Ultra-Fast NVMe Storage Networks for Next Generation Flash Arrays

Sponsored by the Fibre Channel Industry Association (FCIA)

Moderator – Mark Jones FCIA President and Director Tech Marketing Broadcom Inc.

Fibre Channel
Industry Association

- Fibre Channel – Legacy of Interoperability, reliability and robustness
  - Products have been shipping in volume for more than 20 years – 120M ports shipped, 43M in current use.

- Key Factors to Fibre Channel's success
  - Industry-wide participation in FC standards – INCITS T11
  - ~Avg of Two Plugfests per year (39 total) that ensure vendors conform to industry standards

FCIA FC-NVMe Plugfest

- July 23, 2018 - 4th FC-NVMe plugfest
  - 13 Companies/products tested
    - HBAs, Switches, Storage Arrays, Analyzers/Jammers
  - Key Accomplishments
    - Testing of End-end commercial available products
    - Multi vendor interoperability, standards conformance
    - Data Integrity validation over switch multi-hop fabrics
    - Error injection to validate correct FC-NVMe and FC recovery
    - Concurrent FC-NVMe + FC over same Initiator, fabric, target ports
    - "Big Build" overnight stress testing of all of the above.

How to Participate with the FCIA

- ## Https://fibrechannel.org
  - 2018 FC Solutions Guide
  - FC Roadmaps
  - Plugfest Information
  - FC Education links

- ## FC Education
  - www.Brightalk.com - FCIA

- ## Social Media
  - LinkedIn, Twitter @FCIAnews

Agenda

- Curt Beckman – The New Normal in Storage Latency

- Rupin Mohan – NVMe: A New Language for Storage

- Craig Carlson – FC-NVMe Status and Updates

- Dennis Martin – FC-NVMe Test Results

- David Rodgers – Fibre Channel Test and Measurement

- Q&A

The New Normal in Storage Latency

Curt Beckmann

Product Architect

BSN, Broadcom

6

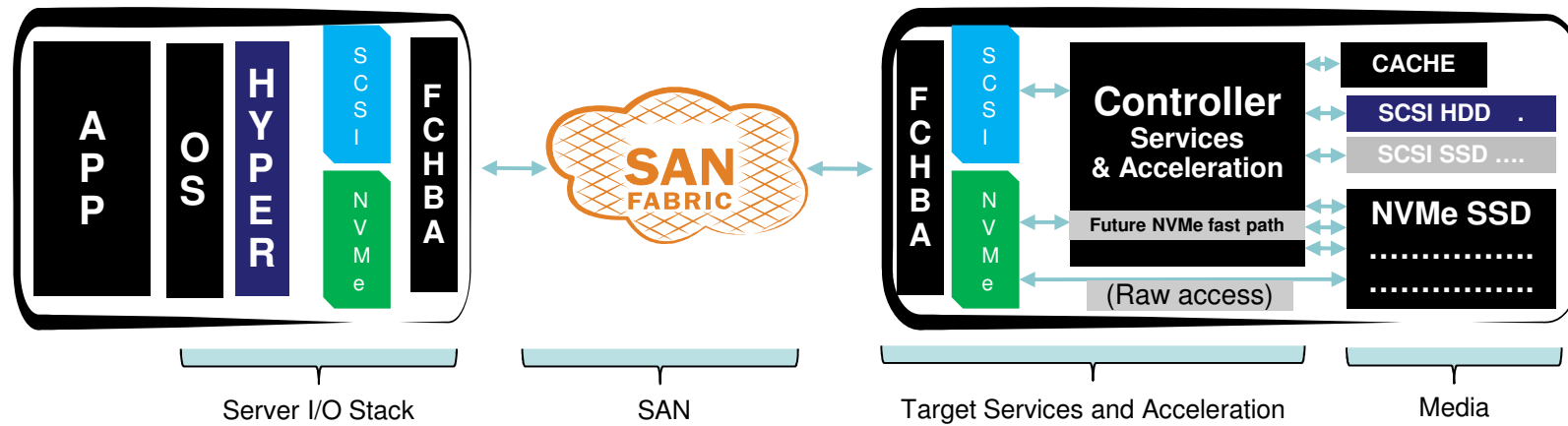# How Flash and NVMe are changing storage latency

- Many elements in networked storage latency
  - Flash and NVMe change which elements matter
- In HDD era, arrays could use cache and spindle count
  - Array acceleration hid SW feature overhead
  - Availability "zero cost" in performance
- Fast SSDs make protocol, feature costs visible
  - Networked storage still has benefits, of course
  - But some apps may prefer speed over features

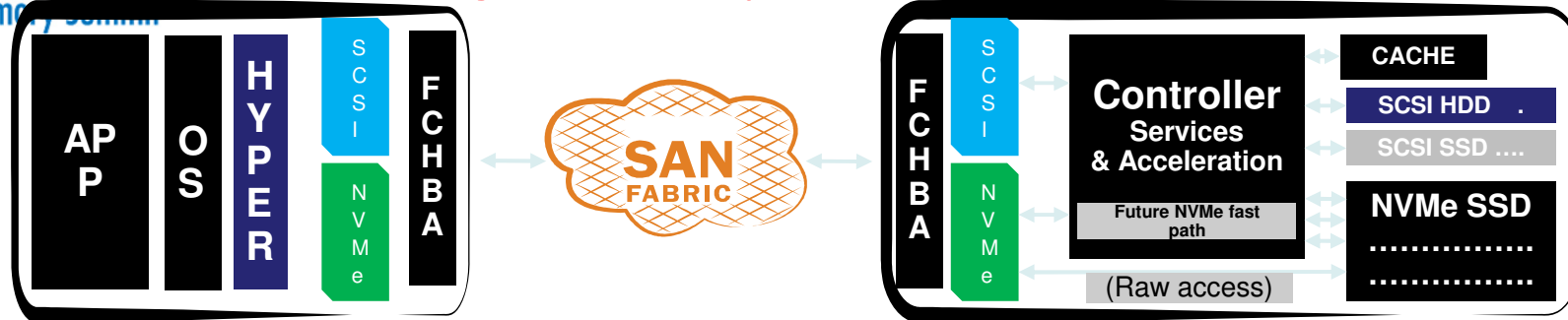# Elements in application storage latency

