NVMe™ over Fabrics – Discussion on Transports

Sponsored by NVM Express® organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards

NVM Express® Sponsored Track for Flash Memory Summit 2018

	Track	Title	Speakers	
NVMe-101-1	8/7/18 8:30-9:35	NVM Express: NVM Express roadmaps and market data for NVMe, NVMe-oF, and NVMe-MI - what you need to know the next year.	Janene Ellefson, Micron J Metz, Cisco	Amber Huffman, Intel David Allen, Segate
	8/7/18 9:45-10:50	NVMe architectures for in Hyperscale Data Centers, Enterprise Data Centers, and in the Client and Laptop space.	Janene Ellefson, Micron Chris Peterson, Facebook	Andy Yang, Toshiba Jonmichael Hands, Intel
NVMe-102-1	3:40-4:45 8/7/18	NVMe Drivers and Software: This session will cover the software and drivers required for NVMe-MI, NVMe, NVMe-oF and support from the top operating systems.	Uma Parepalli, Cavium Austin Bolen, Dell EMC Myron Loewen, Intel Lee Prewitt, Microsoft	Suds Jain, VMware David Minturn, Intel James Harris, Intel
	4:55-6:00 8/7/18	NVMe-oF Transports: We will cover for NVMe over Fibre Channel, NVMe over RDMA, and NVMe over TCP.	Brandon Hoff, Emulex Fazil Osman, Broadcom J Metz, Cisco	Curt Beckmann, Brocade Praveen Midha, Marvell
NVMe-201-1	8/8/18 8:30-9:35	NVMe-oF Enterprise Arrays: NVMe-oF and NVMe is improving the performance of classic storage arrays, a multi-billion dollar market.	Brandon Hoff, Emulex Michael Peppers, NetApp Clod Barrera, IBM	Fred Night, NetApp Brent Yardley, IBM
	8/8/18 9:45-10:50	NVMe-oF Appliances: We will discuss solutions that deliver high-performance and low-latency NVMe storage to automated orchestration-managed clouds.	Jeremy Warner, Toshiba Manoj Wadekar, eBay Kamal Hyder, Toshiba	Nishant Lodha, Marvell Lior Gal, Excelero
NVMe-202-1	8/8/18 3:20-4:25	NVMe-oF JBOFs: Replacing DAS storage with Composable Infrastructure (disaggregated storage), based on JBOFs as the storage target.	Bryan Cowger, Kazan Networks	Praveen Midha, Marvell Fazil Osman, Broadcom
	8/8/18 4:40-5:45	Testing and Interoperability: This session will cover testing for Conformance, Interoperability, Resilience/error injection testing to ensure interoperable solutions base on NVM Express solutions.	Brandon Hoff, Emulex Tim Sheehan, IOL Mark Jones, FCIA	Jason Rusch, Viavi Nick Kriczky, Teledyne

Speakers

Brandon Hoff

Curt Beckmann

Fazil Osman

Praveen Midha

J Metz

Abstract and Agenda

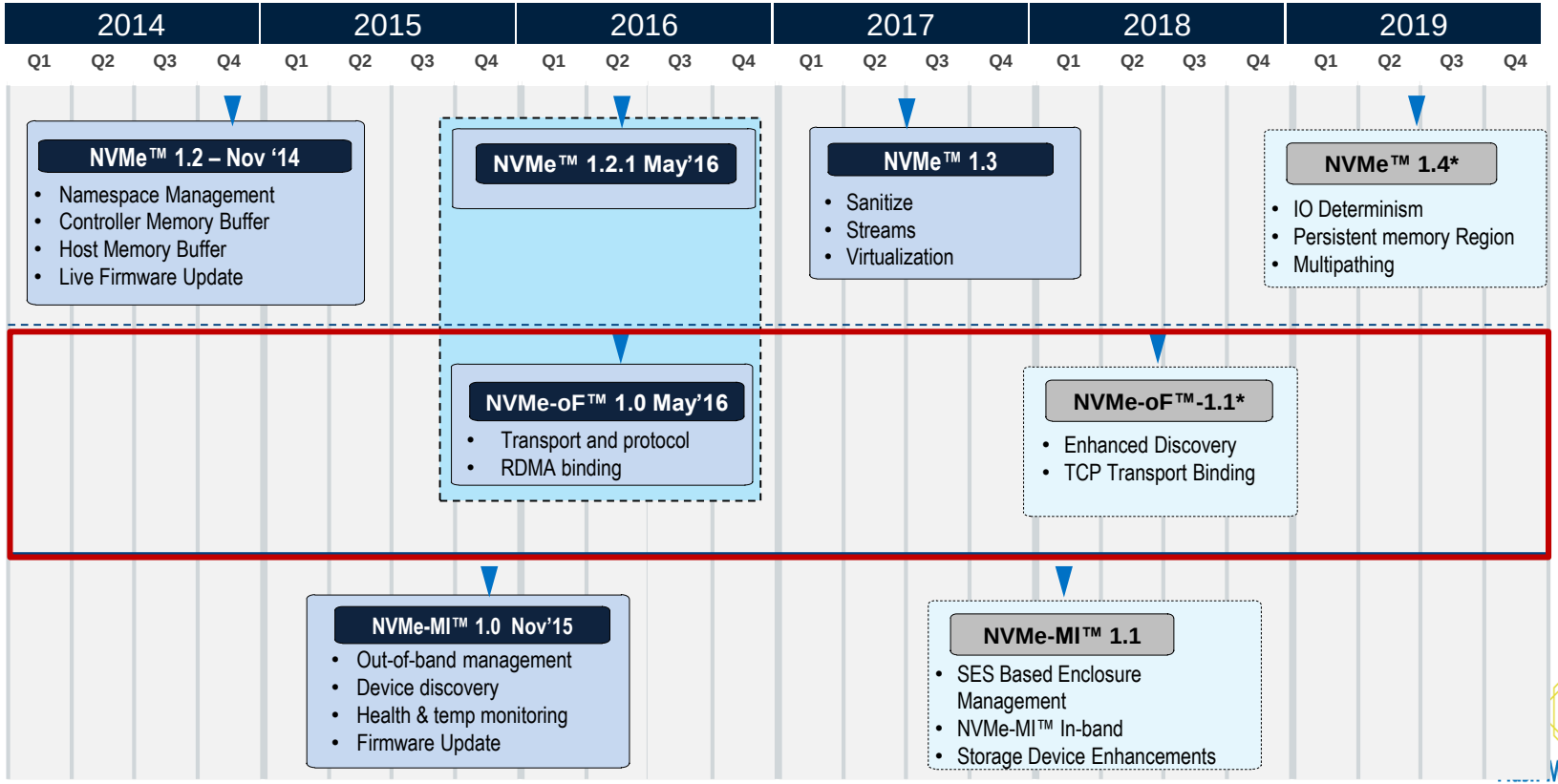
- NVMe-oF® Abstract:
  - NVMe™ over Fabrics is designed to be transport agnostic, with all transports being created equal from the perspective of NVM Express. We will cover for NVMe over Fibre Channel, NVMe over RDMA, and NVMe over TCP.

- NVMe-oF Panel
  - NVMe-oF Overview and Scope of our Panel – Brandon Hoff, Emulex (10 min)
  - NVMe over Fibre Channel (NVMe/FC) – Curt Beckmann, Brocade (10 min)
  - NVMe over RoCE (NVMe/RoCE) – Fazil Osman, Broadcom Classic (10 min)
  - NVMe over iWARP (NVMe/iWARP) – Praveen Midha, Marvell/Qlogic (10 min)
  - NVMe over TCP (NVMe/TCP) – J Metz, Cisco (10 min)
  - Q&A (15min)

NVMe™ over Fabrics

Brandon Hoff, Principle Architect, Emulex

NVMe™ Feature Roadmap



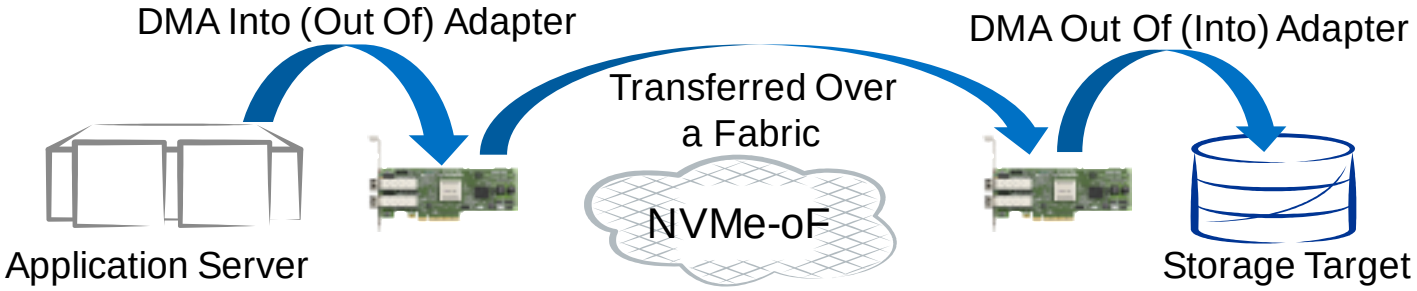
Released NVMe™ specification Planned release

