straw — not so much. If you lose it, you'll get another one.

FC has always been a lossless networking technology. It treats a payload like a loved one. Every FC link is governed by buffer credits that the receiver shares with the sender. The sender knows how many buffers are available at receiving side and doesn't send a frame that the receiver can't handle. By contrast, Ethernet has been a lossy network for decades, treating packets like soda straws, dropping packets in a whole variety of situations and relying on TCP or other mechanisms to replace the lost straws. Data Center Bridging (DCB) is an Ethernet variant that uses PAUSE frames (instead of buffer credits) to avoid packet loss, but DCB still has some notable interoperability challenges.

Security

Obviously, if you're treating your frames like loved ones, you need to protect against inappropriate human actions as well, whether those are simple mistakes or something more troubling. FC offers extra security in a number of ways. FC is already trusted to keep the world's most important data secure, and FC brings this security model to NVMe over Fabrics (NVMe-oF).

REMEMBER

One key advantage of FC comes from its specialized nature. Datacenter-centric FC isn't the protocol of the internet, so hackers can't shove FC frames across the internet to sneak past your firewall and into your datacenter. FC also provides a zoning service that integrates storage access controls into the network. This tried-and-true service works across all enterprise storage vendors, even in multivendor environments.

Through the modernization process, enterprises are facing the reality that their revenue is intertwined with the success or failure of the IT organization. To succeed, IT needs to automate as much as possible to eliminate complexity and reduce costs. IT organizations need to remove tedious, time-consuming, and labor-intensive tasks so they can focus on delivering services to the business that can help deliver additional revenue. Fibre Channel tools deliver automation and intelligence to admins so they don't have to worry about the SAN or dual protocol running and instead can focus on initiatives that are strategic to their organizations.

IN THIS CHAPTER

- Figuring out your starting point
- Making the move to NVMe-oF
- Deploying NVMe-based arrays in production
- Getting to know NVMe over Fibre Channel

Chapter 3
Adopting and Deploying NVMe over Fibre Channel

You've done all your homework. You've read a bunch of manuals, gone to a few seminars, and talked to colleagues. Everyone on your team agrees it's time to move forward with Non-Volatile Memory Express (NVMe) over Fibre Channel adoption in your organization. Only one big question remains: Where do you start? This chapter helps.

Identifying Your Situation

Several factors work in your favor when you implement NVMe over Fibre Channel. You don't need to divide your budget and invest in parallel infrastructure, nor do you have any worries about multivendor interoperability of equipment or new protocols and discovery algorithms to grapple with. You don't have to risk education funds on uncertain internet protocol (IP) and Ethernet NVMe protocol, as you may with competing technologies. That

being said, before going further, you should check a few legacy infrastructure boxes:

- You can deploy NVMe over Fibre Channel on an existing Fibre Channel (FC) infrastructure, provided it's relatively up to date — such as the Fabric Operating System (FOS) 8.1.0 or later. Check with your hardware supplier on interoperability.
- The FC fabric must be Gen 5 (16G) or better yet Gen 6 (32G). Of course, Gen 7 (64G) also supports NVMe over Fibre Channel, and this version incorporates dramatic switching latency improvements over Gen 6, which greatly benefits NVMe workloads.
- Servers using NVMe over Fibre Channel need Gen 6 or Gen 7 host bus adapters (HBAs); Gen 6 or 7 HBAs also work with Gen 5 fabrics.
- Choose a storage device that supports NVMe over Fibre Channel frame types. This can be an NVMe over Fibre Channel-enabled array, or, for early familiarization, a server with an NVMe over Fibre Channel HBA running in target mode can fill the role of storage target.

None of these requirements are unreasonable. For example, if you have servers running performance-sensitive applications and you're still on Gen 4, it's definitely time for an upgrade, regardless of any NVMe over Fibre Channel undertaking.

Considering Your Adoption Strategy

Setting aside the glorious predictions of the NVMe hype-masters, few in the storage and networking communities would dispute the notion that the move to NVMe over Fabrics (NVMe-oF) will be gradual — lasting several years. Unfortunately, this slow transition makes it painful to have a classic FC network sitting side by side with your brand-new Ethernet-based NVMe network. It presents several questions:

- When you buy more storage, what kind do you buy?
- As you build new apps, which environment do you connect them to?

- If you go with Ethernet, which kind: iWARP, RoCEv2 (which doesn't use TCP), or NVME over Transmission Control Protocol (TCP)?

None of these are the obvious solution or have much of a historical adoption record. However, adopting a dual-protocol FC fabric (running concurrent FC Protocol [FCP] and NVMe over Fibre Channel traffic) eliminates or simplifies those questions (see the later section "Exploiting Dual-Protocol FCP and NVMe over Fibre Channel" for more information). FC is widely adopted and enjoys excellent support by storage vendors and other technology providers, making NVMe over Fibre Channel practically a zero-risk proposition.

Protecting high-value assets

Although NVMe over Fibre Channel is new technology, for most storage-oriented usage (as opposed to "working memory"), the goal is to apply the technology to an existing storage asset. With concepts like Big Data Analytics, data mining, data lakes, artificial intelligence (AI), and machine learning (ML) gaining traction, the value of everyone's data assets is climbing rapidly. This makes a top-notch, high-performing storage solution even more crucial than in days past and at the same time highlights the importance of keeping risk to a minimum.

If an application is merely using a copy of a data asset (not modifying or updating the master copy), the application is effectively treating this asset as working memory. Conversely, if the application maintains the master copy of the data, maintaining integrity and availability of the data asset is vital.

Most often, when an existing data asset is involved, you are looking at a "brownfield" (as opposed to a "greenfield") scenario. *Greenfield* is a totally new build where none existed before. *Brownfield* is used to describe an existing environment that should be updated. Brownfield is a newer term in the last 5 to 7 years.

Migration from legacy architecture, which is typically built on a Small Computer System Interface (SCSI) infrastructure, to one based on NVMe should be done in an incremental fashion after lab validation, and allow for the option of rolling back architectural changes. The ideal adoption strategy includes a process and infrastructure that allow for such a model.

Allowing for a marathon shift

With all the buzz about the speed of NVMe and the rapid rise of all-flash arrays, won't the whole world go to NVMe storage by the end of next week? Probably not. Consider how long the transition was for backup media from tape to disk regardless of all the performance advantages with disk. The transition from tape to disk (and lately solid-state drives [SSD] media) progressed for more than ten years. During this period, disk drives continuously increased in capacity with decreased cost; consequently, disk has displaced tape as the mainstream backup media solution. Even if the adoption of NVMe is substantially faster, the transition will take years.

Realistically, some firms or some departments will adopt NVMe rapidly when they're interested only in working memory and have no urgent need to maintain high-value data assets. Often such use cases will be handled with direct-attached NVMe products only, and fabrics won't be required during this phase. Other firms that don't have a compelling need to accelerate their working memory may first adopt NVMe for storage use cases, and of course, there will be some combinations. The point is, there will be different adoption rates for different kinds of usage.

Exploiting Dual-Protocol FCP and NVMe over Fibre Channel

IT people are responsible for company data and productivity and, therefore, are rightly concerned about risk reduction (*de-risking* in geek speak). As IT departments plan to deploy NVMe-based arrays in production, they have two main options.

Option One is creating a new NVMe infrastructure, as shown in Figure 3-1. If they build a new, separate infrastructure, every array purchase after that becomes a gamble because they must choose which infrastructure will access the array. This method isn't recommended.

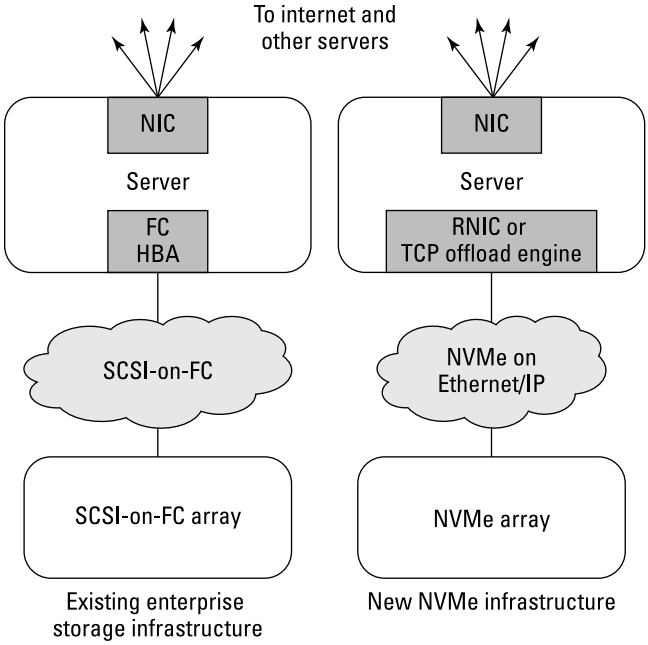


FIGURE 3-1: Building a new separate infrastructure (not recommended).

Option Two — dual-protocol concurrency — is the big win for NVMe over Fibre Channel. By leveraging an existing infrastructure, as shown in Figure 3-2, this "known quantity," such as an FC storage area network (SAN), helps companies easily support dual protocols and removes any risk associated with the inevitable question, "How long will the transition take?"

A dual-protocol approach isn't unreasonable. A strong precedent exists for mixing protocols on top of FC. Since 2001, the established practice supports both FCP and FICON traffic simultaneously on the same FC fabric (see the sidebar, "FICON and FCP" for more info). In fact, with the release of FOS 9.1, there will be support for isolating SCSI traffic from NVMe traffic to allow the best performance of the protocol even over the same link.

