

Compliments of



**BROCADE**<sup>®</sup>  
A Broadcom Company

# NVMe over Fibre Channel

for  
**dummies**<sup>®</sup>  
A Wiley Brand

Boost performance  
with super-low latency

Maintain mission-critical  
storage SLAs

Reduce risk with  
concurrent SCSI/NVMe



**Brian Sherman**  
**Marcus Thordal**  
**Kip Hanson**

**IBM/Brocade**  
**2nd Special Edition**

## About IBM

IBM is a global technology and innovation company headquartered in Armonk, NY. IBM is much more than a “hardware, software, services” company. IBM is now emerging as a cognitive solutions and cloud platform company. More than 25,000 companies choose IBM and Brocade solutions and services for their combined innovations and expertise. IBM helps organizations like yours improve security, collaboration, productivity, and operations. Together they can help you achieve your business goals.

For more information on IBM flash storage solutions: **[ibm.biz/flashstorage](http://ibm.biz/flashstorage)**.

For more information on the IBM b-type SAN Storage solutions: **[ibm.biz/san-btype](http://ibm.biz/san-btype)**.

## About Brocade

Brocade, a Broadcom Company, is the proven leader in Fibre Channel storage networks that serve as the foundation for virtualized, all-flash data centers. Brocade has partnered with IBM since 1998 to provide Fibre Channel storage solutions that deliver innovative, high-performance networks that are highly resilient and easier to deploy, manage, and scale for the most demanding environments. The network matters for storage and Brocade Fibre Channel storage networking solutions are the most trusted, widely deployed network infrastructure for enterprise storage.

**[www.broadcom.com](http://www.broadcom.com)**



# NVMe over Fibre Channel

IBM/Brocade 2nd Special Edition

**by Brian Sherman,  
Marcus Thordal, and  
Kip Hanson**

**for  
dummies<sup>®</sup>**  
A Wiley Brand

# NVMe Over Fibre Channel For Dummies®, IBM/Brocade 2nd Special Edition

Published by  
**John Wiley & Sons, Inc.**  
111 River St.  
Hoboken, NJ 07030-5774  
www.wiley.com

Copyright © 2019 by John Wiley & Sons, Inc., Hoboken, New Jersey

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Trademarks:** Wiley, For Dummies, the Dummies Man logo, Dummies.com, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries, and may not be used without written permission. IBM and the IBM logo are trademarks or registered trademarks of IBM Corporation. Brocade and the Brocade logo are trademarks or registered trademarks of Brocade Inc. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc., is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact [info@dummies.biz](mailto:info@dummies.biz), or visit [www.wiley.com/go/custompub](http://www.wiley.com/go/custompub). For information about licensing the *For Dummies* brand for products or services, contact [BrandedRights&Licenses@Wiley.com](mailto:BrandedRights&Licenses@Wiley.com).

ISBN 978-1-119-60267-5 (pbk); ISBN 978-1-119-60270-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

## Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. Some of the people who helped bring this book to market include the following:

### First Edition Co-author:

Curt Beckmann

**Project Editor:** Martin V. Minner

**Editorial Manager:** Rev Mengle

**Executive Editor:** Katie Mohr

**Business Development**

**Representative:** Karen Hattan

**Production Editor:** Siddique Shaik

**Brocade Contributors:**

Marc Angelinovich,

AJ Casamento,

Howard Johnson,

David Peterson,

David Schmeichel

# Table of Contents

INTRODUCTION .....	1
About This Book .....	1
Foolish Assumptions.....	2
Icons Used in This Book.....	2
<b>CHAPTER 1: Exploring NVMe over Fibre Channel .....</b>	<b>3</b>
Picking Sides: Is It Storage, Network, or Memory? .....	6
Mapping the dichotomy.....	6
On errors.....	8
Accelerating Access to Flash .....	9
Understanding How NVMe Relates to SCSI.....	9
Anticipating Future Benefits of NVMe over Fibre Channel .....	11
Invigorating Your Fabric .....	12
<b>CHAPTER 2: Delivering Speed and Reliability with NVMe over Fibre Channel.....</b>	<b>13</b>
Reviewing Fibre Channel's Place in the Storage Ecosystem.....	14
Evaluating Performance Metrics in Storage Context .....	14
Storage metrics .....	16
Souping up device metrics.....	19
Achieving High Performance .....	20
Making Use of Enhanced Queuing.....	21
Realizing Reliability.....	22
Redundant networks and multipath IO .....	22
Features of a lossless network.....	23
Security.....	24
Good tools matter.....	24
<b>CHAPTER 3: Adopting and Deploying FC-NVMe.....</b>	<b>25</b>
Identifying Your Situation.....	25
Considering Your Adoption Strategy .....	26
Protecting high-value assets.....	26
Allowing for a marathon shift.....	27
Exploiting Dual-Protocol FCP and FC-NVMe.....	28
Zoning and name services.....	31
Discovery and NVMe over Fibre Channel .....	31

	Familiarizing Yourself with NVMe over Fibre Channel.....	32
	Experimenting in your lab.....	32
	Migrating your LUN to a namespace.....	33
	Transitioning to production.....	34
<b>CHAPTER 4:</b>	<b>Comparing Alternatives to NVMe over Fibre Channel</b> .....	<b>35</b>
	The Long and Short of RDMA .....	35
	InfiniBand.....	37
	iWARP .....	37
	Yo, Rocky .....	39
	Evaluating Ethernet-Based NVMe .....	40
	Commodity or premium?.....	41
	Smart shopping.....	42
<b>CHAPTER 5:</b>	<b>Improving Performance with NVMe over Fibre Channel</b> .....	<b>43</b>
	Understanding How FC-NVMe Improves Performance.....	44
	What about the fabric?.....	44
	The host side .....	45
	The storage front end.....	46
	Storage array architecture .....	46
	The storage array back end .....	46
	Handling NVMe support with a software upgrade .....	47
	Improving Performance .....	47
	Considering SAN Design with FC-NVMe in Mind .....	48
	Understanding Why Monitoring Is Important .....	49
	Working with Zoning.....	50
	Knowing What ANA Is and Why It Matters .....	51
	Knowing Which Applications Will Benefit.....	52
	Seeing That Not All Fabrics Are Created Equally .....	53
	Maintaining Performance during Network Congestion .....	55
<b>CHAPTER 6:</b>	<b>Ten NVMe over Fibre Channel Takeaways</b> .....	<b>57</b>

# 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 3 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 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

This book, *NVMe over Fibre Channel for Dummies*, IBM/Brocade 2nd Special Edition, focuses on a relatively small but very important aspect of information technology. NVMe (Non-Volatile Memory Express) 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, FC-NVMe (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. And for the uninitiated, this book introduces you to a technology that may very well be 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, 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 Fibre Channel upper-layer protocols, enabling a low-risk transition from SCSI (short for Small Computer System Interface) to NVMe without the need to invest in experimental infrastructure.

# Foolish Assumptions

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 Facebook account. If not, this book offers plenty of reminders and sidebars to guide you through the more difficult parts, and to help decipher the endless, confusing acronyms lurking around every corner of the IT world.

## Icons Used in This Book

A number of helpful icons are scattered throughout *NVMe over Fibre Channel For Dummies*. These will 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 will make your job lots easier, and prevent having to order late-night pizza delivery because you're stuck in the server room reconfiguring a storage array.