* Subject to change

The Value of Shared Storage and the ‘need for speed’

- The cost of data-at-rest is no longer the right metric for storage TCO
  - The value of data is based on how fast it can be accessed and processed
- NVMe™ over Fabrics increases the velocity of data
  - Faster storage access enables cost reduction through consolidation
  - Faster storage access delivers more value from data
- SSDs are going to become much faster
  - 3D Xpoint Memory, 3D NAND, etc.
  - PMEM, Storage Class Memory, etc.
  - … and innovation will continue

Simplicity of NVMe™ over Fabrics



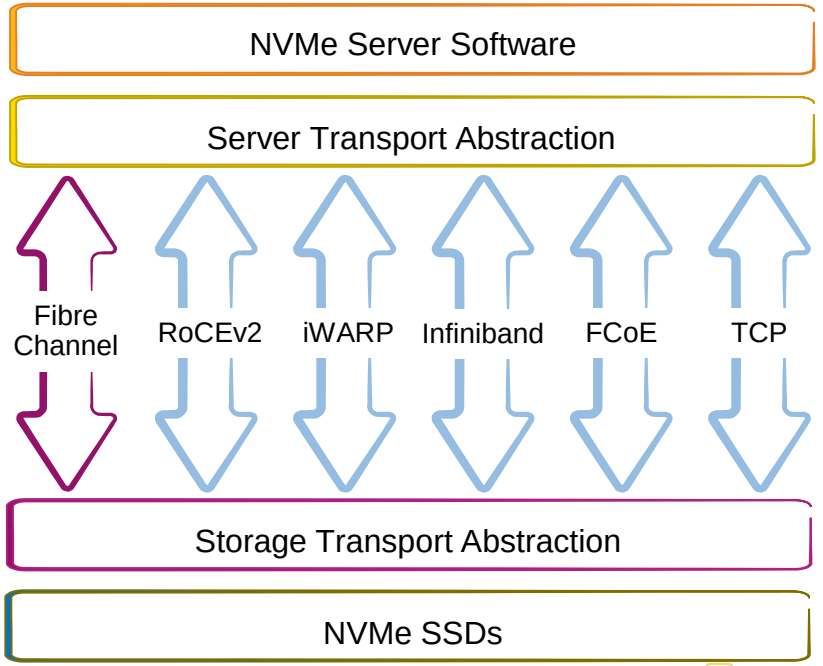
- NVMe-oF™ delivers a new level of performance for today's business-critical applications
- NVMe-oF is, by design, is transport agnostic:
  - Application developers can write to a single block storage stack and access NVMe over Fibre Channel, TCP, or RDMA networks
- Data is DMA'd in and out of the adapters to maximize performance
  - Zero copy is available **today** for Fibre Channel and RDMA protocols for improved performance and there are solutions that can provide zero copy for TCP

Scaling NVMe™ Requires a (Real) Network

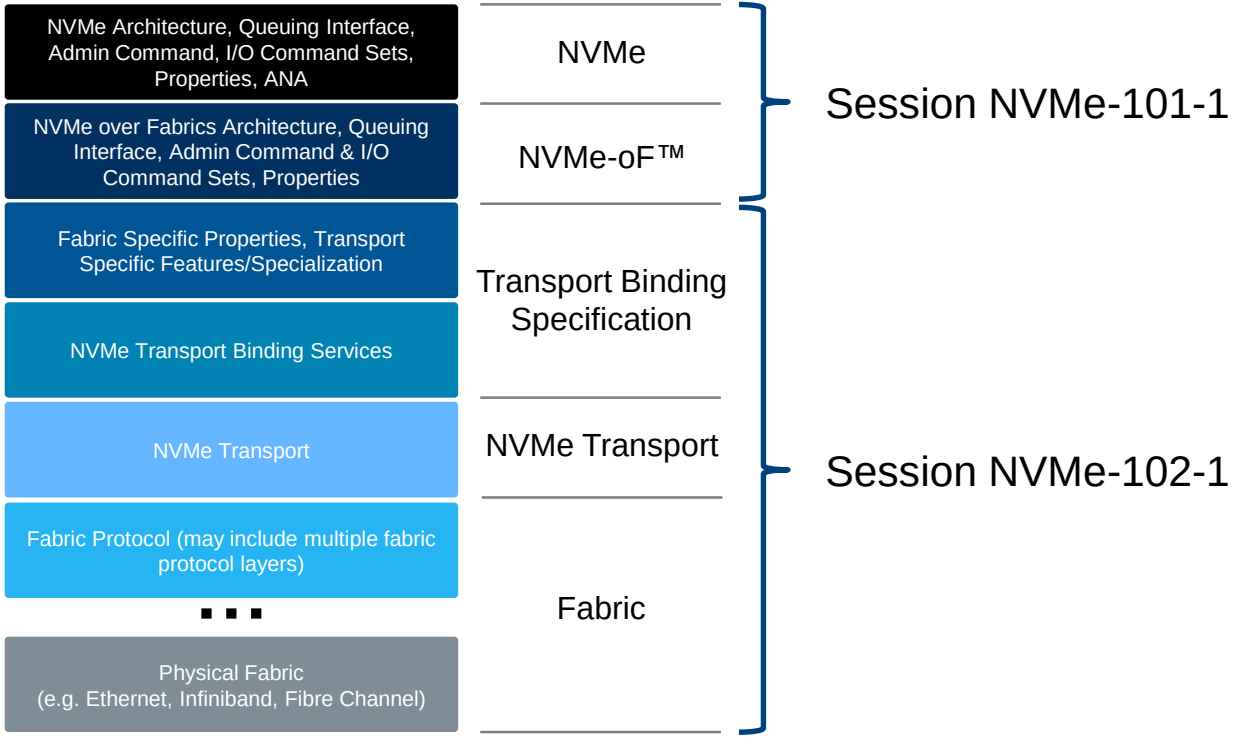
- Many options, plenty of confusion
- Fibre Channel is the transport for the vast majority of today's all flash arrays

  FC-NVMe Standardized in Mid-2017
- RoCEv2, iWARP and InfiniBand are RDMA based but not compatible with each other

  NVMe-oF™ RDMA Standardized in 2016
- FCoE as a fabric is an option, leverages the FC stack integrated into NVMe-oF 1.0
- NVMe/TCP - making it way through the standards



NVMe™ over Fabrics - Architecture



NVMe™ over Fibre Channel

Curt Beckmann, Principal Architect, Brocade

Presentation Topics

- FC-NVMe Spec and Interoperability Update
- Dual Protocol SANs boost NVMe™ adoption
- Performance audit: NVMe/FC v SCSI/FC

FC-NVMe Spec Status

- Why move to NVMe™/FC?
  - It's like SCSI/FC tuned for SSDs and parallelism
  - Simpler, more efficient, and (as we'll see) faster
- FC-NVMe standard effort is overseen by T11
  - T11 and INCITS finalized FC/NVMe early 2018
- Several vendors are shipping GA products
- FCIA plugfest last week: 13 participating companies

Presentation Topics

- FC-NVMe Spec and Interoperability Update
- Dual Protocol SANs boost NVMe adoption
- Performance audit: NVMe/FC v SCSI/FC