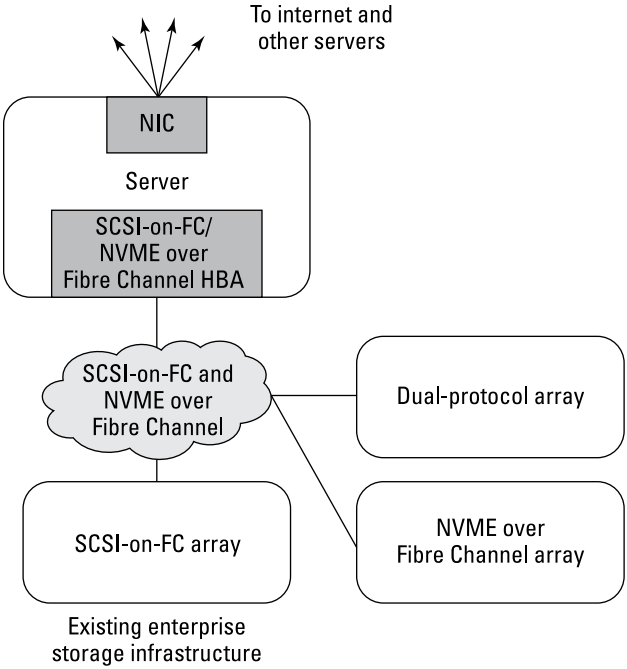


FIGURE 3-2: Leveraging dual-protocol infrastructure (recommended).

Other considerations are

- **Incremental migration:** Often a single application uses many storage volumes and may have different requirements on those volumes. In a dual-protocol environment, individual volumes can be migrated as appropriate. High-value, risk-intolerant assets can remain on trusted infrastructure. Lower-value, latency-sensitive volumes can move to the hottest new targets. Changes can be rolled back easily. In a dual-protocol environment, these changes can be achieved administratively without making disruptive hardware or cabling changes.
- **Dual-protocol publishing:** Master copies of high-value data assets may be maintained on trusted legacy arrays and regularly "published" to the latest rocket-fast NVMe over Fibre Channel arrays. This process allows other applications to use the data as "working memory." Again, this task can be achieved with existing infrastructure.

FC lets IT administrators leverage already familiar elements such as zoning and name services, so no learning curve is required. In addition, dual-protocol concurrency offers opportunities that are otherwise difficult to achieve.

FICON AND FCP

FCP is the SCSI-on-FC protocol used in open systems such as Windows and Linux. FICON is the FC version of a mainframe storage protocol and is the IBM proprietary term *Fiber CONnection* or the mapping of the ANSI FC-SB-3 standard "Single-Byte-Command-Code-Sets-3" over FC. It's how mainframe uses FC. FICON is essentially "ESCON-on-FC" (ESCON stands for Enterprise Systems CONnection and is how mainframes spoke over copper prior to FC). Many firms that use mainframes need to publish mainframe data assets so they can be consumed by open systems. At other times — for example, when consolidating data assets after a merger — they may have a need to ingest an FCP data asset into the FICON world.

To achieve the transfer of assets, special migration servers are configured with dual-protocol (FCP and FICON) HBAs connected to a dual-protocol SAN that connects to both FCP storage arrays and FICON storage arrays. The reality of the FCP/FICON dual-protocol SANs is that they aren't aimed at a long-term transition from one to the other but offer an enduring "bridge" between the two realms. Because the mainframe and open systems are architecturally different, there's no notion of a gradual migration of applications from one side to the other, so normal application servers aren't configured to use both protocols.

Concurrent dual-protocol FCP/NVMe over Fibre Channel is different because both protocols are designed for open systems, and, consequently, you can readily plan for gradual, incremental, non-disruptive migration of an application from one protocol to another. Some apps may benefit from an ability to operate in a dual-protocol mode for an extended period of time, consuming (reading) assets using one protocol and publishing (writing) assets on the other. You may choose to transition other applications relatively quickly, within a matter of days or weeks, and those may spend relatively little time in dual-protocol mode. Either way, the ability to leverage dual-protocol FC gives you tremendous flexibility as you move to deploy NVMe in production.

Zoning and name services

Network administrators use *zoning* to improve security. Two approaches to zoning are available:

- **Port zoning** restricts access based on the specific port of the FC device to which a node is attached.
- **Name zoning** restricts access based on a device's World Wide Name (WWN).

Each method restricts devices from accessing network areas they shouldn't be visiting. The approach that makes the most sense for you depends on your usage. Name zoning usually reduces maintenance effort, except in cases where a sequence of different devices is connected on one port.

You may come across references to *hard* and *soft zoning*, as well. Modern FC fabrics use hard zoning, in which robust silicon-based logic blocks traffic between nodes that aren't allowed to reach each other. In contrast, early products used software for soft zoning. Unable to block traffic, the software hid information. It was like covering your address instead of locking your door. Sensibly, soft zoning is no longer used.

REMEMBER

FCP and NVMe over Fibre Channel can both leverage FC zoning. FC's zone services are implemented in the fabric, which is a different approach from those used by competing technologies. Consequently, the results are more predictable and easier to manage, reducing the opportunity for security holes. Name services, on the other hand, translate obscure computer and device addresses into human-friendly names. They're similar to Domain Name Services (DNS) in Ethernet but are network resident, which greatly simplifies interoperability and management. This has long been the case for FCP and remains true for NVMe over Fibre Channel.

Discovery and NVMe over Fibre Channel

The NVMe-oF specification described a discovery mechanism, but left many details up to the specific implementation. For non-FC

fabrics, this leaves a gaping interoperability chasm, which will likely be as slow to close as previous interoperability challenges such as Priority Flow Control (PFC) and Data Center Bridging Exchange (DCBX). Additionally, without a broadly adopted discovery mechanism and name service in Ethernet fabrics, the NVMe deployments will suffer the same issues there that they do with iSCSI, which is to say that recovery of storage connectivity when a network disruption occurs tends to be manual. Not exactly the best performance and cost option.

FC's ability to deliver a dual-protocol FCP/NVMe over Fibre Channel fabric provides a clearly mapped forward path to interoperability. HBA vendors are creating drivers that leverage FCP for device discovery, then check those devices for NVMe over Fibre Channel traffic support. Enterprise storage vendors who offer SCSI-over-FC arrays are motivated to support this two-step approach as well. Newcomer NVMe array vendors interested in the FC market can leverage the NVM Express standard mapping from SCSI to NVMe, and they may follow the model as well in order to appeal to existing FC customers.

Familiarizing Yourself with NVMe over Fibre Channel

Someday NVMe over Fibre Channel will be an old friend. For now, at least, it's more like a first date. You're not sure how it's going to react to certain stimuli, what level of attention it requires, or whether it's going to unexpectedly overreact to a stressful situation. Go easy. You've already learned all you can about this exciting technology, so now it's time to roll up your sleeves and start playing. Set aside time to become familiar with the nuances of NVMe over Fibre Channel. Go through the various commands and determine how your setup is going to work. Then look beyond your test environment, giving thought to how NVMe over Fibre Channel will fit into your production system.

Experimenting in your lab

Like Dr. Frankenstein, you may be tempted to shout, "It's ALIVE!" the first time you fire up your NVMe over Fibre Channel test environment. Try to resist this level of enthusiasm because it may give people the idea that you were surprised by your success. Instead, calmly nod your head and say, "Yep, that's what I'm talking about," and go get yourself another energy drink.

Not sure where to start? When establishing an NVMe over Fibre Channel test lab, a typical IT department may take the following steps:

1. **Set up a single server with an internal drive, a single switch, and a single NVMe over Fibre Channel-enabled array.**
2. **Connect the server to a lab IP network for access to other lab servers.**
3. **Explore the HBA config and storage array options.**
4. **Configure a volume to use the namespace ID (NSID).**

   The details depend on your NVMe array management tools, but in general, this procedure is similar to provisioning a logical unit number (LUN).
5. **Copy a file from the internal drive to the NVMe volume and back again.**
6. **Run your preferred performance testing application (such as FIO or Iometer) to benchmark your NVMe volume.**

   Compare it to your internal volume.
7. **Lather, rinse, and repeat until your paranoid nature is quelled.**

Migrating your LUN to a namespace

Moving massive quantities of data from one storage technology to another is never much fun. This is especially true when you aren't completely sure of the process. Start small. Migrate a volume or

two from SCSI to NVMe (from LUN to NSID) to become confident in the process.

You should also run some applications on the NVMe volumes (namespaces) in the lab to build up that muscle memory and be certain you haven't forgotten any steps. At this point, you should have a warm fuzzy feeling that you understand how it all works.

Next, expand that comfort level. You've done the basics, so go ahead — open all the storage management apps, SAN management applications, and analytics tools, such as Brocade SANnav, Brocade SAN Health, and Brocade IO Insight. Ensure that these tools have been upgraded appropriately and are NVMe over Fibre Channel-enabled. You should also make sure you're familiar with the alterations that NVMe over Fibre Channel brings to these applications. Lastly, give some thought to which of these tools and features you'll need as you bring NVMe over Fibre Channel online in your production environment.

As with any high-profile production change, think about doing "baseline" performance measurements. Bear in mind that a few hours after flipping the switch to production, you're likely to get a support request because some traditional application is having issues. Was it just a routine human error or an actual hiccup? If it's a hiccup, did your flipping of the switch cause it? Be prepared! You should have enough information by now to recognize whether the support request is linked to the recent change. The baselines you made earlier will help.