REMEMBER

If you're one of those who spends his or her days reading thick hardware manuals, you're bound to forget things along the way. If so, the Remember icon might just be your new best friend.



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.



- » Classic storage versus classic memory
- » Accelerating access to flash
- » How NVMe relates to SCSI
- » The future benefits of FC-NVMe
- » Invigorating your fabric

# Chapter 1

## Exploring NVMe over Fibre Channel

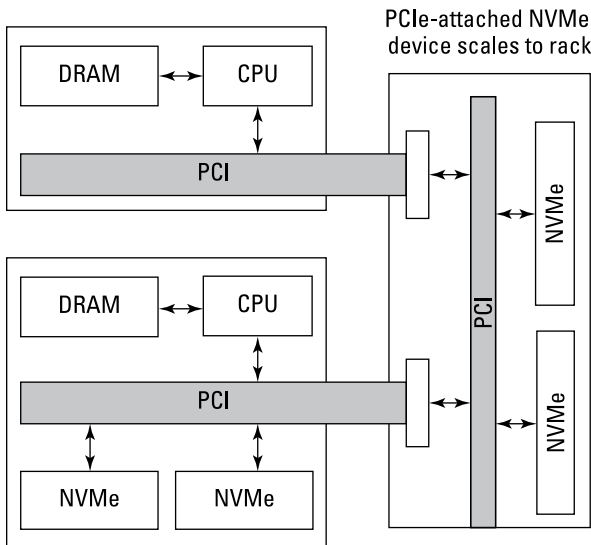
This chapter offers a brief tutorial on NVMe (Non-Volatile Memory Express) over Fibre Channel. We introduce (or reintroduce, for you veterans) a bunch of confusing acronyms, discuss the merits of solid-state storage and see how it compares to memory, and break down the component parts of this exciting,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

NVMe over Fibre Channel is a full-featured, high-performance technology for NVMe-based fabric-attached enterprise storage, but is a no-compromise solution for NVMe working memory use cases as well. (We talk in this chapter about how those use cases differ.) NVMe over Fibre Channel is relatively new, even though its component parts are not. Fibre Channel (FC) has been the leading enterprise storage networking technology since the mid-1990s. Speeds of 16Gbps (called Gen 5) are widespread. Gen 6 Fibre Channel became available in 2016, delivering twice the speed of Gen 5 and a staggering 8x bandwidth on 128GFC links, and it is selling like hotcakes. Gen 7, which will double speeds yet again, is already on the horizon with Gen 7 HBAs entering the market. FC is primarily used to carry the Small Computer System Interface (SCSI) protocol, the historic leader for direct attached PC

or server storage. SCSI on Fibre Channel is called, boldly enough, Fibre Channel Protocol (FCP).

NVMe refers to several related things:

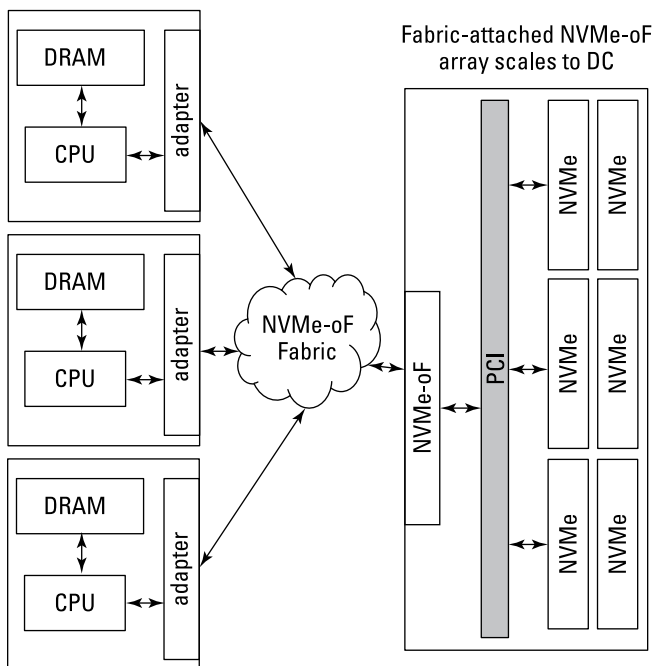
- » An open collection of standards for accessing and managing nonvolatile memory (NVM), especially high-performance solid-state memories such as flash or Storage Class Memory (SCM), such as 3D XPoint
- » That collection's primary specification, which provides a common, high-performance interface for accessing NVM directly over PCI Express (see Figure 1-1) (<https://nvmexpress.org/resources/specifications/>)
- » A nonprofit corporation NVM Express (<https://nvmexpress.org/>) that works to develop and promote the standard, and is supported by a wide range of technology companies



**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 shared access, greater capacity, enhanced data protection, and flexible multi-vendor support. 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 (aka NVMe-oF), which defines how NVMe commands can be transported across different fabrics in a consistent, fabric-independent way (see Figure 1-2). That makes life a lot easier for software developers!



**FIGURE 1-2:** Using NVMe over Fabrics to increase the scale of NVMe.

The NVMe-oF 1.0 specification, released in 2016, described two fabric categories, Fibre Channel fabrics and Remote Direct Memory Access (RDMA) fabrics.

**Fibre Channel fabrics:** NVMe Express chose the T11 standards body, which handles all Fibre Channel standards, to define the new “FC-NVMe” protocol. In mapping NVMe onto Fibre Channel, 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.

**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 NVMe Express to target three existing fabric transports in one effort.



TIP

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

You can find the latest specification at <https://nvmexpress.org/resources/specifications/>.



REMEMBER

Fibre Channel is capable of transporting multiple higher-level protocols concurrently, such as FCP and FC-NVMe (the label for the specific frame type of NVMe-over-Fibre Channel traffic), as well as FICON, a mainframe storage protocol. That bears repeating: FC-NVMe can coexist on your FC 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, Network, or 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. Three other NVMe fabrics (InfiniBand, RoCE, and iWARP) are memory-oriented (they support remote direct memory access or RDMA) while NVMe/TCP is traditional network access. Indeed, recent conferences covering flash and other persistent memory technologies have gushed over the arrival of the “storage/memory convergence.”

### Mapping the dichotomy

Wait a minute . . . memory and storage converging? *What?* 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 are 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), 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
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 is something of a trend toward convergence, rather

than 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 CRC error detection built into drives. Servers (enterprise-grade compute) have 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 service-level agreement (SLA), which is the contract between storage customers and their storage provider. This terminology is often used even within organizations between consumers of storage and their IT departments.



WARNING

Solving the problems of working memory would not 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.

## OTHER DATA ABOUT DATA

For decades, magnetic recording technology 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 default storage in laptops for the past ten years and since 2015 has been transitioning to NVMe attached SSDs.