Dual Protocol SANs boost NVMe™ adoption

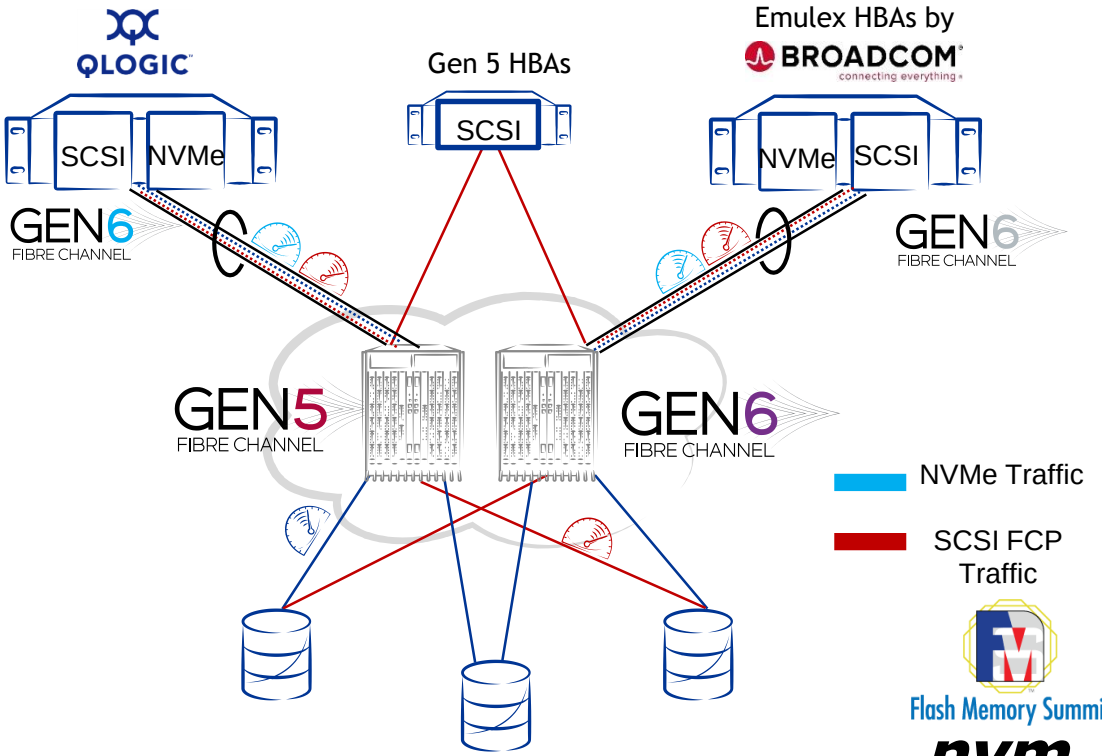
- 80% of today's Flash arrays connect via FC
  - This is where vital data assets live
- High-value Assets require protection
  - Storage Teams are naturally risk averse
  - Risk avoidance is part of the job description
- How can Storage Teams adopt NVMe with low risk?
  - Use familiar infrastructure that speaks both old and new!

Dual Protocol SANs Reduce Risk

- Uses existing, familiar, trusted infrastructure
  - No surprises, no duplication of infrastructure and effort
- Rely on known, established vendors
  - With shared vocabulary and trusted support models
- Continue to use robust FC Fabric Services
  - Name services, discovery, zoning, flow control
- Leverage familiar tools and team expertise
  - No need to start all over from scratch

Dual protocol SANs enable low risk NVMe™ adoption

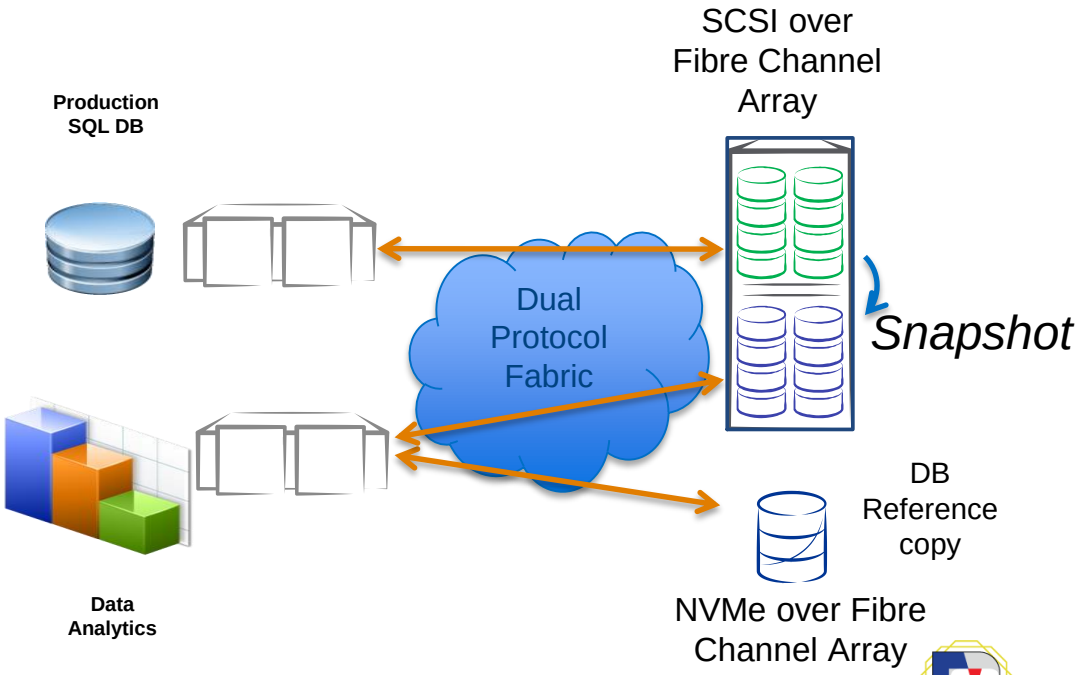
- Get NVMe performance benefits while migrating incrementally “as-needed”
- Migrate application volumes 1 by 1 with easy rollback options
- Interesting dual-protocol use cases
- Full fabric awareness, visibility and manageability with existing Brocade Fabric Vision technology



Sample Use Case: Extract Value from High Value Data Assets

Staged Analytics on Real-World Data Sets

- Using near-live data for analytics is gaining popularity as a way to extract more value
  - But adding traffic loads to live data can impact its performance
- Instead, snapshot data on existing Enterprise Storage
  - Clone the snapshot to NVMe™ NSID
  - Run high performance analytics on the same infrastructure
- Works in many dimensions
  - High performance analytics
  - Easy to operationalize
  - Leverages current infrastructure



Presentation Topics

- FC-NVMe Spec and Interoperability Update
- Dual Protocol SANs boost NVMe adoption
- Performance audit: NVMe/FC v SCSI/FC

Summary of Demartek Report

Purpose: Credibly document performance benefit of NVMe™ over Fibre Channel (NVMe/FC) is relative to SCSI FCP on vendor target

Audited by: Demartek

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

Audit Date: May 1, 2018

- PDF available at: www.demartek.com/ModernSAN

Results of testing both protocols on same hardware:

- Up to 58% higher IOPS for NVMe/FC
- From 11% to 34% lower latency with NVMe/FC

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.

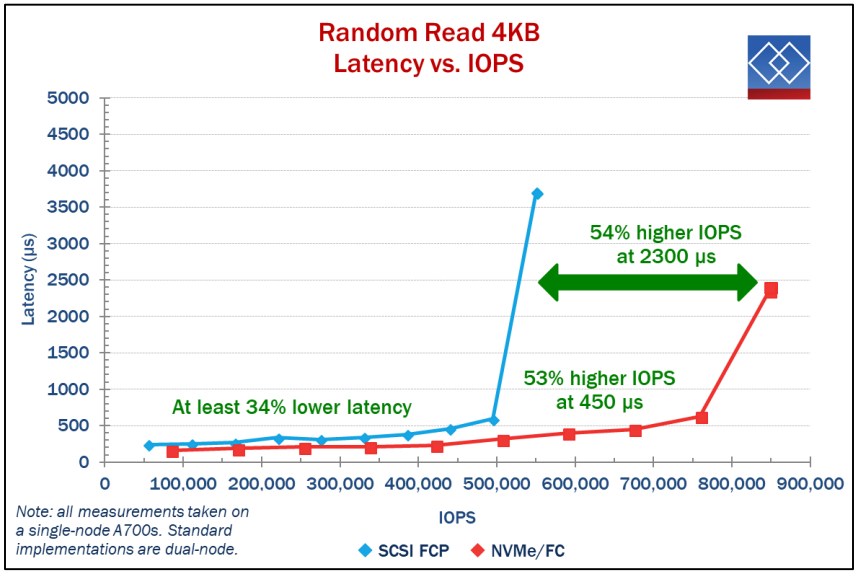
Key Findings and Conclusions

- ***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.