Server I/O Stack | SAN | Target Services and Acceleration | Media

How Flash and NVMe are changing SAN Storage Latency

Media Eras →	Fast HDD (estimate)	SCSI Flash (estimate)	NVMe o FC (anticipated)
Server IO Stack	40 µs	30 µs	7 µs
SAN (no queuing)	*6 µs	3 µs	3 µs
Avg Services	250 µs	150 µs	20 µs
Avg Acceleration	-200 µs	0 µs	0 µs
Media	3 ms	50 µs	10 µs
Raw Access Total	3.0 ms	(?) 83 µs	20 µs
Services Total	3.1 ms	233 µs	40 µs

*16 GFC
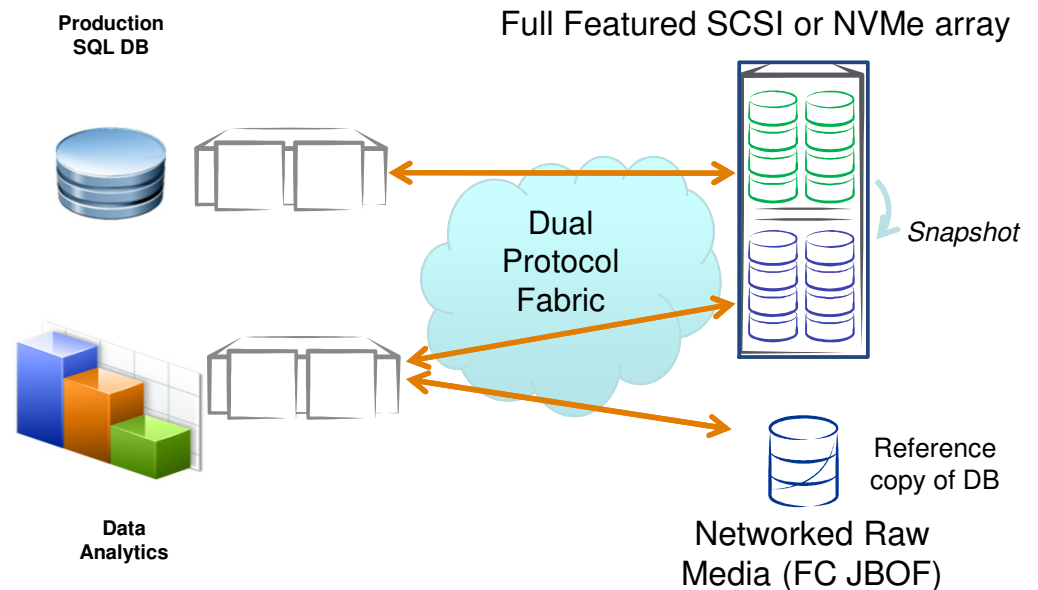
In HDD Era, very little incentive to use raw media with networked storage

In "anticipated" NVMe SSD era, use of raw media may be justified… but many applications may want both!

FCIA, www.fibrechannel.org
Flash Memory Summit 2018
Santa Clara, CA

Concurrent Enterprise / Raw Media Use Case

Analytics on Active DB

- Using sensitive data for ML is an effective mechanism to:
  - Increase revenue
  - Build customer loyalty
- ML is data intensive, and want results soon as possible
  - But active DB needs protection (adds latency)
  - Hammering active DB with ML slows both down
  - What do do?
- Separate the problem:
  - Protect DB master on full-featured volume (as now)
  - Regularly snapshot / Clone the DB to Raw Media
  - Use Raw Media reference copy for ML
  - Do both on the same infrastructure you use today

Production SQL DB

Full Featured SCSI or NVMe array

Dual Protocol Fabric

Snapshot

Data Analytics

Reference copy of DB

Networked Raw Media (FC JBOF)

FCIA, www.fibrechannel.org
Flash Memory Summit 2018
Santa Clara, CA

NVMe: A new language for storage

Rupin Mohan

Director R&D, CTO SAN

HPE Storage

NVMe – Technology Disruption

Traditional Storage Arrays

1. Storage Controller runs SCSI

2. Front end FC/iSCSI

3. Backend SAS/SATA

4. Software Feature Rich

FCIA, www.fibrechannel.org
Flash Memory Summit 2018
Santa Clara, CA

Hybrid Storage Arrays

1. Storage Controller runs SCSI with upgraded back end – Controller does SCSI-NVMe translation with NVMe drives in the backend