# 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.

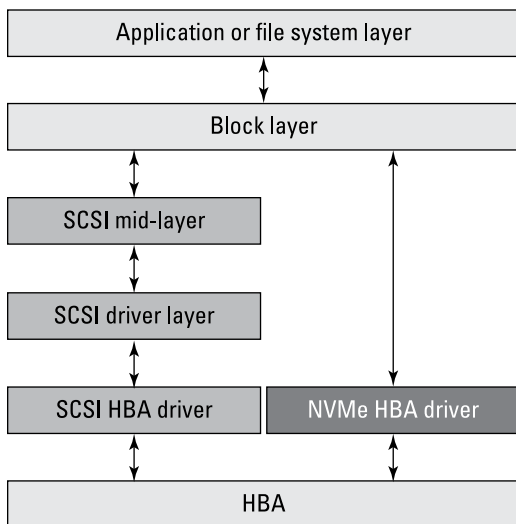
Over time, the density of flash increased dramatically, and effective software algorithms were developed to mask its write challenges. The combination of speed, density, and tolerable write endurance have brought flash to the point where it is the technology of choice for production data and displaces 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 is 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 (see Figure 1-3).



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

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

- » **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.
- » **LUNs and namespace IDs:** LUN stands for logical unit numbers, 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.
- » **Fibre Channel Protocol (FCP):** *Fibre Channel Protocol* is, like *namespace*, an odd moniker for what is really "SCSI over Fibre Channel." FCP is not about basic Fibre Channel; it is about the way SCSI features were implemented on top of Fibre Channel.
- » **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-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 will see FCP in the mix.
- » **Other protocols:** The name *FCP for SCSI-on-FC* might be partly responsible for a perception that Fibre Channel is limited to SCSI, but it's important to remember that SCSI is not 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 Fibre Channel (FC) transport. Some NVMe advocates have pointed to the implementation of SCSI over FC and incorrectly blamed its relatively poky performance on Fibre Channel. 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 Fibre Channel. (We say more about the inherent benefits of Fibre Channel in Chapter 2.) Leveraging Fibre Channel as a transport gives users easy access to all the speed and parallelism of NVMe over

Fabrics with none of the disruption entailed in building parallel infrastructure.

Here are some additional considerations as NVMe over Fibre Channel increases its lead on competing technologies:

- » As flash becomes more storage-oriented, the dominant storage protocol (SCSI) has hampered one of the key advantages of flash, its speed. That will change as more storage vendors embrace NVMe over Fibre Channel.
- » New semiconductor memories such as Intel/Micron's 3D XPoint are just coming onto the scene. They hold out the promise of much faster writes as well as 100x or even 1,000x 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 Fibre Channel 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

- » Reviewing Fibre Channel's place in the storage ecosystem
- » Evaluating performance metrics in the storage context
- » Recognizing the advantage of improved queuing
- » Achieving high performance
- » Realizing reliability

# Chapter 2

# Delivering Speed and Reliability with NVMe over Fibre Channel

**E**thernet 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, Fibre Channel networking managed to grow to the point where nearly all enterprises rely on it for their mission-critical storage needs. FC's success was not 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 stay with or adopt this robust networking technology.

# Reviewing Fibre Channel's Place in the Storage Ecosystem

Ethernet remains the primary transport for communication between servers. There are, however, distinct advantages to using Fibre Channel 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 onramps, but then steers them into the ditch (drops the frames) when the freeway runs out of space. For those cars that don't reach their destination, 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.

Fibre Channel, on the other hand, was designed for well-ordered, reliable transport of data regardless of load. It has a feature similar to the entrance ramp lights most of us grumble about on our daily commute. No frames are lost, and each is delivered in the proper order. The Fibre Channel freeway doesn't have traffic jams or fender benders. This makes Fibre Channel the ideal solution for mission-critical storage requirements.

# 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 the wrong fertilizer is used, you might 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 might end up with a lot more time on your hands. Maybe you can take up gardening?

This section focuses on storage metrics; they are what the storage consumer ultimately cares most about, and they better enable apples-to-apples comparisons between different 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)

## 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 is 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 I/O sizes.

Fabric throughput is usually viewed as how much data can be pushed through the fabric when all ports are running maximum speed. A 64-port, 10G device usually has a throughput of 640 Gbps (double that if input and output are separately counted). But some low-end devices have internal “oversubscription” and cannot 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 does not 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.

The relative importance of these performance indicators depends a great deal on the user application. Systems focused on minimizing response time value latency above the other metrics. Streaming high definition video requires huge amounts of data, so good throughput will be 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.”

Of course, there’s more to overall performance than the speed of the network. If your hard drives are dogs, why bother with Fibre Channel? 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).

## 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 SAS, or serial attached 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, 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 or gigabytes 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 4KB. However, many applications that demand high IOPS use larger I/O sizes, such as 64KB. You need to 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 4KB read operation takes 1 ms, you might expect to be able to perform 1,000 of those I/Os in a second, and indeed that might 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.



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:

- » **Example 1:** Some devices can have excellent throughput or IOPS numbers for sequential accesses, 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 is not 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. IBM and Brocade, a Broadcom Inc. Company, have long recognized that the expectations of Fibre Channel 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, “Fabric Vision Features.”)

## FABRIC VISION FEATURES

The IBM b-type SAN portfolio with Fabric Vision technology provides numerous tools for analyzing and optimizing FC fabric performance and reliability. Here is a summary of its more important features.

**IO Insight:** On supported products, IO Insight proactively and non-intrusively monitors storage device IO latency and behavior through integrated network sensors, providing deep insight into problems and ensuring service levels.

**VM Insight:** Seamlessly monitors virtual machine (VM) performance throughout a storage fabric with standards-based, end-to-end VM tagging. Administrators can quickly determine the source of VM/application performance anomalies, as well as provision and fine-tune the infrastructure based on VM/application requirements to meet service-level objectives.

**Monitoring and Alerting Policy Suite (MAPS):** Leverages prebuilt, rule/policy-based templates within MAPS to simplify fabric-wide threshold configuration, monitoring, and alerting. Administrators can configure the entire fabric (or multiple fabrics) at one time using common rules and policies, or customize policies for specific ports or switch elements. With Flow Vision and VM Insight, administrators set thresholds for VM flow metrics in MAPS policies in order to be notified of VM performance degradation.

**Flow Vision:** A set of flow-oriented tools that enables administrators to identify, monitor, and analyze specific application and data flows in order to simplify troubleshooting, maximize performance, avoid congestion, and optimize resources. Two of the Flow Vision tools are listed next.



**Flow Monitor:** Provides comprehensive visibility into flows within the fabric, including the ability to automatically learn flows and non-disruptively monitor flow performance.

**Flow Mirroring:** Provides the ability to non-disruptively create copies of specific application and data flows or frame types that can be captured for in-depth analysis.

## Souping up device metrics

*Read caching* can help hard disk drives (HDDs) when accessing data both randomly and sequentially. That's because data in cache can be accessed quickly for random reads and the disk drive can read an entire disk track and cache the extra data blocks for sequential reads. Read caches don't help SSD performance as much but are still important because cache provides lower response times than SSDs.

*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 might expect from its latency metric. For example, you might expect a device with a 1 ms latency metric for 256KB reads to have

a 1,000 IOPS metric. However, 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 now “multi-threaded” to generate the overlapping I/Os and take advantage of pipelined device performance. In addition, virtualized servers running lots of apps in parallel also issue many overlapping I/Os. (An interesting side effect is that minor latency changes may not change overall IOPS, depending on where the system bottlenecks are. More on this later in the chapter.)