Even in the absence of a support call, you want to understand how much performance has improved by moving from SCSI to NVMe because that information helps you evaluate and prioritize other moves. Plus, when you get all those emails from jubilant users thrilled with the speed of their applications, you'll be able to tell them precisely how much faster the system is running. Okay, we know that isn't going to happen, but it may earn you a "Job well done!" from your boss. If you don't know the initial performance, you won't be in a good position to pump your hands overhead, thus broadcasting your NVMe over Fibre Channel rock star status.

It's all well and good that you're now an NVMe over Fibre Channel Jedi Knight, but what happens when you go on vacation or problems arise during the night? Unless you like receiving phone calls while skiing the slopes of Aspen, you'd better make sure the rest of your crew is up to speed and good documentation is in place before going live.

Transitioning to production

Hold on tight, you're going live! Setting aside all the nail biting and late-night pizza parties, moving a test system to production is an exciting time. Because your chest now swells with confidence and understanding, it's time to start selecting and prioritizing which applications and volumes are the best place to begin the rollout.

Just as you did in the lab, make sure that all your management tools have been properly refreshed for NVMe over Fibre Channel. You don't want to start a cross-country journey and then realize that you've forgotten the map, the gas tank is empty, and the rear tires are low on tread. You're certainly going to be eager to share the fruits of your labor, but be sure not to shortchange this last — and in many ways most important — part of the process.

Some of this will be drastically simplified for you by the native support of NVMe over Fibre Channel in VMware vSphere 7.x releases. As of the Update 2 (U2) release the performance of VMware on NVMe over Fibre Channel is enthralling. One can provision an NVMe Name Space ID (NSID) on the production array and present it to VMware.

REMEMBER

Be careful to make certain that the zone configuration allows access the storage administrator can then hand over control to the vSphere administrator. That admin can choose the platforms they believe will best benefit from the NVMe technology and perform a live Storage vMotion of that platform from the datastore talking to the SCSI LUN to the mirror of that datastore talking to the NVMe NSID without having to take the application offline. Current verified benchmarks from VMware for this provide

examples of as much as an 80 percent increase in IOPS on a production database without downtime. That effectively makes this the most risk-averse migration in the IT market in more than 20 years.

Finally, schedule the cutover for a time that won't interfere with your customers (those people whose eyes glaze over when you tell them to reboot), and make certain everyone in the company knows about the looming transition. There should be no surprises for anyone (especially you) and all should enjoy the journey down the NVMe over Fibre Channel superhighway.

IN THIS CHAPTER

- Sharing dynamic information with RDMA
- Assessing Ethernet-based NVMe

Chapter 4
Comparing Alternatives to NVMe over Fibre Channel

In most cases, having alternatives is a good thing. This is true whether you're deciding which color carpeting to install in the master bedroom or which way to get home at rush hour. Non-Volatile Memory Express (NVMe) over Fabrics also offers alternatives, although some may not be to your liking. It can run over fabrics such as iWARP, RoCEv2, or InfiniBand, or simply NVMe over Transmission Control Protocol (TCP). This chapter looks at a few of the pros and cons of each and examines performance considerations such as wire speed, architecture and virtualization, and which special features are supported. Of course, performance is meaningless in the face of risk, so this chapter also evaluates predictability and potential disruption.

The Long and Short of RDMA

Remote direct memory access (RDMA) is a protocol designed for use in tightly coupled server environments, especially those that fall into the high-performance computing (HPC) category. If humans used RDMA to communicate, there would be no need

for speech or body language — thoughts and emotions would be shared directly, brain to brain, greatly increasing communication speed and eliminating any chance of misinterpretation.

Fortunately, our brains aren't computers, and we can all keep our thoughts to ourselves. For clustered server applications, however, RDMA is a great way to share dynamic information. One server effectively gives ownership of some part of its memory to a remote server. For many multi-server applications, especially those involving dynamically changing data, this method offers significant performance advantages.

ZERO-COPY FANFARE

When the TCP stack was being developed during the 1980s, a good variety of networking technologies existed. The stack was therefore designed to work with whatever was available, whether it was Token Ring or a phone line. Including clean networking layers made perfect sense for interoperability, and one way to achieve that was the use of intermediate buffering, thus making buffer copies commonplace. As speeds increased, however, most buffer copies were optimized away, except in cases where that practice would break backward compatibility.

In the mid-1990s, a good networking stack could claim single-copy efficiency. A network adapter received frames and wrote them (using DMA) into DRAM buffers associated with the networking stack. (The unavoidable DMA step isn't a DRAM-to-DRAM copy, so it's not counted.) The networking stack would first process the frame, and then copy the "payload" to the memory location desired by the high-level application. For a period of time, this single-copy architecture seemed fully optimized.

Yet by the time FC was being "productized," the game had begun to change. Fibre Channel's main claim was speed, so pressure to optimize was high. Chip technology allowed for more complexity, and the Fibre Channel/SCSI stack was not constrained by the same backward compatibility challenges that IP stacks faced. FC was focused on one "application" (storage), and had a simpler layer structure than TCP/IP/Ethernet. For all these reasons, FC was more motivated, and more able to implement a network adapter/driver/stack architecture that eliminated the single copy. And that's precisely what happened. FC has been quietly delivering "zero copy" for the past two decades.

NVM Express specification for NVMe over Fabrics (Revision 1.1a) describes two types of fabric transports for NVMe (in addition to PCIe): NVMe over Fabrics (NVMe-oF) using RDMA and NVMe-oF using Fibre Channel (FC). Despite this explicit recognition of FC as an NVMe fabric, some RDMA advocates claim that because NVMe over Fibre Channel doesn't use RDMA, it somehow isn't an NVMe fabric, despite the fact that NVMe over Fibre Channel doesn't need RDMA. In addition, at a later stage, a third type of fabric was added to the NVMe-oF specification — one also based on Ethernet and IP but *not* based on RDMA: NVMe over TCP. This cements the idea that RDMA isn't really required to be able to run NVMe over a particular fabric technology. FC uses native direct placement capabilities while also enabling the dual-protocol support that allows for a low-risk transition from SCSI to NVMe. If you have any reservations about those claims, hang on tight: We're about to defuse them in this section.

For more information about the NVM Express specification for NVMe over Fabrics, Revision 1.1a, visit `nvmexpress.org/wp-content/uploads/NVMe-over-Fabrics-1.1a-2021.07.12-Ratified.pdf`.

InfiniBand

InfiniBand (IB) came on the scene more recently than Ethernet or FC. Focused on server cluster communication, IB delivers RDMA natively over a network and focuses on speed instead of mainstream adoption. Special adapters and switches are required to use IB, which is one reason it never reached broad acceptance or partner compatibility except in specialized HPC applications. In fact, only one IB chip supplier at the time of this writing offers InfiniBand products, a factor that discourages new adopters and raises questions about the cost of switching to the InfiniBand protocol. That sole IB chip vendor seems to have recognized this reality, and has therefore focused its NVMe-oF marketing efforts on promoting RoCE.

iWARP

Though not widely deployed, internet wide-area RDMA protocol (iWARP) is nearly 12 years old and is an Internet Engineering Task Force (IETF) standard described in five Request For Comments (RFCs) in 2007, and then updated by three more RFCs as recently as 2014. iWARP is designed to run on top of TCP, which is

categorized as a reliable streaming transport protocol because it includes various techniques to ensure that every byte that's sent has been received by the sender.

WARNING

iWARP's TCP basis isn't ideal for storage because TCP normally ramps up transmission speeds slowly. Because many storage applications have traffic patterns that include so-called "bursty elephant flows," the slow-start behavior of TCP leads to latency challenges and reduced input/output (I/O) operations per second (IOPS) metrics.

TCP has been extended a number of times, and newer versions of TCP stacks have various configurable (and sometimes negotiable) features that older versions lack. Early TCP implemented a "slow start" mechanism that began transmissions slowly in order to avoid overflowing network buffers. If protocol timeouts or receiver acknowledgment messages indicate that some transmissions were not received, traditional TCP retransmits and backs off on transmission bandwidth (it "collapses the TCP window"). Some newer versions of TCP, such as Data Center TCP (DCTCP), have features that work better in a datacenter environment, but these features aren't compatible with wide area network (WAN) usage. This is why datacenter architects and implementers face a challenge if they want to use TCP for high-performance use cases and are faced with one of three options:

- Choose a single complex TCP stack that can be configured differently for different use cases and mandate this complex stack across all operating system (OS) images.
- Choose two or more TCP stacks and manage which image gets which TCP stack.
- Have some OS/hypervisors images that get dual (or more) TCP stacks, mapped internally (probably by IP address) to the desired usage.

Each option is problematic, increasing complexity and placing additional burdens on network administrators.

One final drawback comes with performance. TCP was designed to work across a wide range of networks, which of course includes wide area networks (as referenced in the iWARP acronym). But in order to be effective across a variety of networks, TCP has been designed to attempt to minimize lost messages that can occur

when the sender sends too fast, which generally means "slow down." It's for these reasons among others that iWARP hasn't been well adopted by the networking community.

As a reliable streaming protocol, stand-alone TCP was never intended to guarantee packet or frame alignment. That's because TCP doesn't send a set of packets but rather a stream of individual bytes, removing any certainty that commands placed at, say, byte 8 of the packet will be processed in a timely manner, as the entire stream must first be decoded. TCP is a complex, software-based stack so the alignment question would've stumped the hardware processing of iWARP. To address the alignment problem, one of the iWARP RFCs (RFC 5044) created a fix ("Marker PDU Aligned Framing," or iWARP-MPA) that allows for packet alignment at the cost of increased stack complexity. This is strongly reminiscent of the burdensome Small Computer System Interface (SCSI) stack that NVMe was created to replace.