2. 3D Cross Point for Metadata stores on NVMe stack

3. Front end, FC-NVMe

4. Software Feature Rich

NVMe Storage Arrays

1. Storage Controller only runs NVMe

2. Backend NVMe Drives (PCIe, NVMe over Ethernet/Infiniband)

3. Frontend NVMe (FC-NVMe. NVMe over Ethernet)
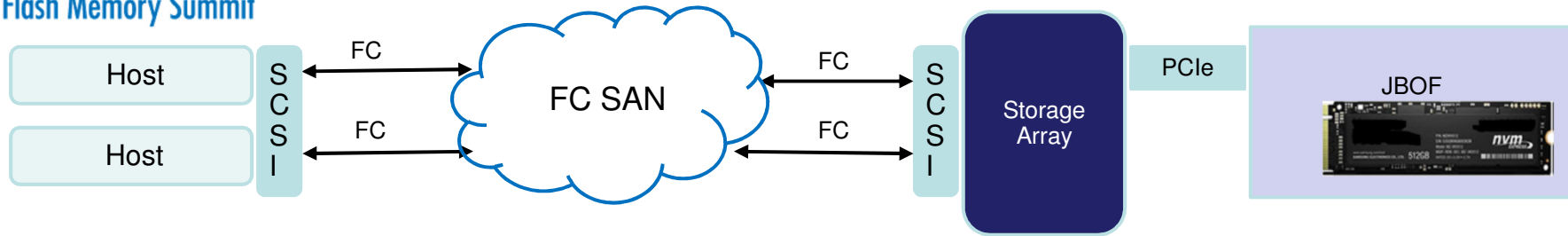
4. Software Features - low
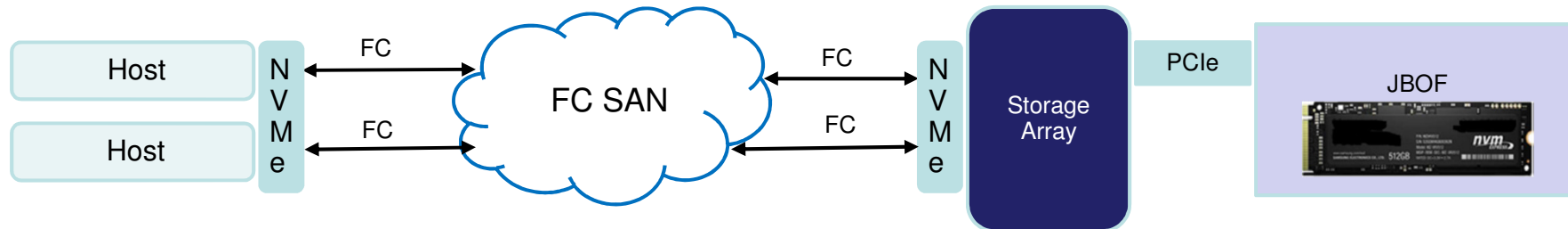
NVMe over Fabrics Use Cases
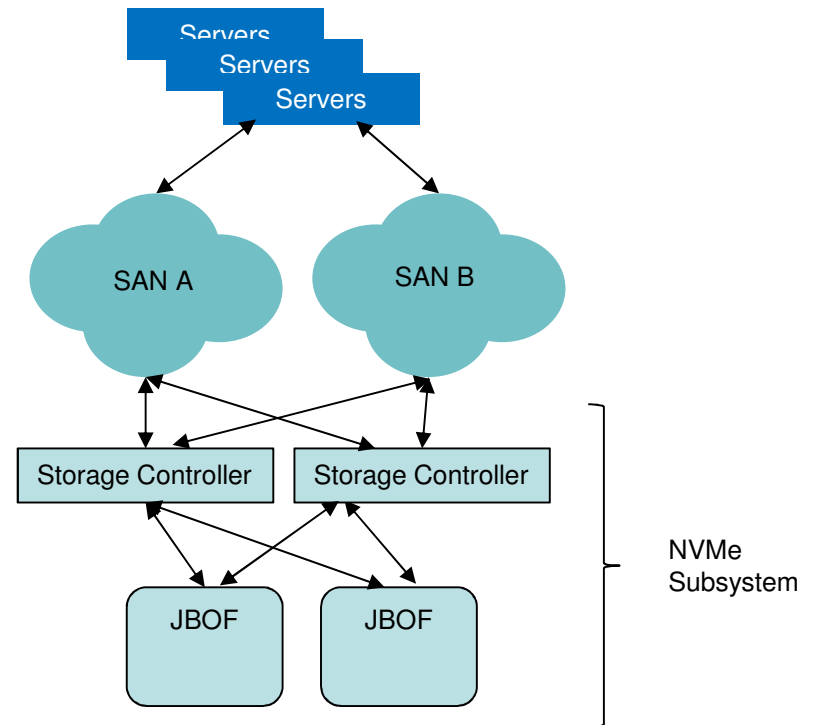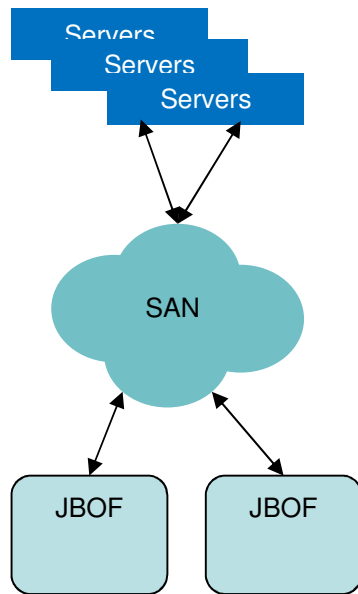
NVMe-oF deployment (FC)