As you’ll see, 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 FC-NVMe, just as it is with Fibre Channel itself. Fibre Channel 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. Fibre Channel 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, Fibre Channel fabrics are typically implemented as

parallel, redundant, active-active fabrics, which offer both reliability and added performance. Similarly, Fibre Channel'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. FC-NVMe leverages all the traditional benefits of Fibre Channel, while providing additional performance benefits inherent in NVMe. The streamlined protocol stack is one benefit that we've mentioned already. The other major improvement is enhanced queuing.

## Making Use of Enhanced Queuing

For decades, storage vendors have competed for leadership on each of three of the metrics described in this chapter, and have long made use of techniques such as caching, parallelism, and pipelining to boost their performance metrics.

We've already mentioned how flash drives have a big advantage in latency over disk drives because they don't have to twiddle their thumbs while the media spins or the read/write head lumbers out to the appropriate track. 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 LUNs (logical unit numbers, 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 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 100x improvement with NVMe, but it isn't hard to imagine a significant boost of 2x or more. You can also expect to see significant CPU utilization reduction because the CPU is no longer waiting for queued IO to complete. The advanced analytics features available with IBM b-type Fibre Channel 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 preceding examples might 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 rather than spend 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 crash on double-bit errors. Servers, like laptops, are also vulnerable to viruses and power failures, so companies don't spend more for killer 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 IO

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 9's or better" reliability, keeping your customers happy.



TIP

Cutting corners is penny wise but pound foolish if you disappoint your customers and they toddle down the road to a competitor.

Enterprise storage vendors understand all that and have invested significant amounts of 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 technology is more mature than a *Saturn V*, and fortunately, the glitches are rarer.)

The vendors have also worked hard to provide multipath I/Os so that 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 Fibre Channel — for many good reasons.

## Features of a lossless network

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

Fibre Channel has always been a lossless networking technology. It treats a payload like a loved one. Every Fibre Channel 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 does not send a frame that the receiver can't handle. By contrast, Ethernet has been 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 (rather than buffer credits) to

avoid packet loss, but DCB still has some notable interoperability challenges.

## Security

Obviously, if you are 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. Fibre Channel offers extra security in a number of ways. Fibre Channel is already trusted to keep the world's most important data secure, and Fibre Channel brings this security model to NVMe over Fabrics.



REMEMBER

One key advantage of Fibre Channel comes from its specialized nature. Datacenter-centric Fibre Channel is not the protocol of the Internet, so hackers cannot shove Fibre Channel frames across the Internet to sneak past your firewall and into your datacenter.

Fibre Channel 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.

## Good tools matter

To err is human, and wise IT administrators know that it's a good career move to use robust tools to reduce the opportunities for, shall we say, demonstrating their humanity. For more info, check out the sidebar on SAN automation and storage integration.

## SAN AUTOMATION AND STORAGE INTEGRATION

IBM b-type SAN automation makes SAN configuration and management simple and error-free through intelligent automation. Many enterprise arrays provide automatic integration of storage provisioning and SAN zoning, delivering end-to-end configuration through a single point of management. These features also apply to FC-NVMe.

- » Identifying your situation
- » Considering your adoption strategy
- » Exploiting dual-protocol FCP and FC-NVMe
- » Familiarizing yourself with NVMe over Fibre Channel

# Chapter 3

## Adopting and Deploying FC-NVMe

**Y**ou'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 NVMe-over-Fibre Channel adoption in your organization. Only one big question remains: Where do you start?

### 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. And you don't have to risk education funds on uncertain IP/Ethernet NVMe protocols, as you might with competing technologies. That being said, you should check a few legacy infrastructure boxes before going further:

- » You can deploy NVMe over Fibre Channel on an existing IBM b-type or Brocade Fibre Channel infrastructure, provided it is relatively up to date (FOS 8.1.0 or later). Check with your hardware supplier on interoperability.

- » The Fibre Channel fabric must be Gen 5 (16 Gbps), or better yet, Gen 6 (32 Gbps). Of course, Gen 7 (64 Gbps) also supports FC-NVMe.
- » Servers using NVMe over Fibre Channel need Gen 6 (32 Gbps) host bus adapters (HBAs); Gen 6 HBAs also work with Gen 5 fabrics.
- » You need a storage device that supports FC-NVMe frame types. This can be an FC-NVMe-enabled array, or, for early familiarization, a server with an FC-NVMe 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 (8 Gbps), it's definitely time for an upgrade, regardless of any NVMe-over-Fibre Channel undertaking. This presents an excellent opportunity to do so.

## 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 will be gradual, lasting several years. Unfortunately, this slow transition would make it painful to have a classic Fibre Channel (FC) network sitting side by side with your brand new Ethernet-based NVMe network. When you buy more storage, what kind do you buy? As you build new apps, which environment do you connect them to? And if you go with Ethernet, which kind: iWARP, RoCEv2, or NVMe over TCP? None of these are the obvious winner, and none have much of a historical adoption record for storage use. However, adopting a dual-protocol Fibre Channel fabric (running concurrent FCP and FC-NVMe traffic) eliminates or simplifies those questions. Fibre Channel is widely adopted and enjoys excellent support by storage vendors and other technology providers, making NVMe over Fibre Channel a relatively low-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, machine learning, and deep learning 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 will be maintaining 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. Migration from legacy architecture, which is typically built on a 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**

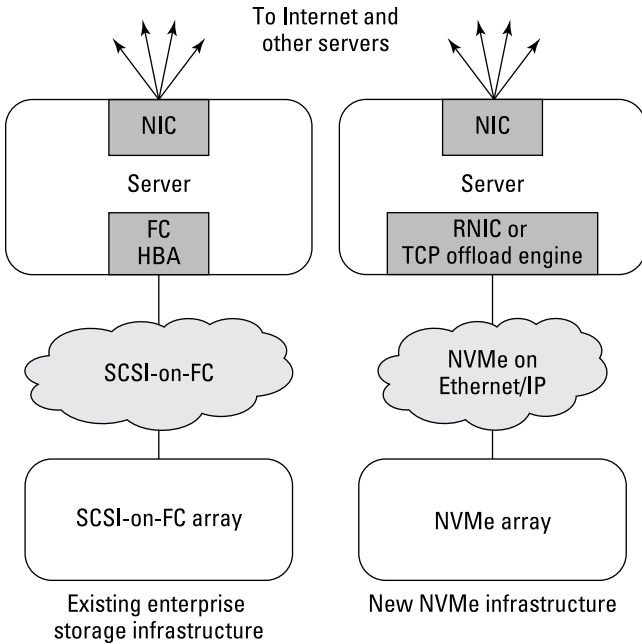
With all the buzz about the speed of NVMe and the rapid rise of all-flash arrays, won't the whole world go to NVMe storage by the end of next week? Probably not. Consider all the claimed advantages of IPv6 and the length of time required for significant adoption of that technology. The transition is still moving very slowly, many years after most infrastructures were IPv6 ready, with full support available on both hardware and software. 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 are 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 will not be required during this phase. Other firms that do not 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 FC-NVMe

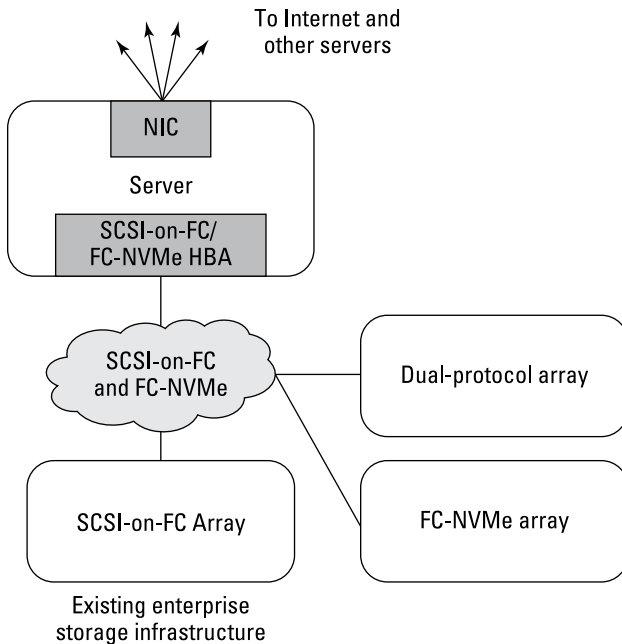
IT people are responsible for company data and productivity, and therefore rightly concerned about risk reduction (“de-risking,” in geek speak).