Demartek demartek.com © 2018 Demartek

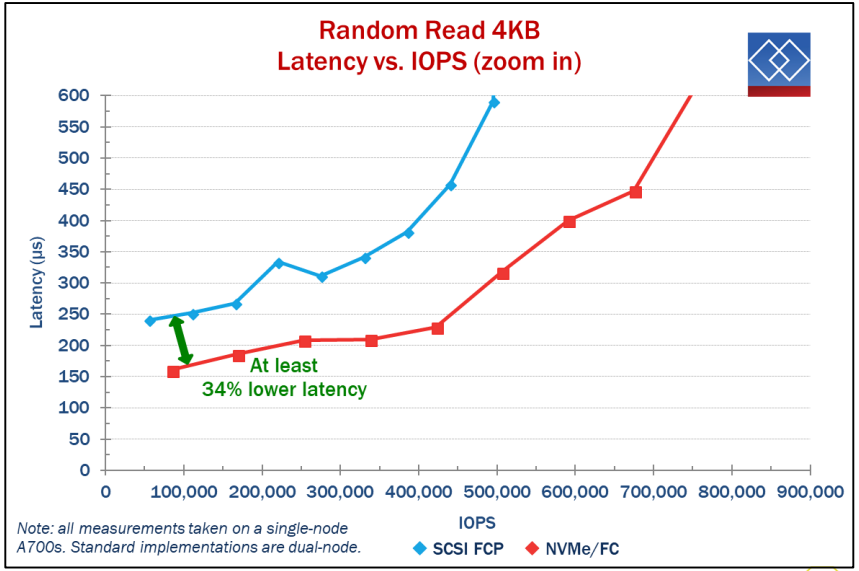
Note: The audit was *not* intended as a test of max array performance

Results: 4KB Random Reads, full scale and zoomed in



This image highlights how NVMe/FC gives **53%** / **54%** higher IOPS with 4KB random read I/Os

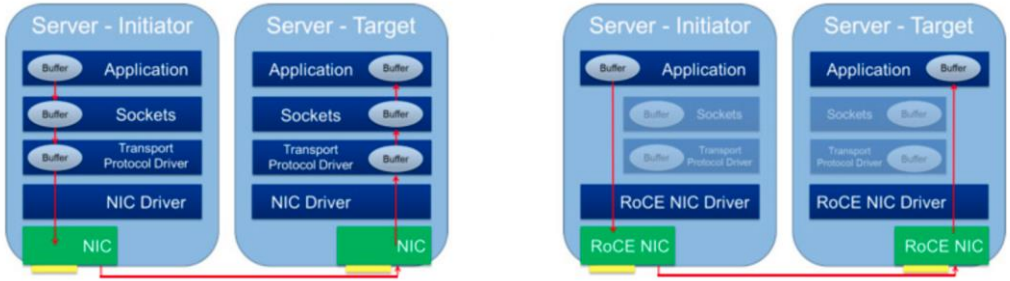
Same data with y-axis expanded to see that NVMe™/FC provides a minimum **34%** drop in latency



NVMe™ over RoCE

Fazil Osman, Broadcom Classic

What is RoCE?



Remote Direct Memory Access (RDMA)

Hardware offload moves data from memory on one CPU to memory of a second CPU without any CPU intervention

RDMA over Converged Ethernet (RoCE)

Runs over standard Ethernet (L2 or L3 network with RoCEv1 or RoCEv2) with very low latencies

Standard Protocol with Multivendor Support

Defined by IBTA

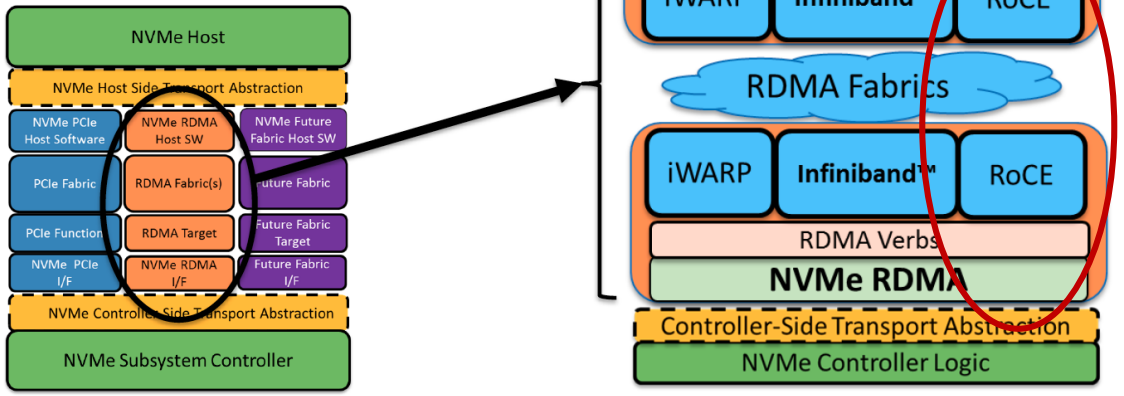
Support from leading NIC vendors – Broadcom, Marvell, Mellanox

Proven Interoperability at UNH and customer deployments

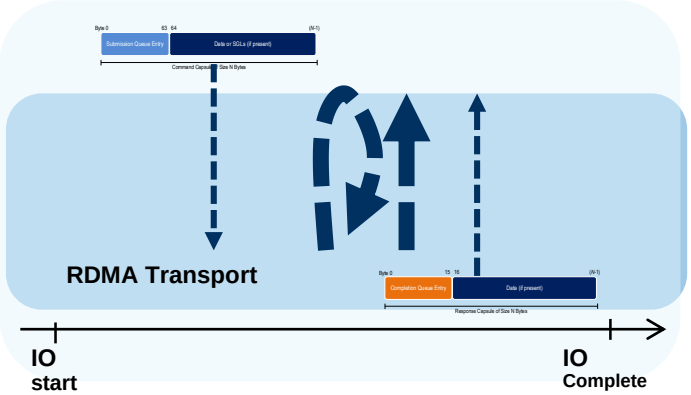
Where RoCE fits in with NVMe-oF™

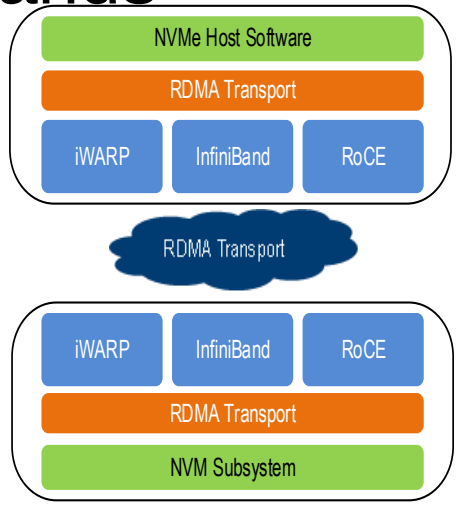
NVMe over RDMA Fabric

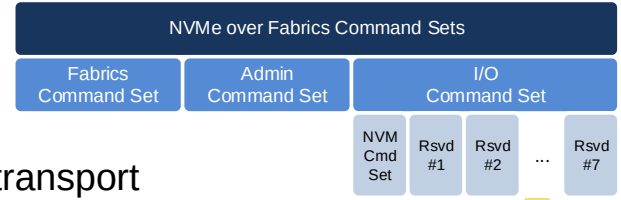
- Upper Level RDMA Block Storage Protocol
- Layered over a common set of RDMA Verbs
- Imperative to support all RDMA provider types
  - Infiniband™
  - Ethernet (iWARP and RoCE)



NVMe™ over RoCe IO Model & Commands







- NVMe commands are encapsulated and sent seamlessly over RoCE transport
- NVMe multi-Queue model
- Fabrics commands may be submitted on the Admin Queue or submitted on an I/O Queue
- Processing requires minimal intervention of target CPUs

NVMe™ over RoCE Advantages

- Ethernet is the converged protocol for the Data Center
  - RoCE is supported by the leading NIC vendors
    - 80% of shipped RNIC 25G+ ports in Q1’18 only support RoCE (Crehan)
  - Proven interoperability at UNH and customer deployments
- RoCE is the lowest latency protocol
  - Sub 5us typical End to End
- Very low CPU utilization when running RoCE
  - Bypasses TCP transport greatly reducing CPU overhead