NVMe storage attached in the backend

NVMe end to end using FC

NVMe-oF deployment (Ethernet)

FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION

Host

Host

NVMe

RDMA or TCP/IP

Ethernet

RDMA or TCP/IP

Ethernet Fabric

RDMA or TCP/IP

Ethernet

RDMA TCP/IP

NVMe

Storage Array

PCIe

JBOF
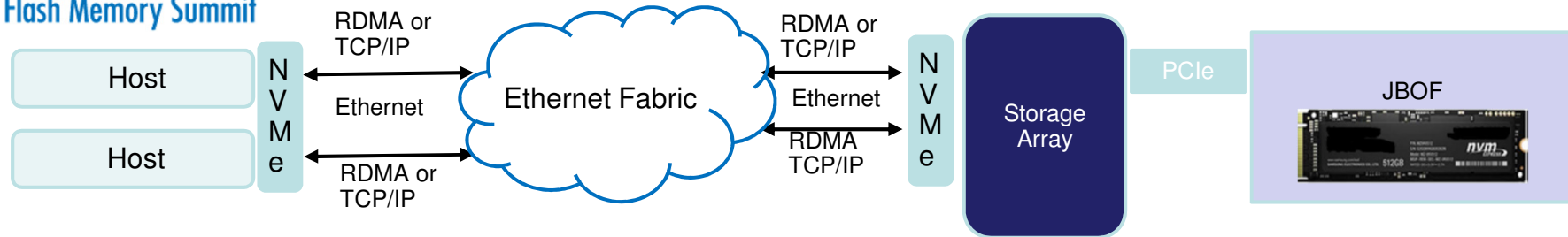
RDMA - RocEv1, RocEv2 and iWARP
(Infiniband not included in pictures)

NVMe-oF in the backend

Host

Host

NVMe

RDMA or TCP/IP

Ethernet

RDMA or TCP/IP

Ethernet Fabric

RDMA or TCP/IP

Ethernet

RDMA or TCP/IP

NVMe

Storage Array

Ethernet

Ethernet Fabric

JBOF

JBOF
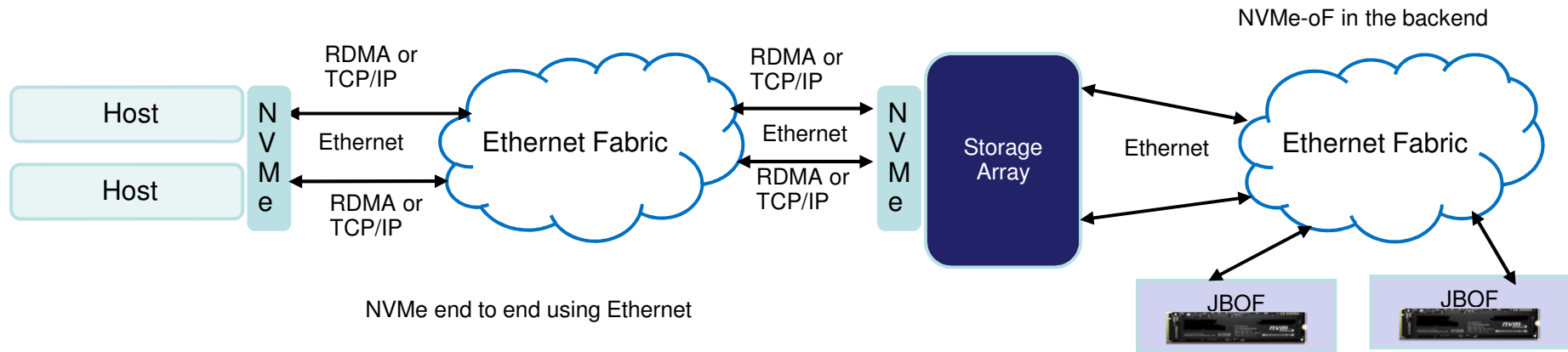
NVMe end to end using Ethernet

FCIA, www.fibrechannel.org
Flash Memory Summit 2018
Santa Clara, CA

The landscape today....

Protocol	Latency	Scalable	Performance	Enterprise Footprint
Fibre Channel	Lower	Yes	High	Reliable / Mature Storage Fabric
RoCEv2	Lowest	Yes	Higher	Negligible
iWARP (Intel)	Medium	Yes	Medium	Negligible
TCP	High	Yes	Medium	Medium with iSCSI
InfiniBand	Lowest	Limited	High	None

FC-NVMe Status and Update

Craig W. Carlson

Marvell

FC-NVMe is real

- **FC-NVMe (Fibre Channel over NVMe)**
  - First revision of standard completed in 2016
  - Products are now available
    - Based on existing trusted hardware/software platforms

Future development

- **FC-NVMe-2 under development now**
  - Major new feature is Enhanced Error Recovery
    - Allows for transport level recovery of lost or corrupted commands
      - Occurrence of this is rare, but not impossible
  - Adds additional reliability to already reliable FC SANs