As IT departments plan to deploy NVMe-based arrays in production, they have two main options. They can either create some new NVMe infrastructure (see Figure 3-1), or they can leverage an existing infrastructure (see Figure 3-2). 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. That’s why dual-protocol concurrency is the big win for NVMe over Fibre Channel.



**FIGURE 3-1:** New separate infrastructure (not recommended).

By leveraging a “known quantity” such as a Fibre Channel SAN (storage area network), companies can easily support dual protocols and remove any risk associated with the inevitable question, “How long will the transition take?”



**FIGURE 3-2:** Dual-protocol infrastructure (recommended).

A dual-protocol approach is not unreasonable. A strong precedent for mixing protocols on top of Fibre Channel exists. Since 2001, it's been an established practice to support both FCP and FICON traffic simultaneously on the same Fibre Channel fabric (see the sidebar, "FICON and FCP").

Other considerations include:

- » **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 FC-NVMe Channel arrays, allowing other applications to use the data as “working memory.” Again, this can be achieved with existing infrastructure.

Fibre Channel lets IT administrators leverage already familiar elements such as zoning and name services. 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 Fibre Channel version of a mainframe (z/OS) storage protocol. Many firms that use mainframes have a recurring need to “publish” mainframe data assets so that 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 are not 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 is no notion of a gradual migration of applications from one side to the other, and so normal application servers are not configured to use both protocols.

Concurrent dual-protocol FCP/FC-NVMe 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 Fibre Channel device to which a node is attached.
- » **Name zoning** restricts access based on a device's World Wide Name (WWN).

Each of these methods restricts devices from accessing network areas they should not be visiting. Which approach 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 will be connected on one port.

You may come across references to *hard* and *soft zoning*. Modern FC fabrics use hard zoning, in which robust silicon-based logic blocks traffic between nodes that are not allowed to reach each other. In contrast, early products used software to do 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.

The point is that FCP and NVMe over Fibre Channel can both leverage FC zoning. Fibre Channel'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 are similar to DNS, but are network resident, greatly simplifying interoperability and management. This has long been the case for FCP, and remains true for FC-NVMe.

## Discovery and NVMe over Fibre Channel

The NVMe over Fabrics 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).

Fibre Channel's ability to deliver a dual-protocol FCP/FC-NVMe 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 FC-NVMe 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 will most likely 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 vendor's interoperability matrix and determine how your particular 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 might 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 might 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 Red Bull.

Not sure where to start? Here are some steps a typical IT department might take when establishing an NVMe-over-Fibre Channel test lab:

- » Set up a single server with an internal drive, a single switch, and a single FC-NVMe enabled array.
- » Connect the server to a lab IP network for access to other lab servers.

- » Explore the HBA config and storage array options you've read so much about.
- » Configure a volume to use the NVMe array. The details depend on your NVMe array management tools, but in general, this procedure is similar to provisioning a LUN. In NVMe, a namespace ID (NSID) is analogous to a LUN in SCSI.
- » Copy a file from the internal drive to the NVMe volume and back again.
- » Run your preferred performance testing application (such as Iometer) to benchmark your NVMe volume. Compare it to your internal volume.
- » 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 namespace ID) 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 IBM Network Advisor, Brocade SAN Health, and IO Insight. Ensure that these tools have been upgraded appropriately and are FC-NVMe 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 FC-NVMe online in your production environment.

As with any high-profile production change, you should think about doing “baseline” performance measurements. Be sure to include host CPU utilization as well as the storage metrics reviewed in Chapter 2. 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? The Boy Scout motto

applies here: Be prepared! You should have enough information by now to recognize whether the support request is linked to the recent change or not. The baselines you made earlier will help.

Even in the absence of a support call, you'll want to understand how much performance has improved by moving from SCSI to NVMe, because that information will help 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 might earn you a "Job well done!" from your boss. If you don't know the initial performance, you will not 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.



TIP

See Chapter 5 for details about the latest performance enhancements with FC-NVMe.

## 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 breast 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 FC-NVMe. 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 anxious 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.

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

- » The long and short of Remote Direct Memory Access
- » InfiniBand
- » iWARP
- » RoCEv2
- » Evaluating 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. NVMe over Fabrics also offers alternatives, although some might not be to your liking. It can be run over fabrics such as iWARP, RoCEv2, or InfiniBand, or simply NVMe over TCP. This chapter looks at a few of the pros and cons of each and examines such performance considerations as wire speed, architecture, 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

RDMA stands for remote direct memory access. It is a protocol designed years ago 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, human 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 adaptor received frames and wrote them (using DMA) into DRAM buffers associated with the networking stack. (The unavoidable DMA step is not a DRAM-to-DRAM copy, so it is 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, Fibre Channel was more motivated, and more able to implement a network adaptor/driver/stack architecture that eliminated the single copy. That's precisely what happened. Fibre Channel has been quietly delivering "zero copy" for the past two decades.

NVM Express (<https://nvmexpress.org>) hosts a white paper that describes two types of fabric transports for NVMe: NVMe over Fabrics using RDMA, and NVMe over Fabrics using Fibre Channel. Despite this explicit recognition of FC as an NVMe fabric, some RDMA advocates will claim that because NVMe over Fibre Channel does not use RDMA, it somehow is not an NVMe fabric, despite the fact that NVMe over Fibre Channel does not need RDMA, as we review later in this chapter. 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.

## InfiniBand

InfiniBand (IB) came on the scene more recently than Ethernet or Fibre Channel. Focused on server cluster communication, IB is designed to deliver RDMA natively. IB is focused on speed rather than 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 over Fabrics marketing efforts on promoting RoCE.

## iWARP

Though not widely deployed, iWARP is nearly ten years old. The acronym stands for “Internet wide-area RDMA protocol,” an IETF (Internet Engineering Task Force) standard described in five 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 is sent has been received by the sender. But iWARP's TCP basis is not 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 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 ACK 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 are not compatible with 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 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 of these three options 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 has not been well adopted by the networking community.



TECHNICAL  
STUFF

As a reliable streaming protocol, stand-alone TCP was never intended to guarantee packet or frame alignment. That’s because TCP does not 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, because the entire stream must first be decoded. TCP is a complex, software-based stack and thus the alignment question would have 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 SCSI stack that NVMe was created to replace.