NVMe-oF™ Transports - iWARP

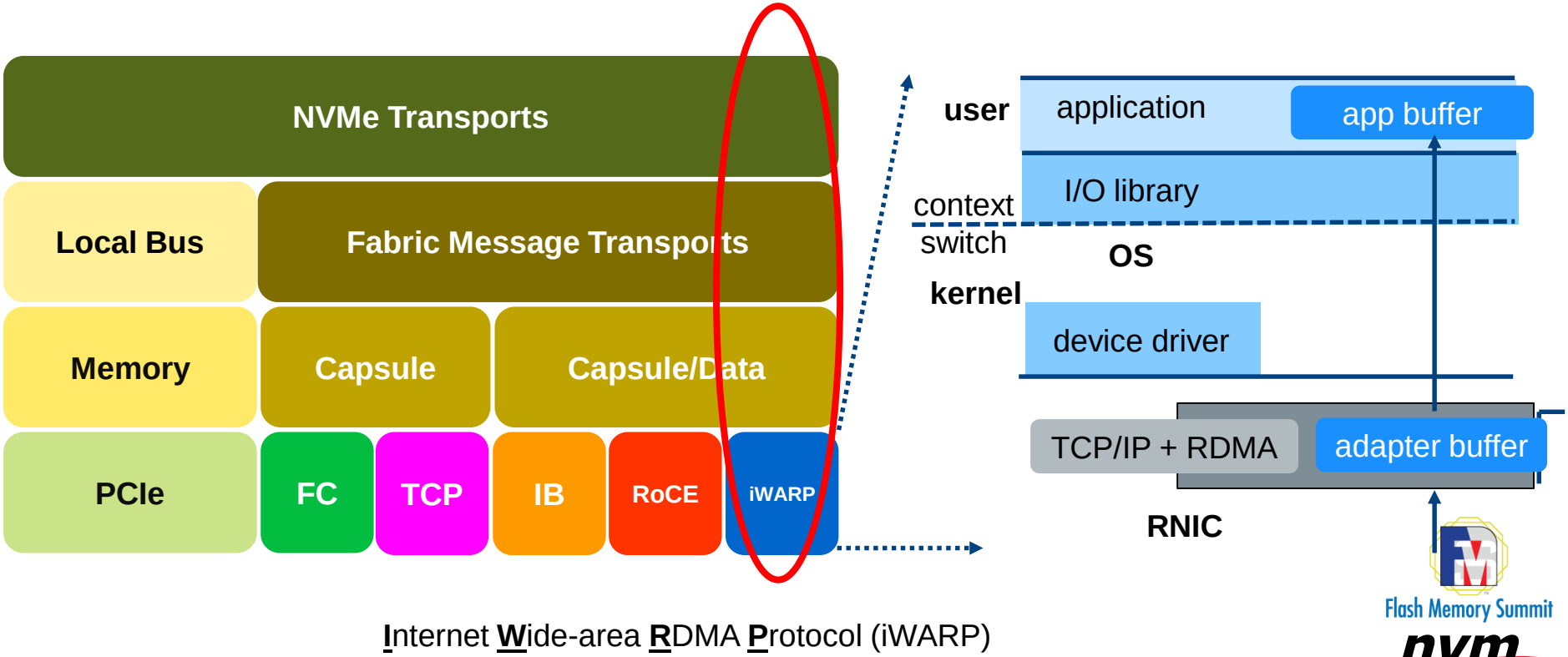
Praveen Midha, Marvell Technologies

Agenda

- What is iWARP?
- Why should I care about iWARP?
- How does iWARP perform?
- Any real world use cases?
- Summary

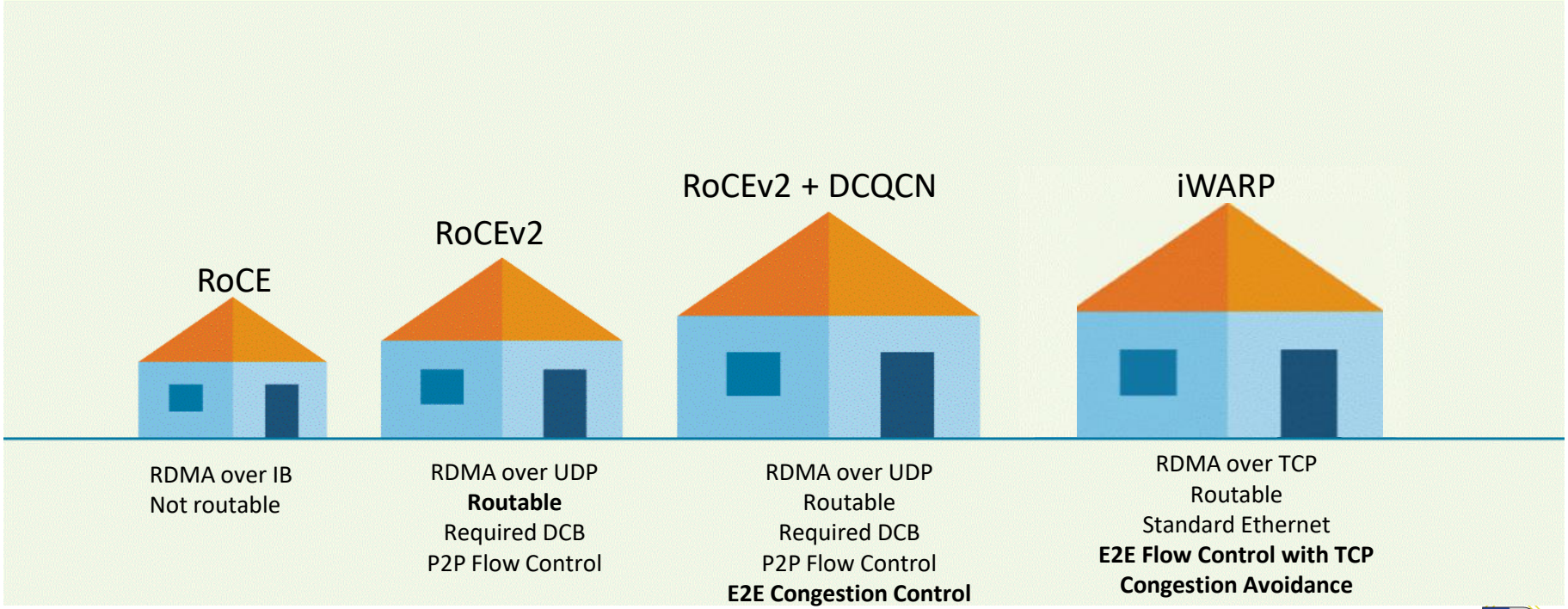
NVMe-oF™ Transport Choices…

Internet Wide-area RDMA Protocol (iWARP)



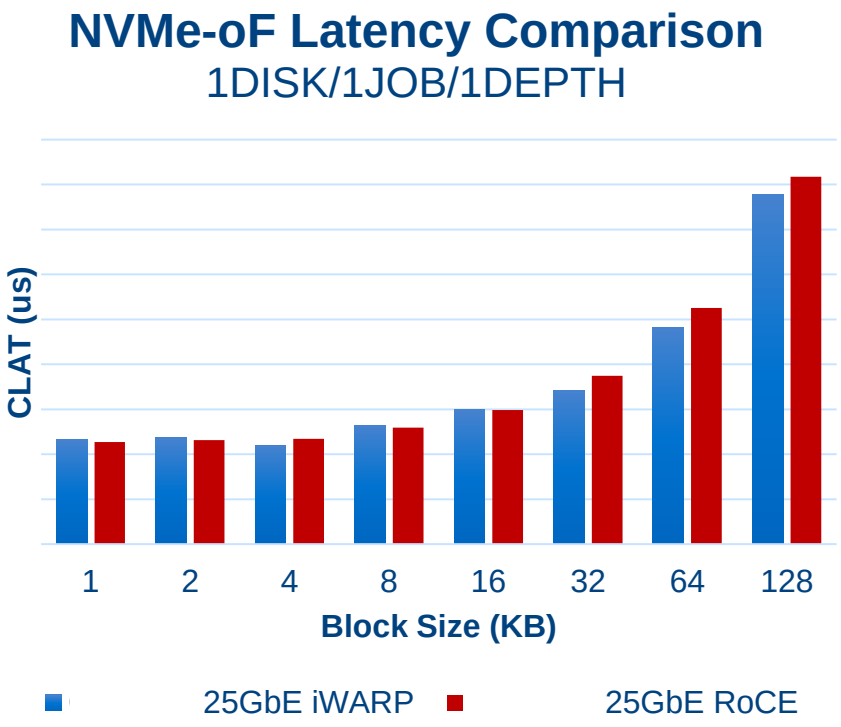
Internet Wide-area RDMA Protocol (iWARP)

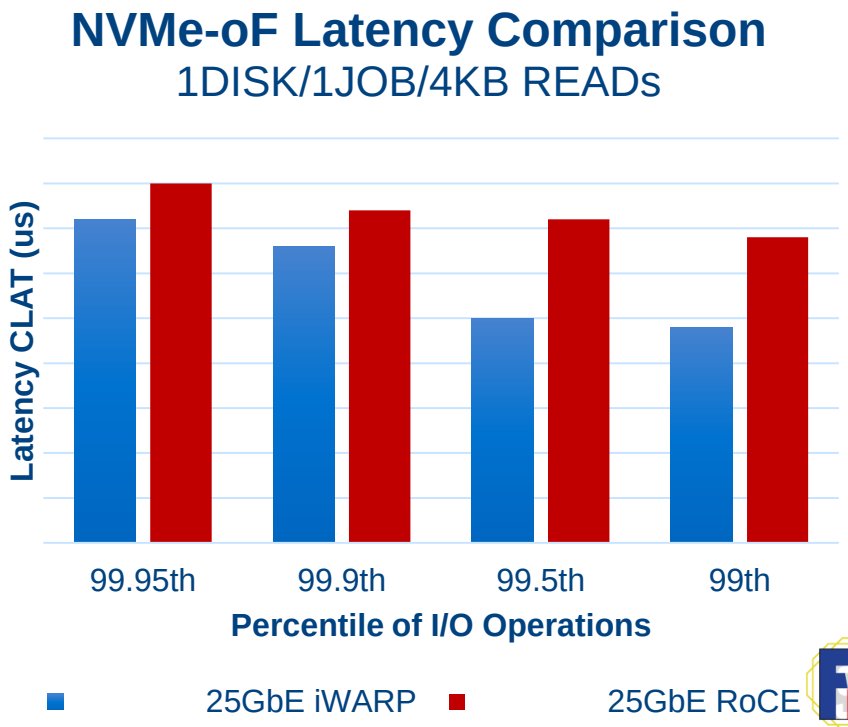
RDMA Scalability Comparison



Internet **W**ide-area **R**DMA **P**rotocol (iWARP)