Fibre Channel

- Ratification of 64GFC serial and 256GFC parallel is under way

- Work started on 128GFC serial with 512GFC parallel following

FCIA Roadmap

Product Naming	Throughput (Mbytes/s)	Line Rate (Gbaud)	T11 Specification Technically Complete (Year)*	Market Availability (Year)*
1GFC	200	1.0625	1996	1997
2GFC	400	2.125	2000	2001
4GFC	800	4.25	2003	2005
8GFC	1,600	8.5	2006	2008
16GFC	3,200	14.025	2009	2011
32GFC	6,400	28.05	2013	2016
128GFC	25,600	4X28.05	2014	2016
64GFC	12,800	28.9 PAM-4 (57.8Gb/s)	2017	2019
256GFC	51,200	4X28.9 PAM-4 (4X57.8Gb/s)	2017	2019
128GFC	25,600	TBD	2020	Market Demand
256GFC	51,200	TBD	2023	Market Demand
512GFC	102,400	TBD	2026	Market Demand
1TFC	204,800	TBD	2029	Market Demand

FCIA, www.fibrechannel.org
Flash Memory Summit 2018
Santa Clara, CA

21

FC-NVMe Test Results

Dennis Martin

President, Demartek

Independent Test Lab and Analyst

◇◇ Demartek

About Demartek

- Industry Analysis and ISO 17025 accredited test lab

- Lab includes enterprise servers, networking & storage (6/12Gb SAS, 10/25/40/100GbE, 8/16/32GFC)

- We prefer to run real-world applications to test servers and storage solutions (databases, Hadoop, VMware, etc.)

- Demartek is an EPA-recognized test lab for ENERGY STAR Data Center Storage testing

- Website: https://www.demartek.com/Testlab/

Demartek

- Full Report
- Short Report
- Video Report
- Infographic
- Commentary
- Speaking

Storage Interface Comparison
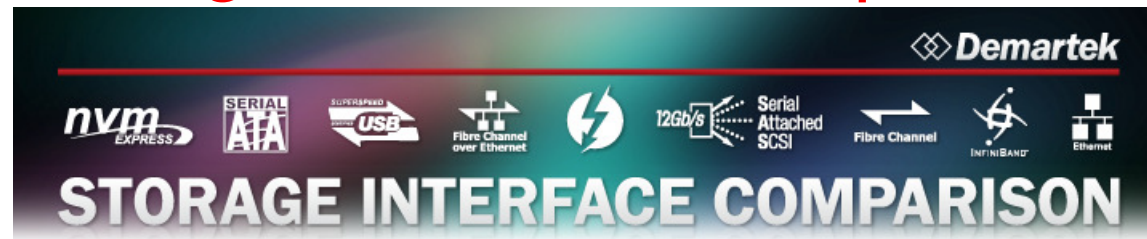
- Free reference page on demartek.com
  - https://www.demartek.com/Storage-Interface-Comparison/
  - Search for "**storage interface comparison**" in your favorite search engin

- Popular page – includes interactive PDF for download

- Provides comparison of storage interfaces
  - FC, FCoE, IB, iSCSI, NVMe, PCIe, SAS, SATA, Thunderbolt, USB
  - Transfer rates, encoding schemes, history, roadmaps, cabling, connectors

- ***We're not a product vendor – we use these technologies in our lab***

FC-SCSI vs. FC-NVMe

- ## May 2018 Demartek Evaluation: *Performance Benefits of NVMe™ over Fibre Channel – A New, Parallel, Efficient Protocol*

  https://www.demartek.com/ModernSAN/

May 2018 — Demartek

Performance Benefits of NVMe™ over Fibre Channel – A New, Parallel, Efficient Protocol

NVMe™ over Fibre Channel delivered **58% higher IOPS** and **34% lower latency** than SCSI FCP. (What's not to like?)

Executive Summary

NetApp's ONTAP 9.4 is the first generally available enterprise storage offering enabling a complete **NVMe™ over Fibre Channel (NVMe/FC)** solution. NVMe/FC solutions are based on the recent T11/INCITS committee **FC-NVMe** block storage standard, which specifies how to extend the NVMe command set over Fibre Channel in accordance with the NVMe over Fabrics™ (NVMe-oF™) guidelines produced by the NVM Express™ organization.

Fibre Channel is **purpose-built for storage** devices and systems and is the de facto standard for storage area networking (SAN) in enterprise datacenters. Fibre Channel operates in a lossless fashion with hardware offload Fibre Channel adapters, with hardware-based congestion management, providing a reliable, credit-based flow control and delivery mechanism, meeting the technical requirements for NVMe/FC.

Today's Fibre Channel adapters have the added benefit of being able to run traditional Fibre Channel Protocol (SCSI FCP) that uses the SCSI command set **concurrently** with the NVMe over Fibre Channel command set in the same adapter, the same Fibre Channel Network, and the same Enterprise All Flash Arrays (AFAs). The NetApp AFF A700s is the first array to support both SCSI FCP and NVMe/FC concurrently on the same port. This provides **investment protection** for existing FC adapters while offering the *performance benefits of NVMe/FC with a simple software upgrade.* Modern Fibre Channel switches and host bus adapters (HBAs) already support both traditional SCSI FCP and NVMe/FC concurrently.