Yo, Rocky

RDMA over Converged Ethernet, version two (RoCEv2) is an odd standard in that it was developed by the InfiniBand Trade Association rather than the Internet Engineering Task Force (IETF) or the Institute of Electrical and Electronics Engineers (IEEE), where most IP and Ethernet standards are developed and maintained.

The RoCEv2 (pronounced "rocky vee two") name says "RDMA over Converged Ethernet," but it's a slight misnomer with version 2, which runs over User Datagram Protocol (UDP) so it's no longer directly coupled to Ethernet. For performance and reliability, however, RoCEv2 recommends Converged Ethernet, a dated term for a lossless Ethernet network. Lossless Ethernet is now more formally referred to as Data Center Bridging (DCB) which includes such interdependent features as Priority Flow Control (PFC), Enhanced Transmission Selection (ETS), and Data Center Bridging Capabilities Exchange (DCBX).

DCB is an ongoing effort to enhance Ethernet with the features designed into FC during the mid-1990s. Although this is a laudable goal, the interoperability of real-world DCB deployments remains rather low, which interacts problematically with the forgiving nature of Ethernet/IP. A misconfigured DCB network can easily go undetected for long periods, operating as a classic best-effort Ethernet network and suffering random packet loss during traffic spikes.

For a convenient comparison of the differences between Ethernet-based options for NVMe and FC, see Table 4-1.

TABLE 4-1 Ethernet-Based versus NVMe over Fibre Channel

Ethernet-Based Options for NVMe	NVMe over Fibre Channel
Developing a new fabric protocol standard	Built on T11-standardized FC fabric protocol
Standards group dealing with challenges of scaling I/O commands, status and data to the datacenter	FC solved these problems when FC protocol (FCP) was developed for SCSI
Transport options: iWARP, RoCEv2 and TCP	Transport is FC; runs on existing Application Specific Integrated Circuit (ASIC)
iWARP and RoCEv2 use RDMA (TCP doesn't)	Doesn't need RDMA, leverages FCP
Complex network configuration if RDMA is enabled	FC is well understood
New I/O protocol, new transport	New I/O protocol, existing reliable transport
Low latency if RDMA used with RNICs	Same zero-copy performance as with FCP
Discovery and zoning services are still in proposal phase	Leverages tried-and-true fabric services

Evaluating Ethernet-Based NVMe

Choosing Ethernet for your storage network's physical layer is challenging. The industry is looking at four Ethernet-based NVMe protocols:

- iWARP
- RoCEv2
- NVMe over FCoE
- NVMe over TCP

To get low latency with iWARP or RoCEv2, you need to install RDMA-enabled Network Interface Cards (RNICs). Big operators in social media will choose the commodity-oriented NVMe over TCP (parallel to today's internet Small Computer Systems Interface [iSCSI]), even if it's slower. Of course, you must buy NVMe arrays that support your fabric choice, but who knows which will win?

Out of all these NVMe over Ethernet technologies, two have positioned themselves as the main challengers to take the Ethernet crown: RoCEv2 and TCP. iWARP never had significant traction even for the use case it was developed for in the first place (HPC), and it's not expected to gain any traction for NVMe. In fact, at the time of this writing, there are no storage vendors supporting NVMe over iWARP on their all-flash array host ports.

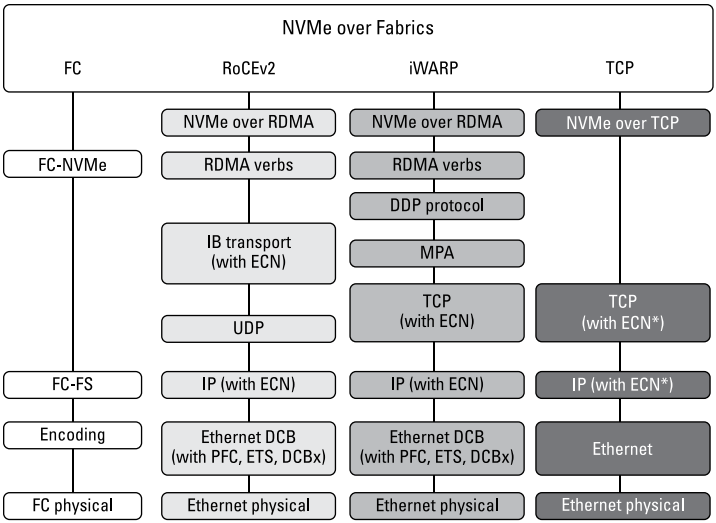
NVMe over FCoE is a niche use case specific to some blade server environments, essentially the only place where FCoE took any significant foothold in the datacenter, and is only supported by one vendor. Plus, FCoE is ultimately FC, and therefore a server running NVMe over FCoE at the end of the day communicates with a storage array running NVMe over native FC. And while RoCEv2 and TCP both run on top of Ethernet, there's no compatibility between these two protocols, meaning that a server that's running NVMe on top of RoCEv2 can't communicate with a storage array that runs NVMe over TCP and vice versa. Therefore, even if the options are reduced, choosing any Ethernet-based NVMe protocol is risky.

REMEMBER

Contrary to the recommendations of the NVM Express white paper on NVMe over Fabrics, Ethernet flow control doesn't use the reliable credit-based flow control mechanisms found in FC, PCI Express (PCIe), and InfiniBand transports. Whether you choose iWARP or RoCEv2, you're choosing a multilayered network, with an associated increase in stack complexity to transport NVMe. Take a look at Figure 4-1. It shows different NVMe fabric stacks.

Ethernet advocates tout benefits like jumbo frames, even though authorities such as Demartek recommend disabling jumbo frames when using RoCEv2. Some datacenters are using VXLAN, which adds extra Ethernet and IP headers and requires extra management of the "maximum Protocol Data Unit" (MaxPDU) setting for every network port. MaxPDU affects IP fragmentation, which in turn affects IPv4 and IPv6 differently. All these layers are required in part for legacy reasons, and in part because

Ethernet/IP is designed for internet scale rather than the data-center scale of Fibre Channel. Opting for a complex multilayer transport is rather an odd choice when the key benefits of NVMe derive from its simplified stack.



* ECN is not required for NVMe over TCP, but it is the prevalent method considered to avoid dropped packets for NVMe over TCP.

FIGURE 4-1: Relative complexity of different NVMe fabric stacks.

Commodity or premium?

Ethernet wins as a great low-cost, best-effort technology. It's easy to deploy for common uses and supports myriad upper-layer protocols and applications. Ethernet's plug-and-play behavior was designed for widespread use, and tens of thousands of well-informed technicians in the industry know how to manage that mainstream Ethernet configuration. Ethernet has simplistic, robust mechanisms like Spanning Tree Protocol that shut down links and guarantee there are no loops that can cause problems with broadcast or multicast storms. These issues might otherwise be commonplace, because broadcasts are a normal part of address learning.

Unfortunately, tree-based topologies aren't ideal in today's data-center traffic flows, so companies tend to use IP routers at the top of server racks. Much of the reason Ethernet is so easy to deploy is that its main customer, TCP/IP, is so resilient and forgiving, at

the cost of modest performance. Layer 2 Ethernet doesn't scale too far, but coupled with IP, it can scale to the internet. Ethernet and IP can also be bought anywhere. Interesting products are available on eBay and Amazon, making it a highly competitive marketplace. FC networking products are largely available from storage vendors, and it's not always easy playing vendors against one another for lower pricing.

Smart shopping

Most storage vendors have invested a significant amount of time in testing their arrays with the networking products they sell. They're familiar with all the enhanced analytics and visibility features that have been designed into the ASICs and the management software. The storage vendor understands the network-resident features like FC Name Services and FC Zoning, including target-driven zoning, features that aren't yet defined for Ethernet-based NVMe fabrics. These vendors are comfortable handling support questions related to the tried-and-true interplay of servers, HBAs, storage arrays, and networks.

TIP

If you acquire your Ethernet/IP networks from the lowest bidder, consider what the support picture will look like when you call the storage provider about some strange issue. Where do you begin? How do you inspect the network to even begin to tackle the problem? Despite widespread recommendations that enterprise storage should be deployed on a dedicated network, IP storage is frequently connected to a shared network. In light of that, consider surveying your existing Ethernet/IP network and evaluating whether you would be able to support a storage service-level agreement (SLA) on such a network.

The widespread success of Ethernet and IP is a double-edged sword. Many of the aspects that make Ethernet and IP widespread and commoditized are problematic when you need a more specialized, premium network. These two protocol suites are obviously the leading answer for the internet, the campus, homes, and mobile devices, places where FC doesn't make sense. Ethernet and IP even work well in the datacenter for multiprotocol best-effort communication needs. However, enterprise storage in the datacenter is a more demanding use case. That's why "good enough" is anything but, and investing in valuable assets makes sense.

IN THIS CHAPTER

- Knowing what performance improvements to expect
- Influencing your SAN design
- Seeing the importance of monitoring
- Altering how you zone
- Explaining ANA and why it matters
- Exploring application use cases
- Understanding not all fabrics are created equal
- Ensuring performance during congestion

Chapter 5
Improving Performance with NVMe over Fibre Channel

When you're ready to test the waters with Non-Volatile Memory Express (NVMe) over Fibre Channel, what level of performance improvements can you expect compared to Small Computer System Interface (SCSI)/Fibre Channel Protocol (FCP), and why? To help you answer this question, this chapter looks at where you can expect performance improvements to come from. NVMe is designed for storage with characteristics similar to those of memory and consequently is a more efficient protocol than SCSI. This chapter then shows you how this method applies in an end-to-end solution from the application running on the server through the Fibre Channel (FC) storage area network (SAN) to the storage array.