NVMe-oF™ Latency – Single I/O



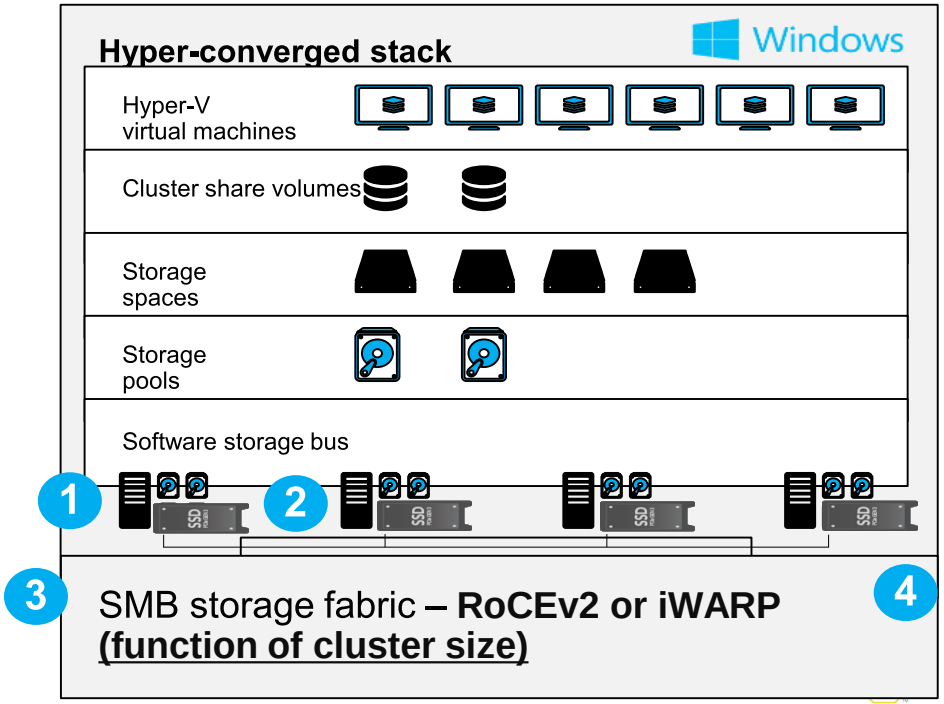


Internet **W**ide-area **R**DMA **P**rotocol (iWARP)

Storage Spaces Direct (S2D) – Hyper-Converged

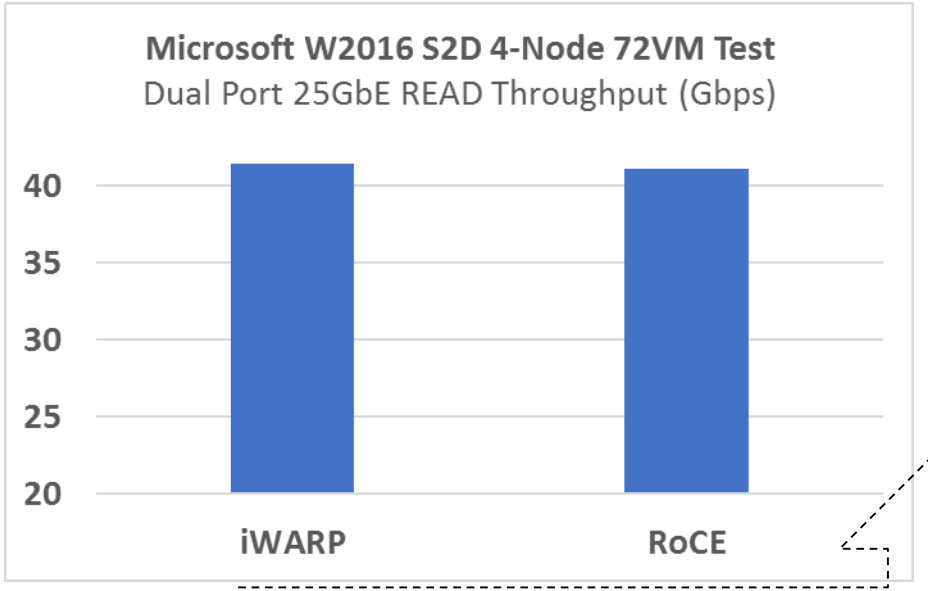
Hyper-Converged storage **and** compute with Storage Spaces Direct

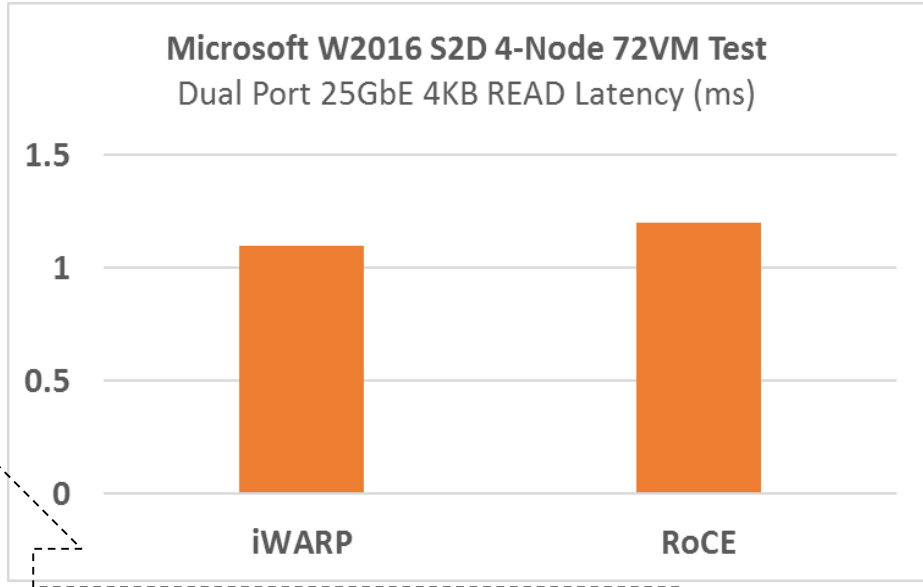
1. Industry standard x86 servers, with **local** SSD/NVMe™ and HDD. Servers are connected together with 10/25GbE.
2. Compute and storage resources scale and are managed together. Typically small to medium sized scale-out deployments.
3. RoCE/RoCEv2 based RDMA with lossless Ethernet network enables a low latency SMB Storage Fabric – ideal for overprovisioned or well managed network infrastructure
4. iWARP based RDMA with standard Ethernet enables a scalable, low touch SMB Fabric – ideal for large scale or congestion prone network infrastructures



Internet **W**ide-area **R**DMA **P**rotocol (iWARP)

S2D Performance – iWARP vs RoCE





Storage at Microsoft

The official blog of the Windows and Windows Server storage engineering teams

Storage Spaces Direct with Cavium FastLinQ® 41000

September 21, 2017 by clausjor // 1 Comments

https://blogs.technet.microsoft.com/filecab/2017/09/21/storage-spaces-direct-with-cavium-fastlinq-41000/

Summary - iWARP

is one of several transport choices for deploying NVMe-oF™

Wide Area Networks supported

Assumes standard Ethernet – no DCB!