For this test report, Demartek worked with NetApp and Broadcom (Brocade and Emulex divisions) to

demonstrate the benefits of NVMe over Fibre Channel on the NetApp AFF A700s, Emulex Gen 6 Fibre Channel Adapters, and Brocade Gen 6 Fibre Channel SAN switches.
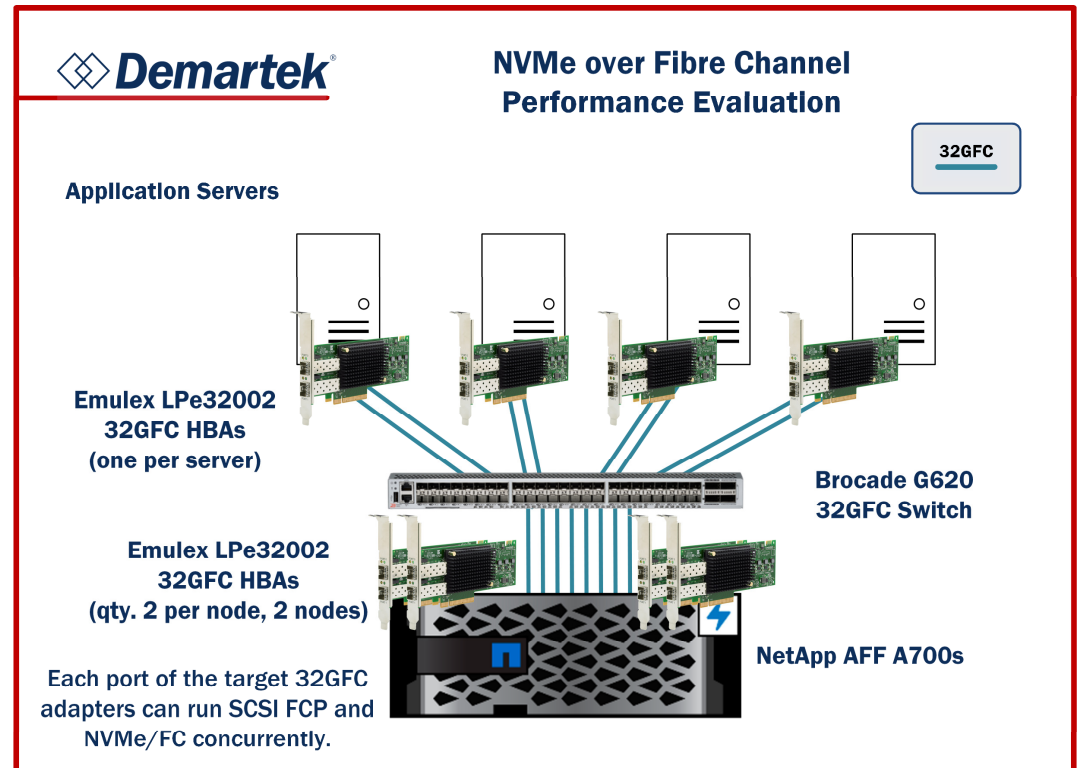
Key Findings and Conclusions

> **NVMe/FC enables new SAN workloads:** Big data analytics, Internet of Things (IoT) and A.I. / deep learning will all benefit from the faster performance and lower latency of NVMe/FC.

> **NVMe/FC accelerates existing workloads:** Enterprise applications such as Oracle, SAP, Microsoft SQL Server and others can immediately take advantage of NVMe/FC performance benefits.

> *Test results:* in our tests, we observed up to **58% higher IOPS** for NVMe/FC compared to SCSI FCP *on the same hardware*. We also observed minimum differences, depending on the tests, of 11% to 34% lower latency with NVMe/FC.

> **NVMe/FC is easy to adopt:** All of the performance gains we observed were made possible by a software upgrade.

> **NVMe/FC protects your investment:** The benefits we observed were with existing hardware that supports 32GFC.

> **NVMe/FC Datacenter consolidation:** More work can be completed in the same hardware footprint with increased IOPS density.

The Test

- Comparison of FC-SCSI to FC-NVMe

- Same hardware, different protocol

Demartek

NVMe over Fibre Channel Performance Evaluation

32GFC

Application Servers

Emulex LPe32002 32GFC HBAs (one per server)

Brocade G620 32GFC Switch

Emulex LPe32002 32GFC HBAs (qty. 2 per node, 2 nodes)

NetApp AFF A700s

Each port of the target 32GFC adapters can run SCSI FCP and NVMe/FC concurrently.

Results: Random Read 4KB

Random Read 4KB
Latency vs. IOPS

54% higher IOPS
at 2300 μs

At least 34% lower latency

53% higher IOPS
at 450 μs

Latency (μs)

IOPS

Note: all measurements taken on
a single-node A700s. Standard
implementations are dual-node.

◆ SCSI FCP   ◆ NVMe/FC

https://www.demartek.com/ModernSAN/

Zoom-in: RR 4KB

Random Read 4KB
Latency vs. IOPS (zoom in)

At least
34% lower latency

Note: all measurements taken on a single-node
A700s. Standard implementations are dual-node.