Understanding How NVMe over Fibre Channel Improves Performance

Figure 5-1 shows an end-to-end storage chain extending from an application on the host to storage media on the storage array. NVMe over Fibre Channel contributes to better performance in the following areas:

- **Host side:** How NVMe over Fibre Channel performs on the server compared to SCSI/FCP
- **Storage array front end:** How NVMe over Fibre Channel performs better on the storage array target ports compared to SCSI/FCP
- **Storage array architecture:** How the storage array architecture handles NVMe compared to SCSI/FCP
- **Storage array back end:** How using NVMe attached solid-state drives (SSDs) in place of Serial Attached SCSI (SAS) and Serial Advanced Technology Attachment (SATA) SSDs improves performance by removing the translation from NVMe to SCSI

This section covers each area of Figure 5-1 in more detail.

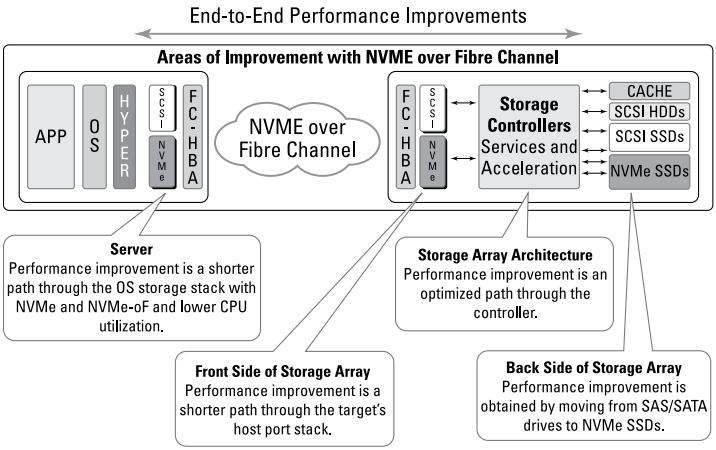


FIGURE 5-1: Areas of performance improvement with NVMe over Fibre Channel.

REMEMBER

You may be thinking, "Wow! But wait, what about the FC SAN?" Remember that the SAN supports SCSI/FCP and NVMe over Fibre Channel equally, and there's no difference in performance across the FC network between transporting FCP and NVMe over Fibre Channel. The transport happens with the same low latency and high performance that you're used to. Still, enhancements are being added to the FC standard to augment high-performing NVMe over Fibre Channel in case low-level errors occur (check out Chapter 6 for more on Sequence Level Error Recovery and other new benefits of NVMe over Fibre Channel).

The host side

On the host side, you see performance improvement because the NVMe protocol is more lightweight than SCSI, resulting in a leaner driver stack that executes more quickly by using fewer resources than SCSI. NVMe over Fibre Channel requires less CPU processing time than FCP for the same workload, which gives you more CPU cycles for your applications.

The storage array front end

On the storage array front end, you can expect the same sort of driver performance improvement as on the server side. With the more lightweight NVMe protocol compared to SCSI, getting to the storage controller happens quicker. Sometimes you hear this improvement described as a "shorter path" through the driver/protocol stack.

Storage array architecture

Traditionally, storage array controllers increased performance by striping input/output (I/O) across spinning media, and they provided storage services, including data protection through encryption and dedupe. The array controllers did this without adding any noticeable latency because the media was so much slower than the storage services running on the controller.

This practice is changing with SSDs, particularly with NVMe attached SSDs. Instead, storage services are becoming visible from a latency perspective. New array architectures entering the

market are designed to keep the array controller out of the data path and offer the ability to deselect storage service for applications with low-latency requirements.


TIP

You can switch to NVMe over Fibre Channel with an existing array for many compelling reasons. Investment protection and the ease with which NVMe over Fibre Channel is made available — with the first products in market — by simply upgrading the storage array to the latest firmware level makes it a straightforward approach to adopt and implement NVMe over Fibre Channel.

The storage array back end

With many arrays in the market today still predominantly using SAS/SATA attached SSDs, you can reap performance improvement by using NVMe attached SSDs, which makes sense only when the overall array architecture can deliver the performance end-to-end through the array. Using the latest flash media technologies commonly referred to as *storage class memory* (SCM), such as 3D-XPoint (pronounced "three dee cross point"), makes sense in arrays designed to deliver the utmost low latency and highest I/O operations per second (IOPS) end-to-end.

PCIe GEN4

A new technology that's become recently available is PCIe Gen4, which will play a role in new storage array design decisions. Since mid-2018, new NVMe SSDs entered the market that are PCIe Gen4 compatible. The PCIe Gen4 standard doubles the transfer rate currently available with PCIe Gen3. Now that PCIe Gen4 x86 motherboards are available, it's just a matter of time before new NVMe storage arrays will enter the market. These arrays will better utilize the performance improvement of each individual SSD by doubling the bandwidth per PCIe lane by using Gen4 instead of Gen3. When looking at the balance of the environment, consider where the bottlenecks will occur. Given that the individual Gen 4 PCIe card is a 64G pipe back to the CPU and memory complex, would you throttle it down to 25GE by putting on Ethernet? Or would you maintain a 1:1 throughput by implementing Gen 7 Fibre Channel?

Handling NVMe support with a software upgrade

The first storage arrays in the market to deliver NVMe over Fibre Channel made it easy for you to start using NVMe over Fibre Channel. All you needed was a simple software upgrade of the storage array controllers to the latest firmware version, and off you went provisioning NVMe namespace IDs (NSIDs) and logical unit numbers (LUNs) concurrently on the storage arrays. By using Gen 6 or already available Gen7 host bus adapters (HBAs) in your server, and a Gen 5 or newer fabric, you're ready to use NVMe over Fibre Channel.

Seeing How Much Performance Improves

If you read the section "Understanding How NVMe over Fibre Channel Improves Performance," you discovered the performance improvements and where they come from in an end-to-end NVMe over Fibre Channel solution. Now, this section shows you how big these improvements can be.

Many vendors with products in the market have demonstrated and documented improvements across latency, IOPS, and CPU utilization on the host side with NVMe over Fibre Channel compared to SCSI/FCP. Although the numbers differ between the systems and tests, application execution speed improved 30 to 50 percent, and IOPS increased 25 to 50 percent while consuming 30 percent less CPU resources for the same workload. These vendors showcased a typical Online Transaction Processing (OLTP) workload profile for transactional databases.

You may be thinking, "Wow, that's great! With a simple software upgrade on an NVMe over Fibre Channel capable array, we can achieve 30 to 50 percent faster application execution!" Therein lies the rub. If you're satisfied simply because the storage chain is now faster, you miss the greatest opportunity NVMe over Fibre Channel presents, which is that the NVMe benefits are dramatic in the server, the fabric, *and* the storage and not simply the storage chain itself.

THINKING ABOUT APPLICATIONS DESIGNED FOR NVMe

Imagine a program that can decompose a mathematical problem to parallel stream data requests — perhaps as many as 32 — that would return to be processed on a chip — possibly a NVIDIA chip — with multiple graphics processing units (GPUs) on it. Functionally, the GPUs would be large floating-point engines. If the application could then aggregate the output of those 32 streams, imagine what might be done to cycle times on financial analysis, threat analysis, or rendering. The possibilities are endless.

Chapter 2 discusses enhanced queuing with NVMe supporting 64k queues with a depth of 64k commands each, by designing modern applications that are already multithreaded to take advantage of the SSDs and NVMe protocol that support parallel queues by using multiple concurrent threads to perform IO. The application can achieve significant gains in IO performance. Additionally, in a highly virtualized environment, an individual VM could be assigned its own queue meaning it would no longer be in contention with other virtual machines.

Clearly some time will pass before applications across the board are redesigned to utilize FC-NVMe and SSDs in an optimal way. One application or layer where this change would be obvious — and this improvement is likely to happen soon — is in the hypervisor layer that virtualizes the server hardware in your datacenter. Increased parallel storage IO threads will likely deliver storage performance that is an order of magnitude better with virtual machines.

Considering SAN Design

How does running NVMe over Fibre Channel influence your SAN design? From a SAN design perspective, the areas to pay attention to are somewhat interrelated. In the same way the transition

to all-flash arrays (AFAs) in the datacenter has increased the IO density of the AFA storage ports, by delivering much more IOPS per port, the host-to-target port ratio has likewise increased. This increase continues with NVMe over Fibre Channel. The combination of a higher host port-to-storage target ratio and more IOPS per storage port heightens the risk of having oversubscribed hosts that exhibit slow-drain behavior and negatively impact other high-performing applications.

A *slow-drain device* is a host or storage array that doesn't return buffer credits in a timely manner to the switch. This causes frames to back up through the fabric and also causes fabric congestion. In a fabric, many flows share the Inter-Switch Links (ISLs), as well as virtual channels (VCs) on the ISLs. However, the credits used to send traffic or packets across the ISL or link are common to all the flows using the same VC on the link. Consequently, a slow-draining device may slow the return of credits and impair healthy flows through the same link.

To mitigate the impact of a device in a slow-drain state, the Brocade Slow Drain Device Quarantine (SDDQ) feature enables Monitoring Alerts Policy Suites (MAPS) to identify a slow-draining device automatically and quarantine it by moving its traffic to a lower priority VC in the fabric, thereby avoiding adverse impact on healthy flows. Having SDDQ enabled (part of Fabric Vision) is critical when implementing NVMe over Fabrics (NVMe-oF). In addition, the latest FC standards have added mechanisms for the fabric to send notifications about these types of congestion situations to the end devices, allowing them to take corrective action to mitigate the congestion by throttling or other means.