Reliable connected communication provided by congestion-aware TCP protocol

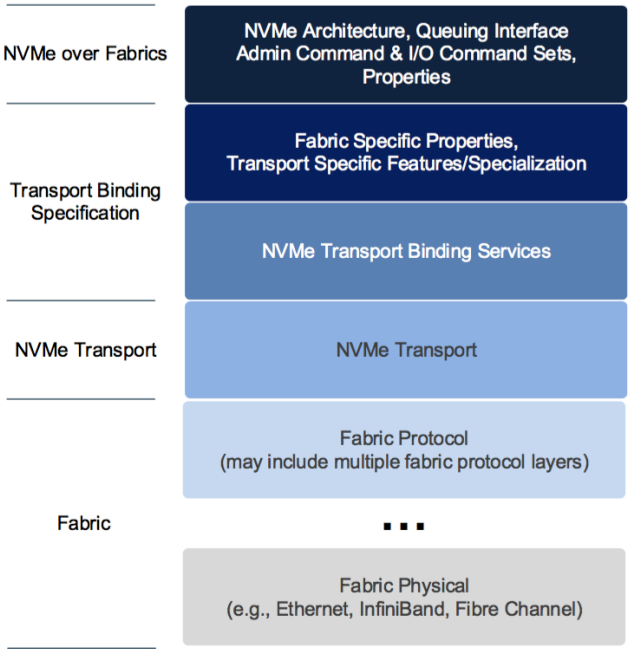
Performs as well as RoCE/RoCEv2

Internet **W**ide-area **R**DMA **P**rotocol (iWARP)

NVMe™ over TCP

J Metz, Cisco

What's Special About NVMe-oF™: Bindings



What is a Binding?

- "A specification of reliable delivery of data, commands, and responses between a host and an NVM subsystem for an NVMe™ Transport. The binding may exclude or restrict functionality based on the NVMe Transport's capabilities."

I.e., it's the "glue" that links all the pieces above and below (examples):

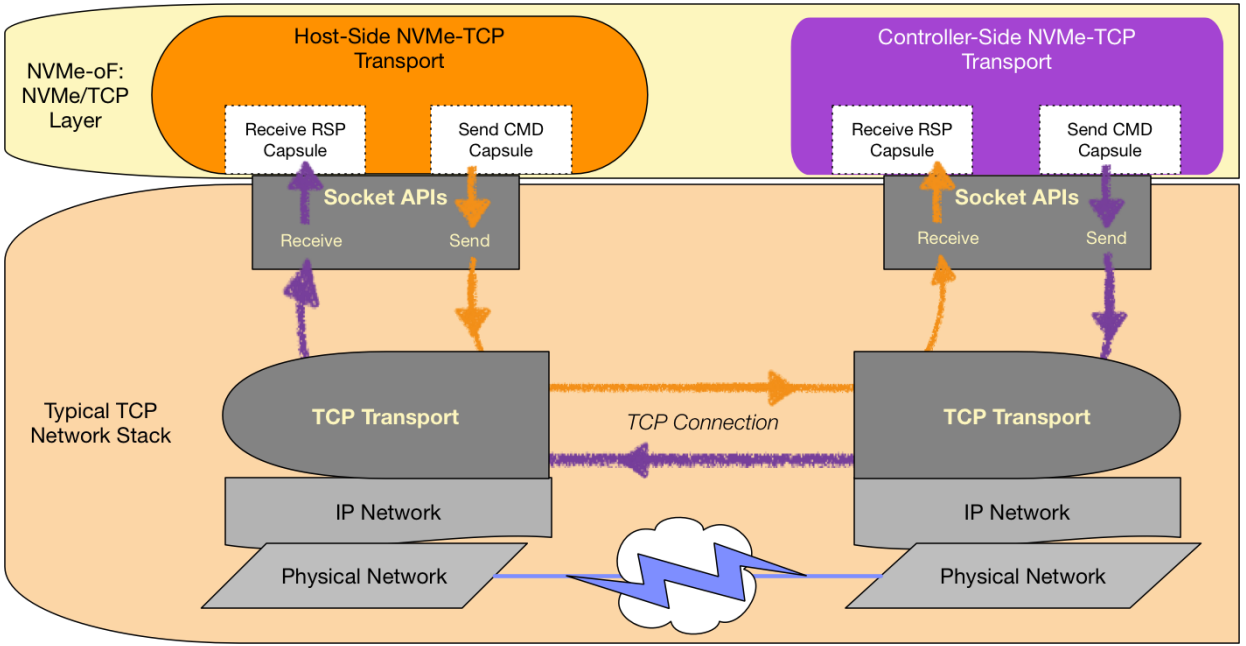
- SGL Descriptions
- Data placement restrictions
- Data transport capabilities
- Authentication capabilities

NVMe™/TCP in a Nutshell

NVMe-oF™ commands sent over standard TCP/IP sockets

Each NVMe queue pair mapped to a TCP connection

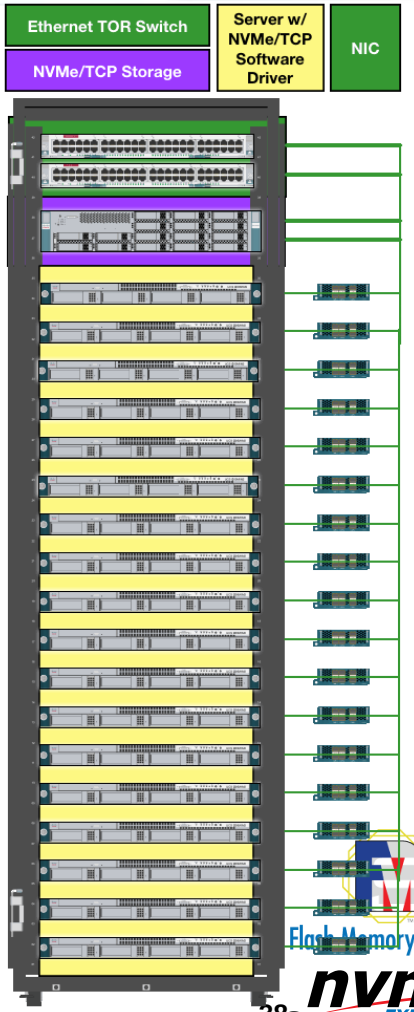
TCP provides a reliable transport layer for NVMe queueing model



NVMe™/TCP Data Path Usage

Enables NVMe-oF™ I/O operations in existing IP Datacenter environments

- Software-only NVMe Host Driver with NVMe-TCP transport
- Provides an NVMe-oF alternative to iSCSI for Storage Systems with PCIe NVMe SSDs
  - More efficient End-to-End NVMe Operations by elimination SCSI to NVMe translations
  - Co-exists with other NVMe-oF transports
    - Transport selection may be nased on h/w support and/or policy



NVMe™/TCP Control Path Usage

Enables use of NVMe-oF™ on Control-Path Networks (example: 1g Ethernet)

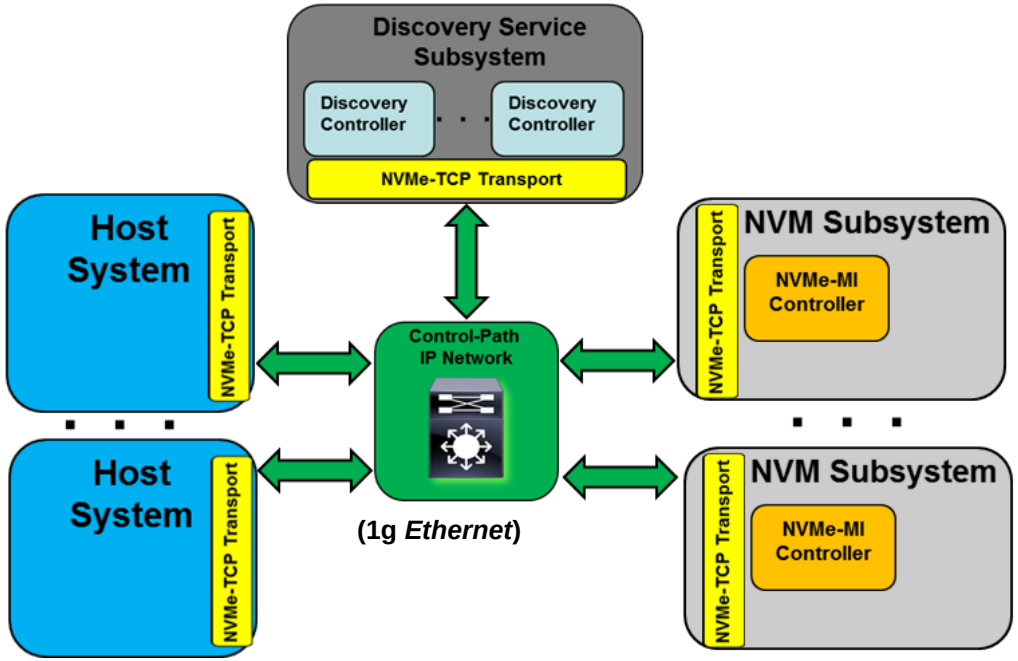
Discovery Service Usage

Discovery controllers residing on a common control network that is separate from data-path networks