## Yo, Rocky

RoCEv2 is short for RDMA over Converged Ethernet, version two. It's pronounced "Rocky vee two," and if you call it anything else you'll be laughed out of the server room. It is an odd standard in that it was developed by the InfiniBand Trade Association rather than IETF or IEEE, where most IP and Ethernet standards are developed and maintained.

The RoCEv2 name says "RDMA over Converged Ethernet," but it is a slight misnomer with Version 2, which runs over UDP and thus is 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 Fibre Channel during the mid-1990s. Although this goal is laudable, 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 Fibre Channel, see Table 4-1.

**TABLE 4-1 Ethernet-Based versus Fibre Channel**

Ethernet-based options for NVMe	NVMe over Fibre Channel
Developing a new fabric protocol standard	Built on T11-standardized Fibre Channel fabric protocol
Standards group dealing with challenges of scaling I/O commands, status, and data to the datacenter	Fibre Channel solved these problems when FCP protocol was developed for SCSI

*(continued)*

**TABLE 4-1 (continued)**

Ethernet-based options for NVMe	NVMe over Fibre Channel
Transport options: iWARP, RoCEv2 and (recently) TCP	Transport is Fibre Channel; runs on existing ASICs
iWARP and RoCEv2 use RDMA (TCP does not)	Does not need RDMA; leverages FCP
Complex network configuration if RDMA is enabled	Fibre Channel 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

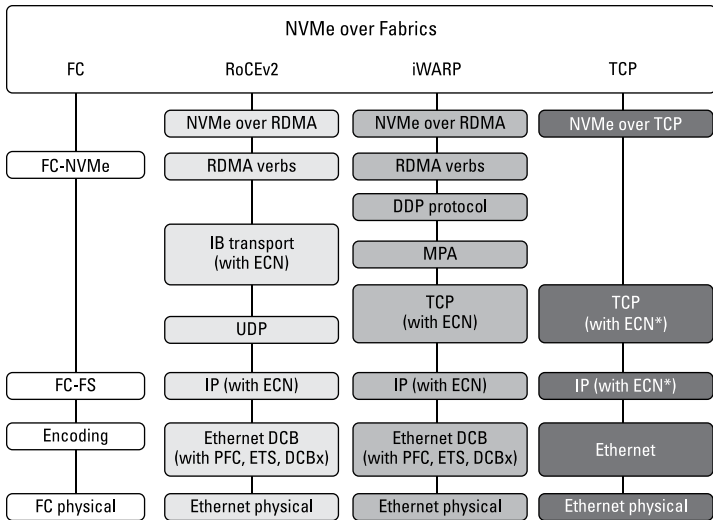
## Evaluating Ethernet-Based NVMe

Choosing Ethernet for your storage network’s physical layer is challenging, for several reasons. First, the industry is looking at four Ethernet-based NVMe protocols, iWARP, RoCEv2, NVMe over FCoE, and NVMe over TCP. To get low latency with iWARP or RoCEv2, you need to install RDMA-enabled NICs (RNICs). Big operators like Facebook will choose the commodity-oriented NVMe over TCP (parallel to today’s iSCSI), even if it’s slower. Of course, you must buy NVMe arrays that support your fabric choice, but who knows which will win? The upshot is that, until the dust settles, choosing any Ethernet-based NVMe protocol is risky.

Second, contrary to the recommendations of the NVMeExpress.org white paper on NVMe over Fabrics, Ethernet flow control does not use the reliable credit-based flow control mechanisms found in Fibre Channel, PCI Express, and InfiniBand transports.

Third, whether you choose iWARP or RoCEv2, you are choosing a multilayered network, with an associated increase in stack complexity (see Figure 4-1) to transport NVMe. 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 datacenter 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 is 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 are not ideal in today’s datacenter 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. Fibre Channel networking products are largely available from storage vendors, and it is not always easy playing vendors against one another for lower pricing.

## Smart shopping

The flip side is that most storage vendors have invested a significant amount of time in testing their arrays with the networking products they sell. They are 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 Fibre Channel Name Services and Fibre Channel Zoning, including target-driven zoning, features that are not 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 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 Fibre Channel 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
- » Taking the shortest path to NVMe over Fibre Channel
- » Considering SAN design with FC-NVMe in mind
- » Knowing what ANA is and why it matters
- » Exploring application use cases
- » Seeing that not all fabrics are created equal

# Chapter 5

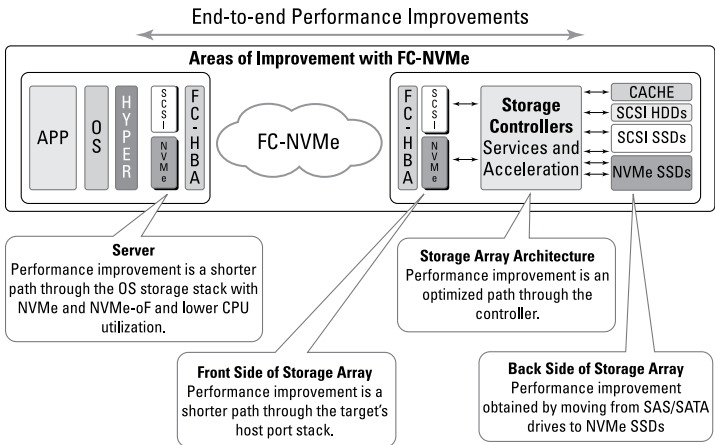
## Improving Performance with NVMe over Fibre Channel

**N**ow that you are ready to test the waters with NVMe over Fibre Channel, the big question is what level of performance improvements you can expect compared to SCSI/FCP, and why. To help you answer this question, this chapter begins with a look at where you can expect performance improvements to come from. NVMe is designed to exploit flash 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 Fibre Channel SAN to the storage array.

# Understanding How FC-NVMe 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. The areas where NVMe over Fibre Channel contributes to better performance are:

- » **Host side:** How FC-NVMe performs on the server compared to SCSI/FCP
- » **Storage array front end:** How FC-NVMe 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 SSDs in place of SAS/SATA attached SSDs improves performance compared to SCSI/FCP



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

## What about the fabric?

You may be thinking, “Wow! But wait, what about the Fibre Channel SAN?” Remember that the SAN supports SCSI (FCP) and FC-NVMe equally and there is no difference in performance across the Fibre Channel network between transporting FCP and FC-NVMe.

The transport happens with the same low latency and high performance that you are used to. Still, enhancements are being added to the Fibre Channel standard to augment high-performing FC-NVMe in case low-level errors occur (see the sidebar). The following sections review each point in a little more detail.

## The host side

On the host side, you will see significant CPU utilization improvement because the NVMe command set is more streamlined than SCSI, has multiple queues per core, multiple commands per queue, and separate asynchronous submission/completion queues, resulting in a leaner and meaner driver stack that executes more quickly while using fewer resources than SCSI. FC-NVMe requires less CPU processing time than FCP for the same workload, thus providing more CPU cycles for your applications. Over time, applications will exploit this leaner and meaner driver stack to further reduce application latency.

## SEQUENCE LEVEL ERROR RECOVERY FOR FC-NVMe

What is Sequence Level Error Recovery and why does it matter? In T11, the FC-NVMe-2 standard is under development and includes an enhancement for FC-NVMe called Sequence Level Error Recovery. 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 NVMe over Fibre Channel moves toward end-to-end storage latencies in the tens of microseconds with Storage Class Memory, the result will be remarkably better error recovery than with any other fabric technology.