Another important step is to evaluate the switch port-to-ISL (fan in) ratios to validate that adequate bandwidth is available for peak spikes of traffic. With NVMe over Fibre Channel, the boundary can easily be pushed higher. As a result, you may need to add ISLs between switches in your existing SAN when increasing the footprint of NVMe over Fibre Channel storage arrays.

Understanding Why Monitoring Is Important

Monitoring is a cornerstone when managing a high-performance infrastructure such as a SAN. Adding NVMe over Fibre Channel to your SAN makes monitoring your SAN even more important because it enables you to identify issues before they affect application performance. Monitoring also helps you troubleshoot and pinpoint the root cause and path to resolution when an issue arises.


TIP

Having MAPS enabled is the baseline. But complementing MAPS with Brocade IO Insight capability on Gen 6 platforms adds built-in device I/O latency and performance instrumentation in Flow Vision. With the latest Brocade Gen 6 products, as well as with the entire Gen 7 portfolio, the IO Insight capabilities include NVMe over Fibre Channel protocol-level, non-intrusive, real-time monitoring and alerting of storage I/O health and performance. The additional visibility delivers deep insights into problems that may arise and helps maintain service levels.

Working with Zoning

Zoning applies the same way SCSI/FCP target access does for NVMe over Fibre Channel targets. Depending on the storage array, implementing NVMe over Fibre Channel can alter how you must zone for NVMe Controller and NSID access because some storage arrays implement the NVMe Controller target ports as logical interfaces (child World Wide Port Number [WWPN] behind the physical target port).


TIP

In principle, zoning works the same way as Node Port ID Virtualization (NPIV) logins in the fabric. As a result, you need to zone the hosts with provisioned NSIDs to the logical interface ports for the NVMe controller(s).

PEER ZONING

Peer zoning is a new type of zoning supported since the release of Fabric Operating System (FOS) 8.1. Peer zoning simplifies single initiator — single target zoning functionality without the need to define the zones individually. It allows one or more principal devices to communicate with the rest of the devices (nonprincipal devices) in the zone as if they were a single-initiator zone. Nonprincipal devices in the zone can communicate with the principal devices only, but they can't communicate with each other — and principal devices can't communicate with each other. This approach establishes zoning connections that are set up as single-initiator zoning with the operational simplicity of one-to-many zoning and reduced zone database usage.

A peer zone can have one or multiple principals. In general, storage ports are assigned as principals. Multiple principal members in a peer zone are used when all the nonprincipals (initiators) in the zone share the same target (storage) ports.

Peer zones aren't mutually exclusive with traditional zones; multiple zoning styles can coexist within the same zoning configuration and fabric. Peer zoning dramatically reduces the administrative burden of zoning in your fabric.

Knowing What ANA Is and Why It Matters

Multipath IO support with NVMe over Fibre Channel can be a confusing topic, but the key point is that symmetric multipathing is part of the NVMe specification and also applies to NVMe over Fibre Channel.

Take a step back and consider how multipathing is supported with SCSI/FCP. On enterprise class storage arrays, the predominant storage array controller architecture is designed so all the paths to a single LUN, regardless of which controller target port is used, are equally optimized, thus providing symmetric multipathing. The challenge is that many midrange storage arrays provide active/active access across two storage array controllers to a single LUN, but in fact only one of the two controllers owns the LUN at any given point. The result is that paths through one controller are considered optimized and preferred — through the

controller that owns the LUN — while paths through the other controller are non-optimized and not preferred.

Asymmetric Logical Unit Access (ALUA) was created to ensure that hosts use the optimized paths and only send IO on non-optimized paths when the optimized path isn't available or not functional for LUN access. ALUA is a SCSI standard that's implemented in the OS as well as on the storage array to ensure the OS always uses the optimized path unless it is unavailable or has failed.

For NVMe over Fibre Channel, an equivalent standard is provided with Asymmetric Namespace Access (ANA) as part of the NVMe specification 1.4. The ANA protocol defines how the storage array communicates path and subsystem errors back to the host so the host can manage paths and failover from one path to another.

REMEMBER

As you begin implementing NVMe over Fibre Channel, if you're working with a storage array that uses ANA, be sure to check that ANA is available on the OS version on the server side.

Knowing Which Applications Will Benefit

You likely have a handful of business applications in mind that you want to migrate to NVMe over Fibre Channel. After all, which application can't benefit from more performance? These are the usual suspects among enterprise applications, always hungry for more IOPS at lower latency, and if you can free up some CPU cycles on the server in the process, that's even better! In this category are transactional database systems such as Oracle and MS SQL Server as well as SAP HANA and NoSQL DBMS systems, and you may have an application in mind that's specific to your business.

Early adopters in this space tend to focus on applications that are heavy on the analytics side, such as machine learning (ML), artificial intelligence (AI), or simply the ability to perform real-time analytics on a transactional database system without affecting the primary purpose of the application. On the horizon, as enterprises and application developers garner experience with the vast queuing abilities with NVMe over Fibre Channel and develop new or re-architected applications designed to take full advantage of the potential of NVMe over Fibre Channel and Storage Class Memory storage systems, our guess is that the world will see application capabilities that today we can only dream about.

FACTORS PUSHING ML TO NVMe OVER FIBRE CHANNEL

Here's a quick look at the factors that are driving machine learning (ML) toward NVMe over Fibre Channel:

- NVMe over Fibre Channel offers the ability to decompose compute and storage to scale independently and enable storage capacity beyond each server's local capacity.
- More flexible cluster capabilities are available, with servers accessing the same/shared data set.
- ML applications need access to existing data placed on storage arrays in the SAN, such as data from transactional systems.
- ML has become business critical and requires data protection and/or high availability — even if "just" working on a copy of primary data sets.

Seeing That Not All Fabrics Are Created Equally

In any network design, you must evaluate the network needs over the projected lifespan of the infrastructure, which is often four to seven years for datacenter and storage networks. When doing so, follow best practices design guidelines regarding network redundancy, resiliency, and symmetric/homogenous topology without inherent bottlenecks. Likewise consider the cost of the network and ongoing operations — the total cost of ownership (TCO).

TIP

The best practice is to plan for up to 80 percent utilization of the network, which should leave room to accommodate high traffic spikes without performance degradation. The reality is how a network performs under load is a combination of the network design and the network technology used.

Enterprise Storage Group (ESG) published a paper comparing enterprise workload performance with AFAs using SCSI over FC and internet Small Computer Systems Interface (iSCSI) over Ethernet networks. One of the tests showed the impact of network utilization on performance as the utilization increased from

60 percent to 80, 90, and then 100 percent. The test revealed the impact of network congestion on FC and Ethernet with the host and the storage remaining the same.

Figure 5-2 shows the performance was normalized per the throughput performance when there was no other traffic on the network than the monitored enterprise workload. The results show that the enterprise workload is impacted when the network becomes congested, but the impact is drastically different on the FC network. There, the enterprise workload performance is degraded approximately 20 percent when the network congestion is between 80 percent and 100 percent. In comparison, with iSCSI on Ethernet, seen for Array B in Figure 5-2, the enterprise workload performance starts degrading at 60 percent congestion and then practically falls off the cliff with increased congestion and crashing at 100 percent.

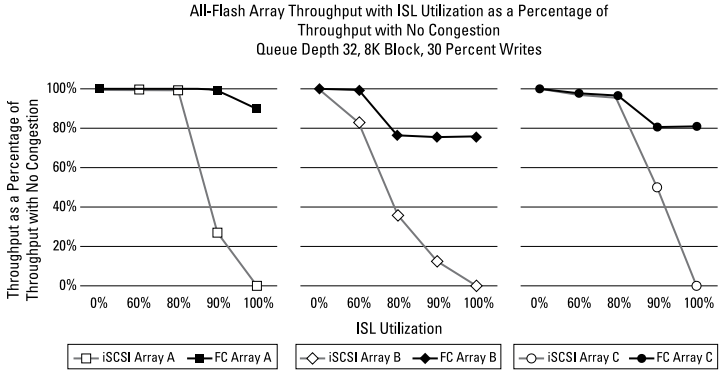


FIGURE 5-2: Comparing enterprise workload performance.

The tests in the ESG paper show it's a misperception that performance impact from congestion is solely a matter of bandwidth. Any network (unless it's uneconomically oversized) has periods of congestion. In cases where a network is sized well to sustain the workloads on the network, the periods of congestion should be momentary unless performance is degraded because of link failures or other component failures.

For networks transporting storage traffic, it's of utmost importance to know what the behavior or impact is when congestion occurs. Applications don't perform well when storage traffic performance is degraded, and they do even worse when traffic

is disrupted. The result can be application crashes. For high-intensive transactional systems, recovery after a database crash is typically time consuming. Meanwhile, the application is down, and business is at a halt.

REMEMBER

Degraded network situations occur multiple times during the lifetime of an IT infrastructure. These problems can result from failed cables, optics, switches, or human error. During these events, high-performing business critical applications must still be available and must perform as intended.

Maintaining Performance During Network Congestion

Why does iSCSI throughput drop rapidly and come to a halt as the network becomes saturated, while FC continues to provide significant throughput? Because the two fabrics have different flow control and congestion control mechanisms. With FC, the network is responsible for guaranteed delivery and doesn't allow packets to drop (a *lossless* network), but iSCSI depends on TCP/IP to ensure delivery because Ethernet provides only transport, not guaranteed delivery.