◆ SCSI FCP   ◆ NVMe/FC
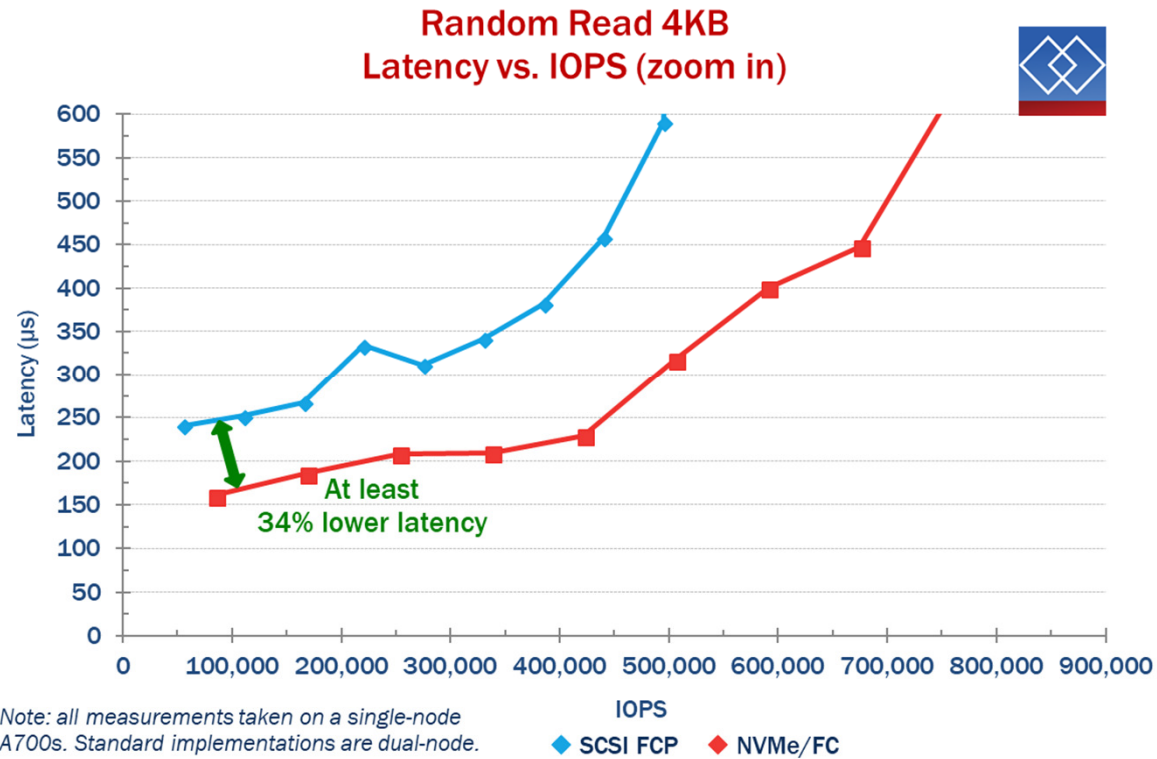
Oracle 80-20 8KB

Oracle 80-20 8KB
Latency vs. IOPS

47% higher IOPS
at 1085 μs

58% higher IOPS
at 375 μs

At least 18% lower latency

Latency (μs)

IOPS

Note: all measurements taken on
a single-node A700s. Standard
implementations are dual-node.

◆ SCSI FCP   ◆ NVMe/FC

https://www.demartek.com/ModernSAN/
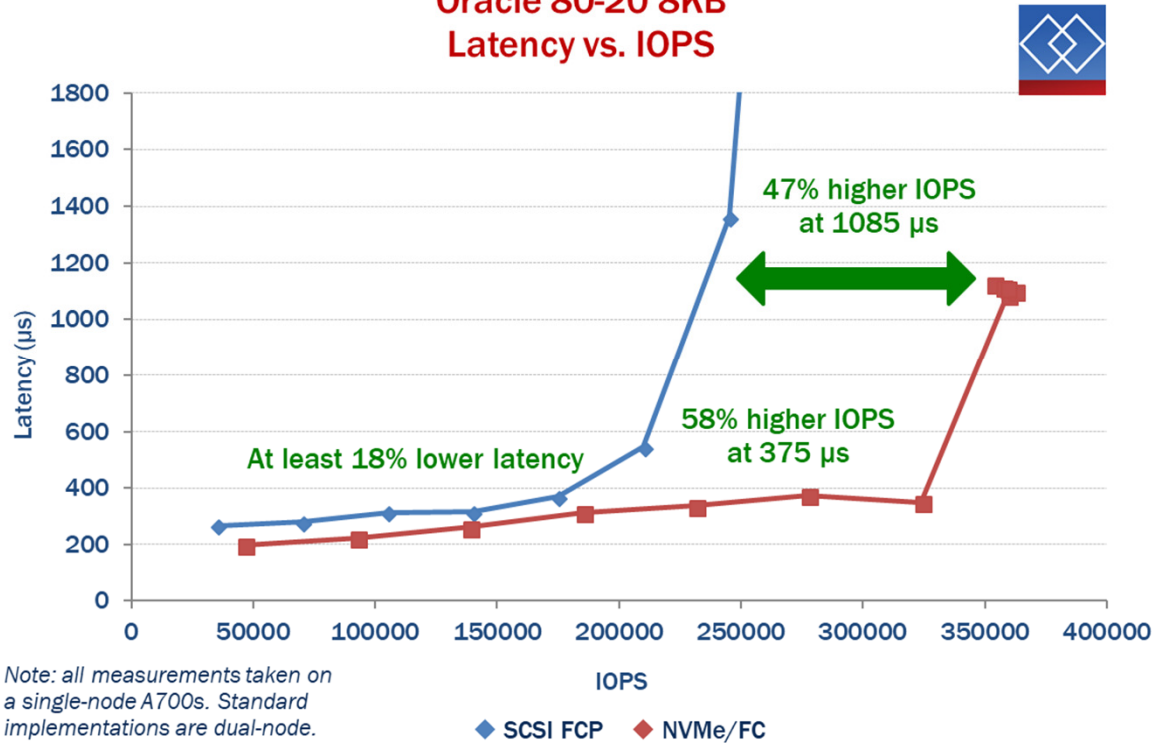
Test & Measurement in Fibre Channel From Inception to Support

Protocol Awareness is Required from Initial Phy Development for Successful NVMe Application Support

David J. Rodgers

30

High-Speed Fabric Designs and Protocol Analysis

Basic Premise:

Mission Critical Storage demands, i.e. NVMe/oF, are fueling the exponential growth of Fibre Channel speeds, protocols, port counts and densities. The challenge to meet the demands of users and applications requires adaptation and evolution of test and measurement tools and practices.