For the most geeky bragging rights, you can look up some of the other highlights in the FC-NVMe-2 standard (<https://nvmexpress.org/resources/specifications>):

- Admin Command Determinism
- T10 Protection Information processing enhancements
- Submission Queue flow control processing enhancement

## The storage front end

On the storage array host port side (the front end), with the leaner NVMe protocol compared to SCSI, getting to the storage controller happens more quickly. 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 IOs across spinning media plus provided storage services, including data protection and, for example, encryption, compression, and dedupe without adding any noticeable latency because the media was orders of magnitude 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 beginning to enter 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

There are many compelling reasons to simply switch to FC-NVMe with an existing array. The ease of how FC-NVMe is made available by simply upgrading current storage arrays to the latest firmware level makes it a straightforward approach to adopt and implement NVMe over Fibre Channel.

## The storage array back end

With almost all arrays in the market today predominantly using SAS/SATA attached SSDs, you have an opportunity to reap performance improvement by using NVMe attached SSDs, which deliver lower latency and higher IOPS. Keep in mind this practice makes sense only when the overall array architecture can deliver the performance end-to-end through the array. Using the latest flash media technologies like 3D-TLC and 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 IOPS end-to-end.

Another new technology on the brink is PCIe Gen4 and Gen5, which will play a role in new storage array design decisions. For more information about PCIe Gen4 and what it means in the context of NVMe, see the sidebar, “PCIe Gen4.”

## PCIe GEN4

Since the middle of 2018, new NVMe SSDs entering the market are PCIe Gen4 compatible. The PCIe Gen 4.0 standard specifies a 16 Gbps data link speed with up to 16 links or lanes delivering 64 Gbps, which is double PCIe Gen3's maximum of 32 Gbps. Once PCIe gen4 x86 motherboards become available, the market likely will begin to offer new NVMe storage arrays delivering unprecedented performance. A PCIe Gen5 standard is underway with a target date in 2019 and provides for 128 Gbps from a 16-lane implementation.



REMEMBER

In addition, rearchitecting multi-threaded applications and hypervisors to take advantage of multi-queue properties of NVMe is an area that will likely see improvements over time.

### Handling NVMe support with a software upgrade

The first storage arrays in the market to deliver FC-NVMe make it easy for you to start using FC-NVMe. All you need is a simple software upgrade of the storage array controllers to the latest firmware version, and off you go provisioning NVMe NSIDs and LUNs concurrently on the storage arrays. Using Gen 6 or (already available) Gen 7 HBAs in your server, you are ready to use NVMe over Fibre Channel.

## Improving Performance

Now that you understand why you should see performance improvements and where these come from in an end-to-end NVMe-over-Fibre Channel solution, this section shows how big these improvements can be.

Both vendors with products in the market at the time of writing have demonstrated and documented improvements across latency, IOPS, and CPU utilization on the host side with FC-NVMe compared to SCSI/FCP. Although the numbers differ between the systems and tests, they both demonstrated improvement of up to 30-50 percent faster application execution, as well as up to 25-50 percent more IOPS while consuming 30-50 percent less CPU resources for the same workload. Both demonstrations showcased a typical OLTP workload profile for transactional data bases.

## 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 an NVIDIA chip — with multiple 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.

You are likely thinking, “Wow, that’s great! With a simple software upgrade on an FC-NVMe capable array you can achieve 30-50 percent faster application execution!” Therein lies the rub. If you are satisfied simply because the storage chain is now faster, you miss the greatest opportunity NVMe over Fibre Channel presents.

Chapter 2 introduces 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.

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 than with virtual machines.

## Considering SAN Design with FC-NVMe in Mind

One question that may come to mind is how running NVMe over Fibre Channel influences your SAN design. From a SAN design perspective, the areas to pay attention to are somewhat related.

In the same way the transition to All Flash Arrays in the datacenter has increased the IO density of the AFA storage ports, delivering orders of magnitude more IOPS per port, the host-to-target port ratio has likewise increased. This increase will continue with FC-NVMe. 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.



TIP

A *slow drain device* is a host or storage array that does not return buffer credits in a timely manner to the switch. This causes frames to back up through the fabric and thus causes fabric congestion. In a fabric, many flows share the ISLs, as well as 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 IBM b-type Slow Drain Device Quarantine (SDDQ) feature enables 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.

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 FC-NVMe, 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 FC-NVMe storage arrays.

## Understanding Why Monitoring Is Important

Monitoring is a cornerstone when managing a high-performance infrastructure such as a SAN. Being aware of the importance of monitoring, you likely already use all or part of the IBM b-type Fabric Vision suite of tools. 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.

Having MAPS enabled is the baseline. Complementing MAPS with IO Insight capability on Gen 6 platforms adds built-in device input/output (I/O) latency and performance instrumentation in Flow Vision. With the latest IBM b-type Gen 6 products, the IO Insight capabilities include FC-NVMe 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 in the same way for SCSI/FCP target access as for FC-NVMe targets. Depending on the storage array, implementing NVMe over Fibre Channel can alter how you must zone for NVMe Controller and NSID access. The reason is that some storage arrays implement the NVMe Controller target ports as logical interfaces (child WWPN) behind the physical target port (WWPN).

### PEER ZONING

Peer Zoning allows a "principal" device to communicate with the rest of the devices in the zone. The principal device manages a Peer Zone. Other "non-principal" devices in the zone can communicate with the principal device only; they cannot communicate with each other.

In a Peer Zone setup, principal-to-non-principal device communication is allowed, but non-principal-to-non-principal and principal-to-principal device communication are not allowed. This approach establishes zoning connections that provide the efficiency of single initiator zoning with the simplicity and lower zonedb memory usage as with one-to-many zoning. Typically, a Peer Zone has a single principal device and one or more non-principal devices, but configurations having multiple principal devices are allowed. Peer Zones are not mutually exclusive with traditional zones; multiple zoning styles can coexist within the same zoning configuration and fabric.





TIP

In principle, zoning works the same way as you likely are already familiar with for 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).

## Knowing What ANA Is and Why It Matters

Multipath IO support with FC-NVMe 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 is not available or not functional for LUN access. ALUA is a SCSI standard that is 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 are 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 cannot 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 likely have an application in mind that is 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, artificial intelligence, 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 rearchitected 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 and

Internet of Things (IOT) or edge devices generating vast amounts of unstructured data.

- 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, 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).



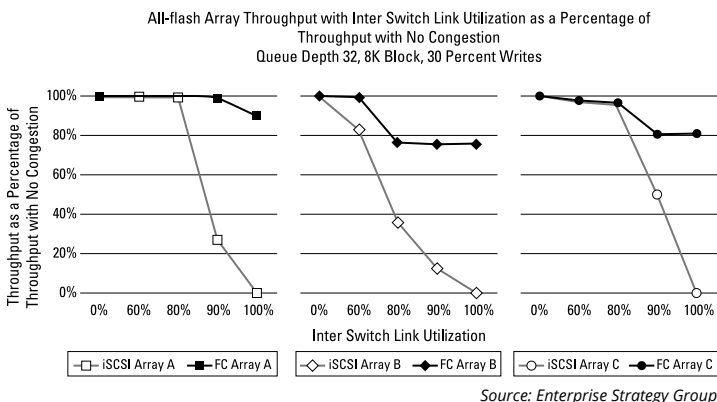
TIP

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 all-flash arrays (AFAs) using SCSI over Fibre Channel and 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 Fibre Channel and Ethernet with the host and the storage remaining the same.