NVMe-MI™ Usage

NVMe-MI endpoints on control processors (BMC, ..) with simple IP network stacks

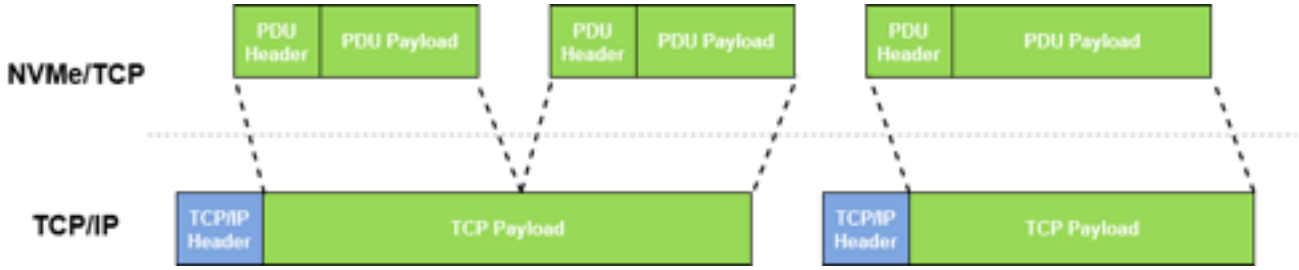
NVMe-MI on separate control network



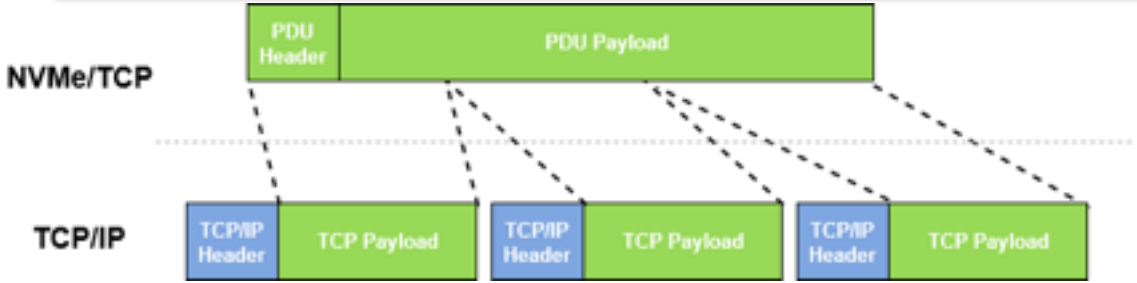
Source: Dave Minturn (Intel)

How NVMe™/TCP Works

Multiple NVMe/TCP data units in a single TCP/IP packet

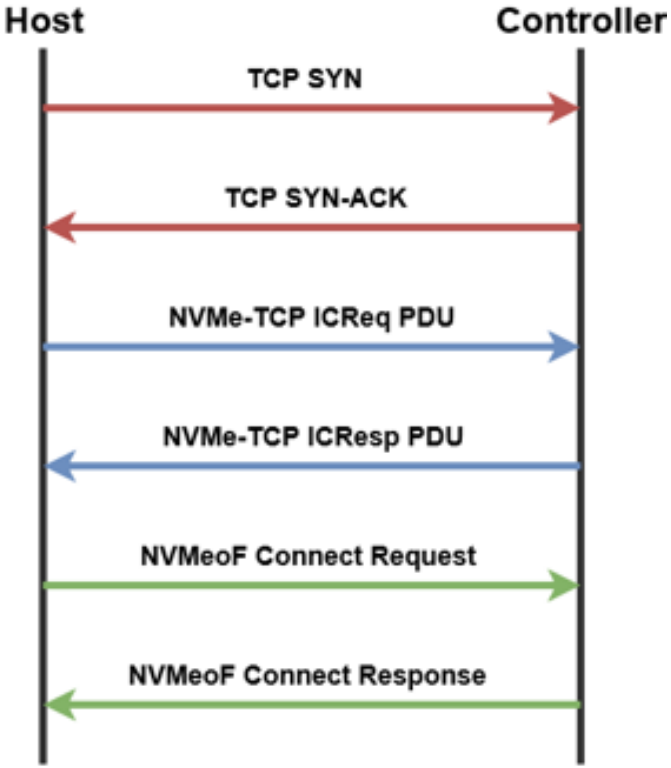


- TCP accepts data in the form of a data stream and breaks the stream into units
- A TCP header is added to a unit creating a TCP segment
- A segment is then encapsulated in an Internet Protocol (IP) datagram creating a TCP/IP packet



Single NVMe/TCP data unit spanning multiple TCP/IP packets

NVMe™/TCP Message Model



NVMe/TCP connection is associated with a single Admin or I/O SQ/CQ pair

- No spanning across queues or across TCP connections!

Data transfers supported by:

- Fabric-specific data transfer mechanism
- In-Capsule data (optional)
  - Allows for variable capsule sizes

All NVMe/TCP implementations support data transfers using command data buffers

Potential Issues With NVMe™/TCP

Absolute latency higher than RDMA?

Head-of-line blocking leading to increased latency?

Delayed acks could increase latency?

Incast could be an issue?

Lack of hardware acceleration?

Only matters if the application cares about latency

Protocol breaks up large transfers

Acks are used to 'pace the transmission of packets such that TCP is "self-clocking"

Switching network can provide Approximate Fair Drop (AFD) for active switching queue mgmt, and Dynamic Packet Prioritization (DPP) to ensure incast flows are serviced as fast as possible

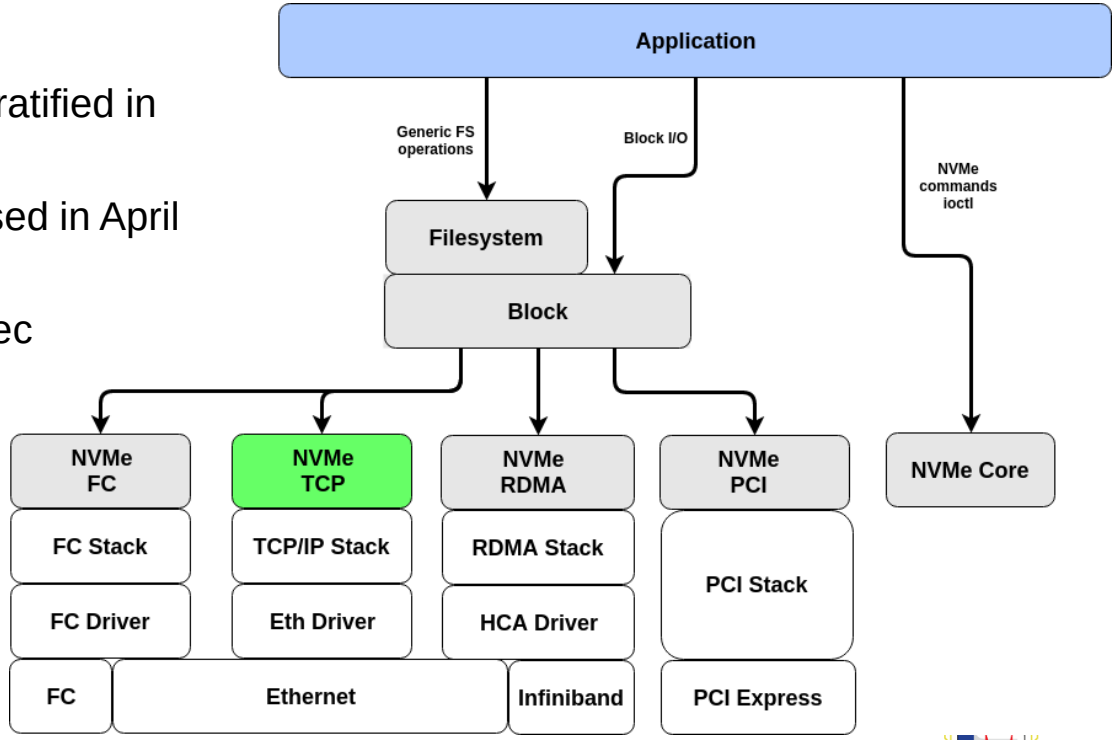
Not an issue for NVMe/TCP use-cases

NVMe™/TCP Standardization

Expect NVMe over TCP standard to be ratified in 2H 2018

The NVMe-oF™ 1.1 TCP ballot passed in April 2017

NVMe Workgroup adding TCP to spec alongside RDMA



Contact Information

For more information please contact the following:

Brandon Hoff	brandon.hoff@broadcom.com	
Curt Beckmann	curt.beckmann@broadcom.com	
Fazil Osman	fazil.osman@broadcom.com	
Praveen Midha	Praveen.Midha@cavium.com	
Fazil Osman	fazil.osman@broadcom.com	
J Metz	jmmetz@cisco.com	@drjmetz


nvm
EXPRESS®
Architected for Performance