Specific to Phy Layer Designs, link interfaces have evolved to include improved communications schemes and adapted corresponding high-speed transmitter training and equalization practices.

The impact on hardware designs requires protocol awareness beginning with initial design, through validation, and in the field after deployment.

Universal T&M Considerations

- **Common to all stages of Fibre Channel Fabric Development, Deployment, and Support for NVMe**
  - What issue(s) are we trying to understand and correct?
  - When and How does the issue manifest?
  - Is the issue reproducible?
  - Can root cause be definitively determined?
  - What are the curative measures?
  - Can you test the 'fix'?
  - What are the Cost considerations to vendors, customers?

Fibre Channel T&M Today

- **The physical communications "Channel" must be stable**
  - Minor Imperfections, once considered 'routine' and unremarkable are no longer "minor"

- **Vendor Interoperability is required!**
  - From Switch to HBA to Interconnect options, Vendor offerings must work together

- **Specification Conformance**
  - FC Physical/Communications Layer
  - NVMe iterations

- **Line-rate Capture/Analysis Tools Needed**
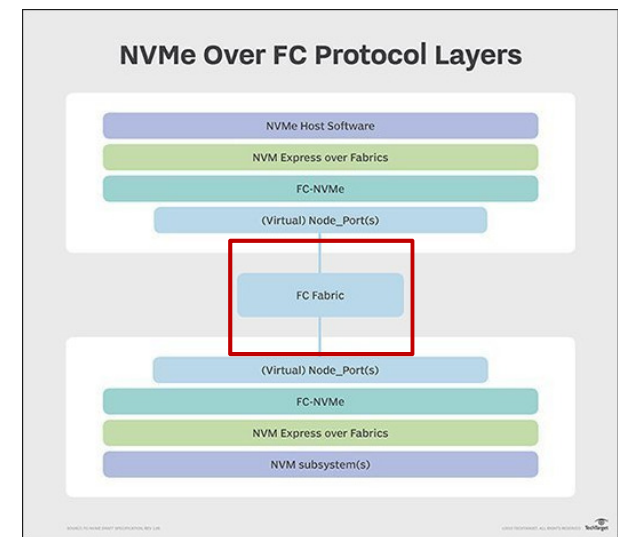 - **The ability to use a "neutral" oberver**

NVMe Over FC Protocol Layers

Image courtesy of TechTarget.com

Fibre Channel T&M Future

- **New Tools and Processes for PHY Testing**
  - **High Speed Real Time and/or Sampling Scopes**
    - Up to 100GHz Today!
- **Specialized Traffic Generation Capabilities Supporting**:
  - Physical Coding Sublayer – 8b/10b, 64b/66b, 256b/257b encoding
  - Speed-Negotiation, Transmitter Training Sequences
- **New Line Rate analysis capabilities supporting:**
  - "Pass Through" tapping
  - Bit-level Capture
- **The "Channel" must be "smart"**
  - Protocol is inherent in the physical layer!

Thank You!

Q&A

Fibrechannel.org

BIOs

Curt Beckman BIO

- Principal Architect – Broadcom Inc
- Author – NVMe over Fibre Channel Dummies Guide

Criag Carlson BIO

- Co-Chair INCITS T11
- Chair of T11 FC-NVMe working group
- FCIA Board of Directors
- NVM Express Board of Directors
- SNIA Technical Council

Rupin Mohan BIO

- *Rupin Mohan is a Director of R&D and CTO of Storage Networking (SAN) at HPE Storage. Rupin leads a global engineering and product management teams responsible for development of Storage Networking products. Rupin has filed for 30+ patents at HPE. He is a Board Member of and Marketing Chairman for FCIA. Rupin completed his MBA from MIT Sloan School of Management as a Sloan Fellow. He also holds a MS in Engineering from Tufts University and BE in Computer Engineering from Delhi Institute of Technology.*

Dennis Martin BIO

- Dennis Martin is the founder and President of Demartek, a computer industry analyst organization with its own ISO 17025 accredited test lab. Demartek focuses on lab validation testing and performance testing of storage and related hardware and software products and is recognized by the EPA as an official test lab for ENERGY STAR Data Center Storage. Dennis has been working in the Information Technology industry since 1980.

David Rodgers Bio

- David J. Rodgers is Senior Product Marketing Manager with Teledyne LeCroy Protocol Solutions Group,
- where he focuses on defining, designing, deploying, and marketing a broad range of high-speed serial
- analysis test and measurement products for Ethernet and Fibre Channel SANs and LANs.

- With over 30 years of leading-edge computer industry experience, Rodgers represents Teledyne LeCroy Test
- and Measurement solutions in various industry standards groups, including participation on the Board of
- Directors for the Ethernet Alliance and the Fibre Channel Industry Association, and in the IEEE and T11
- standards bodies, and involvement in the USB, PCIe and NVMe special interest groups. He began his career
- in the test and measurement space as an original member of the USB Implementers Forum and pioneer
- marketer of the first protocol specific analyzers for the USB specifications.