Consequently, when the network is congested, packets are dropped at the Ethernet level, and the TCP/IP protocol counteracts by retransmitting dropped packets and attempts to adapt to the lossy characteristics of the network. This includes TCP/IP packet acknowledgements from the receiver to the sender containing the receive window, which is equal to the amount of buffer space available on the receiver. This information tells the sender how much data can be in flight between the two ends of the communication. However, the receive window accounts only for the receiver's buffer space and not for any intermediary network nodes. Therefore, as the network becomes congested, an intermediary node may run out of buffer space and start dropping packets, which requires retransmission.

WARNING

Dropped packets and retransmissions can cause cascading congestion as retransmissions consume increasingly more of the available bandwidth, leaving less throughput for new data blocks and in some cases prevent storage exchanges from completing. In the worst case, as the test results demonstrate, iSCSI transmission

effectively stops, as the dropped packets and retransmissions consume all available bandwidth and timeouts are propagated to the SCSI layer.

In contrast, FC operates on a link-by-link, buffer-to-buffer accounting system. When devices (hosts, storage arrays, and switches) are connected, each end of each link communicates the amount of buffer space available. A sender is responsible for tracking how much of the link's receiver buffer space the sender is consuming and whether there are still buffers available to send. This is known as *buffer credits.* Each frame sent decreases the sender's buffer credits, and each frame acknowledgement increases the sender's credits. A sender can't send more data if the receiver buffer count is zero (if it runs out of buffer credits). Therefore, as the network becomes congested, an intermediary node may run out of buffer space, causing the upstream sender to stop sending, which proceeds in turn all the way back to the originator of the communication.

FC's end-to-end flow control protocol includes intermediary nodes, which use fair share algorithms to ensure each sender gets its fair share of the available throughput as buffer space becomes available. Consequently, as demonstrated by these tests, FC traffic continues to flow even as congestion approaches 100 percent.


REMEMBER

The bottom line is that during network congestions, there's no reason to believe that NVMe over TCP (Ethernet) will perform any better than the ESG paper demonstrated for iSCSI.

IN THIS CHAPTER

- Expanding the ecosystem
- Detecting and reporting intermittent physical issues
- Enabling error recovery at the sequence level
- Tagging flows with unique IDs

Chapter 6
Realizing the New Benefits for NVMe over Fibre Channel

Knowing what the changes are to functionality and specifications is part of the story. But just as important is understanding how those changes are represented in actual implementations and what benefits and caveats the end-user may need to be aware of. But beyond the specification changes is the question of how things get deployed. There's an ecosystem around Fibre Channel (FC) that includes servers, storage, networking, and software. How this ecosystem is being expanded by the functionality changes is critical. This chapter offers you a brief update on the new features and functions that have arrived for Non-Volatile Memory Express (NVMe) over Fibre Channel.

Ecosystem Expansion

Even as little as two years ago, the Ethernet community posed a serious question about whether NVMe over Fibre Channel would ever become a player in the market. Now, NVMe over Fibre

Channel has become the most pervasive interface for enterprise storage arrays from the storage Original Equipment Manufacturers (OEMs) exceeding the offerings of platforms that include the use of either RoCEv2, iWARP, or the nascent NVMe over TCP/IP market.

NVMe over Fibre Channel's success occurred due to five factors:

- **Ability to support concurrent NVMe and SCSI traffic**

  The ability to support concurrent NVMe and Small Computer System Interface (SCSI) traffic on the same server, host bus adapter (HBA), fabric, and storage is critical. Customers can't walk away from the existing investments in SCSI platforms due to financial implications and the stability of the environment. With NVMe over Fibre Channel, both types of traffic may be run on the same HBA across the same switched fabric to the same port on the storage array. This process provides the best possible investment protection to the customer both in the existing environment and for the cost of running a trial with the new technology. Over time, the infrastructure team can migrate those applications and workloads that will best see a return from the new technology without having to abandon the current running environments.

- **Ease of migration**

  An example of the ease of migration is evident in VMware's vSphere environments (ESXi v7.x) that first became available in April of 2020. This was VMware's first native operating system (OS) support release for NVMe over Fabrics, but what it allows as a process in terms of workload migration from SCSI to NVMe is brilliant. With any of the several storage array platforms that are currently supporting NVMe over Fibre Channel, it's simply a matter of configuring a namespace ID (NSID) on the array and presenting that NSID storage capacity to ESXi v7.x in the normal fashion. No special storage area network (SAN) behavior exists for this; it's normal SAN operations such as zoning to make certain that the storage element can be seen by the server. After that's done, the vSphere administrator chooses which of the workloads they want to migrate to the new technology and proceeds to perform a live Storage vMotion for the chosen platform to the new storage. Effectively, without the need of

an application service window (a major win for the SAN administrator), the application can be pointed at the new mirror of the datastore running on NVMe and see an 80 percent (or more) increase in performance based on VMware benchmarks. This accomplishment is achieved without the need to touch a single cable and without the need to negotiate an application outage window because no outage exists.

- **Broad OS support**

  Software support has actually been one of the major hinderances in the adoption of NVMe technology. However, the number of OSes that now support NVMe over Fibre Channel include SUSE Linux, RedHat Linux, VMware, IBM/AIX, and Windows. With the Linux implementations being put into the upstream development code, NVMe over Fibre Channel is available to all versions of Linux moving forward.

- **Application applicability**

  One of the primary benefits of application applicability has been low latency. Early use cases include taking a snapshot of a database that's running in a production SCSI environment and copying it into NVMe in order to run analytics on the data. As customers build more confidence in the ecosystem and the robustness of the implementations, more latency-sensitive applications will move into these environments. What workloads and how rapidly they get moved over will initially depend on the review of "best return" for early applications. But at some tipping point, net new applications of certain classes, Online Transaction Processing (OLTP), order entry, and so on will simply be placed into the new technology.

- **Simultaneous scale and performance**

  As opposed to alternative NVMe over Fabrics implementations such a NVMe over RoCEv2 where a single switch may reasonably provide excellent performance to a small cluster but large-scale deployments in a traditional spine-leaf configuration can be more problematic or TCP-based implementations can certainly scale but have frequent issues with high latency, NVMe over Fibre Channel has both performance and scale with extremely low latency — effectively providing the best of all environments.

Fabric Notifications

What would you think if the fabric and the attached devices could simply tell each other when they're detecting errors with the physical links and hardware? How comforting would it be to a SAN administrator to know that the devices are constantly monitoring for physical errors, and they're telling each other when those errors are impacting data flows? Enter Fabric Notifications!

Early in 2021, the FC community completed the definitions of a method of enhancing the ability of the SAN infrastructure to heal itself, such as identifying when a marginal link is impacting data flow (for example, the infamous "sick but not dead" path). This capability is embodied in a new element of the architecture known as *Fabric Notifications*.

This new architecture defines a system of messages that leverages the capabilities of the fabric and end devices to detect and report the occurrence of the intermittent physical issues. The fabric elements monitor the system for persistent issues and then generate messages sent to the impacted devices. The devices then make a more informed decision about what do about the error and automatically take the most effective action.


TIP

The Fabric Notifications architecture, called the Fabric Performance Impact Notification (FPIN), inserts key information into the system and distributes it to the participating devices and their peers (for example, zoned in devices). It provides the ability for the devices to manage their level and degree of participation (such as only receive and react to events they care about). The simplicity and control of the architecture allows participating devices to control the level and degree of participation, which supports evolution of the solutions over time. Specific device solutions can adjust their participation based on the vendor's experience in the enterprise.


TIP

Pairing the high-performance and low-latency characteristics of NVMe solutions with the advanced characteristics of Fabric Notifications for FC SANs enables the optimal confluence of performance and simplicity. The NVMe stack is provided visibility into the operation of the transport and the attached devices, which provides visibility and intelligence that can be used to optimize the overall solution. Because the intelligence afforded by Fabric Notifications

is exchanged throughout the fabric, it enables the self-learning, self-optimizing, and self-healing activities of an autonomous SAN and provides NVMe with an optimal infrastructure — one that reduces the time that NVMe applications may perform at sub-par levels resulting in lost revenue or degraded service.

Sequence Level Error Recovery

Way back when, when FC Protocol (FCP) was developed, everything was in order end-to-end, and people took a big hammer approach to Ethernet recovery and any type of FC error, so error detection recovery was the method that worked. Then came exchange-based routing, where exchanges may have been delivered out of order. And this change became the root of the problem.

The big hammer approach doesn't work well for customers or for FCP-based error detect recovery. A new functionality was desired, and that's where Sequence Level Error Recovery (SLER) was born.

But what exactly is SLER for FC-NVMe and why does it matter? The goal is to enable error recovery at a sequence level without having to pass the error up to the storage protocol level (NVMe). To recover from errors, the NVMe initiator/target adapters agree on retransmission of lost or corrupted commands. The benefit of enabling the transport layer to recover lost or corrupted commands is that error recovery happens much faster as a result, with little or no impact on storage performance. As the technology for FC-NVMe moves toward end-to-end storage latencies in the tens of microseconds with storage class memory, the result is better error recovery than with any other fabric technology.

VMID Tagging

An element of the FC standard that's been around for more than six years is virtual machine ID (VMID) tagging, which is the ability to tag the input/output (IO) of a virtual machine at a FC frame level granularity. The constraint has always been that in virtualized environments, the host OS handles all the IO. The guest's OS hasn't been visible to the fabric. The challenge with this scenario is the "bully" versus the "victim" issue. For example, of the

30 VMs hosted on one ESXi image, do you know which ones are consuming the majority of the IO and which ones are suffering as a result? In some instances, it may be that the application that's consuming the most in fact needs to for the success of the application environment. However, without real performance data, it's impossible to be sure if that's the case.