As shown in Figure 5-2, 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 Fibre Channel 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 is 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 is of utmost importance to know what the behavior or impact is when congestion occurs. Applications do not 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.



TIP

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 Fibre Channel continues to provide significant throughput? Because the two fabrics have different flow control and congestion control mechanisms. With Fibre Channel the network is responsible for guaranteed delivery and does not 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. Thus, as the network becomes congested, an intermediary node may run out of buffer space and start dropping packets, which requires retransmission.

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, Fibre Channel 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 if there are still buffers available to send. Each frame sent decrements the receiver buffer count, and each frame

acknowledgement increments the receiver buffer count. A sender cannot send more data if the receiver buffer count is zero. Thus, 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.

Fibre Channel'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. Thus, as demonstrated by these tests, Fibre Channel traffic continues to flow even as congestion approaches 100 percent.



REMEMBER

The bottom line is that during network congestions, NVMe over TCP (Ethernet) performs in the same way as the ESG paper demonstrates for iSCSI.

## Chapter 6

# Ten NVMe over Fibre Channel Takeaways

**Y**ou're ready to get serious about NVMe over Fibre Channel. Here are ten key points to get you started:

- » **Reduce risk to business operations by extending the existing SCSI/FC network to include NVMe/FC.** Installing an entirely new non-Fibre Channel (FC) fabric infrastructure to adopt NVMe is an all-or-nothing approach. Additionally, which Ethernet based implementation do you deploy — iWARP, RoCEv2, or TCP? 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.
- » **Improve server CPU utilization and ROI with NVMe/FC.** NVMe provides significant host CPU utilization benefits by exploiting NVMe's streamlined IO commands, multiple queues per core, and asynchronous submission/completion queues. Improved resource utilization delivers greater host and storage resource utilization so that you can get the most from your resources and achieve a more scalable and more cost-efficient IT infrastructure. You may see a reduction in software core licensing costs, as well.

- » **Drive more application workload because NVMe's input/output operations per second (IOPS) will have as much impact as its 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.
- » **Install NVMe-ready devices today to allow applications to exploit the next generation of low latency flash — Storage Class Memory (SCM).** This next generation of memory has significantly lower latency than today's flash technology. SCM also comes with a higher cost and in order to get the full low latency benefit, it can only be achieved with the NVMe protocol. You will leverage a tiered infrastructure blending SCM and traditional flash to optimize data placement and cost of the infrastructure.
- » **Leverage the multiple NVMe flash media form factors to cost optimize your IT infrastructure.** NVMe media comes in a variety of form factors, including add-in PCIe cards, 2.5-inch SFF drives, M.2 and NF1 expansion, NVDIMM, and others. Different vendors leverage different forms of the media in their solutions to optimize storage system performance. Understand how the technology deployments affect your application, because that's what matters to you!
- » **Allow the NVMe future to come at its own pace.** Relative to SCSI, the NVMe protocol is fairly young, having been born in 2014 and will continue to quickly evolve. There has been a rapid set of enhancements with the industry adoption of flash technology and will continue with the NVMe specifications. Expect client adoption to start taking hold in 2019 and NVMe to replace SCSI in the coming years. A significant evolving area is for all the OS vendors to exploit the built-in multipathing capability of ANA.
- » **Include vendor support in your NVMe-FC 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 thorough interoperability testing and, when there are issues, industry leading support. IBM is an



industry leader with complete end-to-end NVMe storage strategy across the entire block, file, and object storage portfolio. Demand enterprise testing and support for your NVMe solutions, so be sure to ask all your vendors about their NVMe storage and fabric interoperability testing, deployment plans, and backend support capabilities.

- » **Fibre Channel (FC) and Ethernet/IP are optimized differently.** Fibre Channel was born and raised during an era of explosive Ethernet and IP growth. FC succeeded because it was purpose-built and 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 for data traffic, but, like the long-standing SCSI protocol, come with additional baggage over the years.
- » **Dedicate a separate storage fabric for best results.** Storage experts know that reference architectures from mission-critical storage vendors call for dedicated storage fabrics. This is why a dedicated FC fabric with HBAs is the low-risk choice for your storage traffic compared to running storage over a shared Ethernet/IP fabric, which requires gambling on either RDMA-capable NICs or TCP offload engine technology (TOE). For best performance and highest availability, we recommend a dedicated storage fabric for either FC or Ethernet, or both.
- » **Maximize your Fibre Channel fabrics and minimize your migration by running both SCSI/FC (FCP) and FC-NVMe.** By supporting dual protocols (FCP/FC-NVMe), Fibre Channel fabrics offer big advantages by providing ultra-low latency data access 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 is likely to take years.

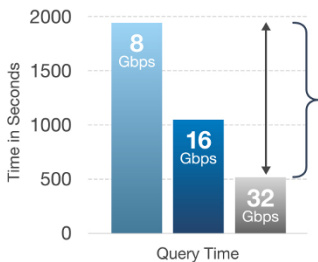
Want to get  
**RACE PERFORMANCE** from your  
**IBM FLASHSYSTEM SOLUTION?**

# Turbo Charge your Network!

## Upgrade to IBM b-type 32 Gbps Gen 6

NVMe-Ready storage networking that maximizes flash performance!

### Benefits for upgrading to Gen 6 storage infrastructure



Connectivity to 8 Gbps Flash Storage

**71%**  
Faster query  
completion

**4x**  
Complete more  
database queries

\* [www.demartek.com/Demartek\\_Emulex\\_LPe32000\\_Gen6\\_FC\\_Evaluation\\_2016-03.html](http://www.demartek.com/Demartek_Emulex_LPe32000_Gen6_FC_Evaluation_2016-03.html)

To learn more about IBM FlashSystem and b-type SAN solutions:

[ibm.biz/flashstorage](http://ibm.biz/flashstorage)

[ibm.biz/san-btype](http://ibm.biz/san-btype)



# Move into the NVMe over Fibre Channel future

NVMe (Non-Volatile Memory Express) is today's new massively parallel, ultra-low latency memory protocol. NVMe over Fibre Channel extends this blockbuster new protocol to the scale of enterprise storage. Fibre Channel, the premium datacenter fabric, can transport both NVMe and 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



**BROCADE**<sup>®</sup>  
A Broadcom Company


**Brian Sherman** is an IBM Distinguished Engineer. **Marcus Thordal** is Principal Solution Architect at Brocade. **Kip Hanson** is a one-time IT geek turned freelance writer. He spends his days making difficult subjects easy to understand, while lamenting the loss of floppy disks and DOS prompts.

Go to **Dummies.com**<sup>®</sup>  
for videos, step-by-step photos,  
how-to articles, or to shop!

ISBN: 978-1-119-60267-5  
Not For Resale



for  
**dummies**<sup>®</sup>  
A Wiley Brand

 Also available  
as an e-book

# **WILEY END USER LICENSE AGREEMENT**

Go to [www.wiley.com/go/eula](http://www.wiley.com/go/eula) to access Wiley's ebook EULA.