One of the most common efforts at mitigation of such an imbalance is to vMotion (move the virtual machine location) the application that first complains about performance in the hope that one is either moving the victim away from the bully or vice versa. As of March 2021, the first of the NVMe capable arrays supports the ability to "tag" the flow with a unique ID traceable back to the VM, which effectively gives the ability to measure/monitor an individual flow to the VM itself. This capability is unique to FC at this time, but it represents an amazing window into the operational characteristics of the applications. The implications of this are incredibly far reaching. Visibility to the application itself — instead of just to the group of applications resident on a particular host — has been somewhat of a "holy grail."

The only hypervisor to date that has implemented VMID tagging is VMware. However, additional hypervisors are expected to be supporting this function in 2022.

IT infrastructure is in place to service the application base. But how do you best do this without visibility to that application flow? This is the benefit that VMID Tagging brings. Expectations are that over time VMID tagging will become integrated into a number of management tools. This integration allows IT infrastructures to know which VMs may be safely stacked in a given environment without implications. Also, the performance and growth of those applications need to be tracked and proactively dealt with instead of waiting for the underlying infrastructure to fail in its ability to service the application. Long-term profiling of application performance will be a particular benefit to applications that look to do load balancing in the infrastructure.

IN THIS CHAPTER

- Seeing the impact of NVMe's IOPS
- Scaling NVMe adoption
- Understanding how FC and Ethernet/IP are optimized
- Having a dedicated storage fabric
- Recognizing the advantages of dual FCP/FC-NVMe fabrics

Chapter 7
Ten NVMe over Fibre Channel Takeaways

You're ready to get serious about Non-Volatile Memory Express (NVMe) over Fibre Channel. In this chapter, we cover ten key points to remember about NVMe over Fibre Channel. Keep in mind the following:

- **Storage and memory have different needs.** You probably want to support both. Enterprise storage is all about protecting high-value data assets while providing fast access. Working memory, on the other hand, is coupled to computation, where error elimination is secondary to minimizing latency.
- **Enterprise storage eases handling memory errors.** Mainstream server architectures correct single-bit memory errors while regarding double-bit error correction as not worth the cost. Instead, rely on known-good images kept on enterprise storage and rerun an application when uncorrectable errors occur.
- **Reduce risk with a transitional adoption strategy.** Installing an entirely new non-Fibre Channel (FC) fabric infrastructure to adopt NVMe is an all-or-nothing approach. It presents risks for your long-term, high-value data assets as

well as your budget. A better tactic is to extend your existing infrastructure, providing an as-needed, gradual transition that protects your data and investments while leveraging existing IT skills.

- **NVMe input/output operations per second (IOPS) has as much impact as latency.** Most of the buzz over NVMe has been focused on its amazingly low latency because this measure is easy to benchmark, and the early focus of NVMe is on memory use cases. But as architectures move toward storage and massively parallel processing, IOPS will matter more, increasing the need for robust datacenter fabrics, analytics features, and the excellent vendor support for which FC is known.
- **Scale NVMe adoption with a tried-and-true fabric.** As NVMe adoption grows beyond direct-attached memory needs to datacenter-scale storage, IT architects will need a robust, credit-based lossless fabric to deliver predictable speed and reliability. Experimenting with immature fabrics doesn't help the transition to NVMe.
- **The NVMe future will come at its own pace.** A quick glance at how long the backup media transition has been (see Chapter 3) reminds you that the vast installed Small Computer System Interface (SCSI) infrastructure will continue to deliver value for years to come. That's why a sound SCSI-to-NVMe transition strategy must provide flexibility, whether the move lasts two years or two decades.
- **Include vendor support in your adoption plan.** Reliable IT infrastructure takes far more than sexy technology; products from multiple suppliers must also work well together. Enterprise vendor support for SCSI-based storage works because of interoperability testing, so be sure to ask your vendors about their NVMe fabric interop test plans.
- **FC and Ethernet/IP are optimized differently.** FC was born and raised during an era of explosive Ethernet/IP growth. FC succeeded because it was optimized for one premium use case: reliable, high-speed delivery of bursty storage traffic. Conversely, Ethernet and IP won as commoditized technologies that work anywhere, but, like SCSI, they come with a lot of baggage.

TIP

- **Dedicated storage fabrics are recommended.** Storage experts know that reference architectures from mission-critical storage vendors call for dedicated storage fabrics.

  An FC fabric that can support both SCSI and NVMe concurrently is the low-risk choice compared to an Ethernet/IP fabric, which requires gambling on either Remote Direct Memory Access (RDMA)-capable Network Interface Cards (NICs) or Transmission Control Protocol (TCP) offload engine technology (TOE).

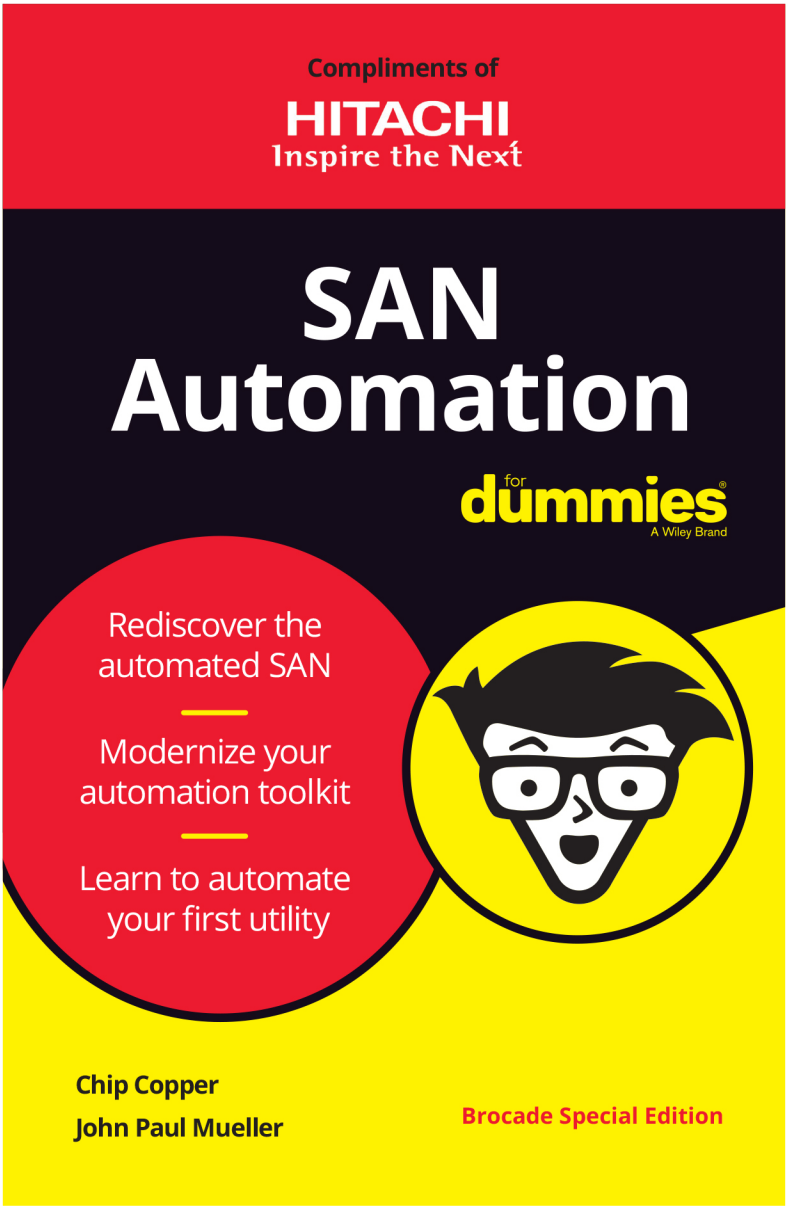
- **Dual FC Protocol (FCP)/FC-NVMe fabrics offer big advantages.** Dual-protocol FCP/FC-NVMe fabrics offer ultra-low latency for working memory needs, as well as the as-needed, low-risk SCSI-to-NVMe migration you need for high-value storage assets. Such fabrics also simplify your storage purchasing decisions during the SCSI-to-NVMe transition — a process that's likely to take years.

  You can continue to use the same technologies, best practices, and tools to provision NVMe-based storage, dramatically reducing the learning curve and associated risk.

REMEMBER

Visit www.broadcom.com

to download *SAN Automation For Dummies*



Move into the NVMe over Fibre Channel future

Non-Volatile Memory Express (NVMe) is a massively parallel, ultra-low latency memory protocol. NVMe over Fibre Channel extends this protocol to the scale of enterprise storage. Fibre Channel, the premium datacenter fabric, can transport both NVMe and Small Computer System Interface (SCSI) concurrently, giving you a low-risk NVMe adoption path that protects your high-value data assets. This book gets you started with NVMe over Fibre Channel, helps you create an adoption strategy, and shows you the way forward.

Inside...

- Adopt NVMe at your pace with low risk
- Deliver speed and reliability
- Increase IOPS with enhanced queuing
- Protect high-value storage assets
- Leverage concurrent FCP and NVMe
- Simplify storage purchasing decisions
- Analyze and optimize with Fabric Vision

HITACHI Inspire the Next BROCADE A Broadcom Company

AJ Casamento a 40+ year IT veteran in consulting, engineering, and product development, helps customers with technology. **Marcus Thordal,** 20+ years in storage architecture and product management, provides expertise to customers on infrastructure solution designs. Both AJ and Marcus work for Brocade.

ISBN 978-1-394-15984-0
90000
9 781394 159840

for dummies® A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.