How Ethernet RDMA Protocols iWARP and RoCE Support NVMe over Fabrics

John Kim, Mellanox
David Fair, Intel
January 26, 2016

SNIA™
Ethernet Storage Forum

Who We Are

John F. Kim
Director, Storage Marketing
Mellanox Technologies

David Fair
Chair, SNIA-ESF
Ethernet Networking Mktg Mgr
Intel Corp.

SNIA Legal Notice

SNIA
Ethernet Storage Forum

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

- How RDMA fabrics fit into NVMe over Fabrics
- RDMA explained and how it benefits NVMe/F
- Verbs, the *lingua franca* of RDMA
- Varieties of Ethernet RDMA explained
- Deployment considerations for RDMA-enhanced Ethernet

How RDMA Fabrics Fit Into NVMe over Fabrics

SNIA™
Ethernet Storage Forum

We Are Not Covering NVMe Over Fabrics Here Today

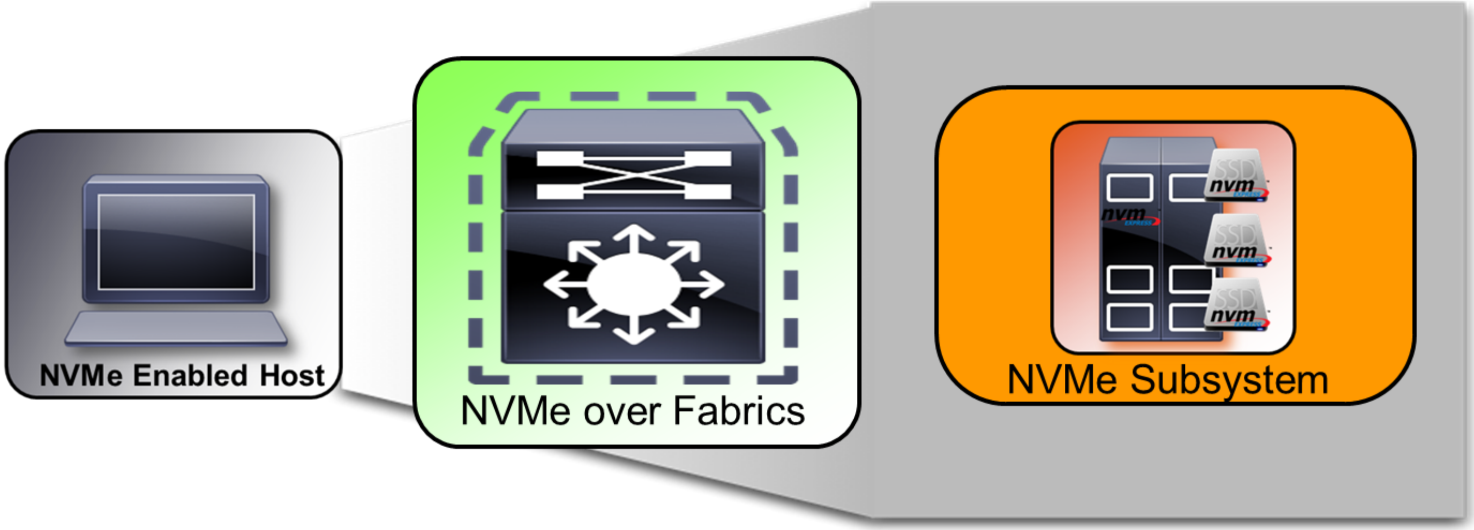
- For a comprehensive introduction to NVMe/Fabrics, please watch the SNIA-ESF webcast "*Under the Hood with NVMe over Fabrics*" produced December 2015 by J Metz (Cisco) and Dave Minturn (Intel)

- Posted on the SNIA-ESF website under "Webcasts On Demand": http://www.snia.org/forums/esf/knowledge/webcasts

- We are focusing on how RDMA fits into NVMe/Fabrics
  - A detailed understanding of the NVMe/F spec is not required

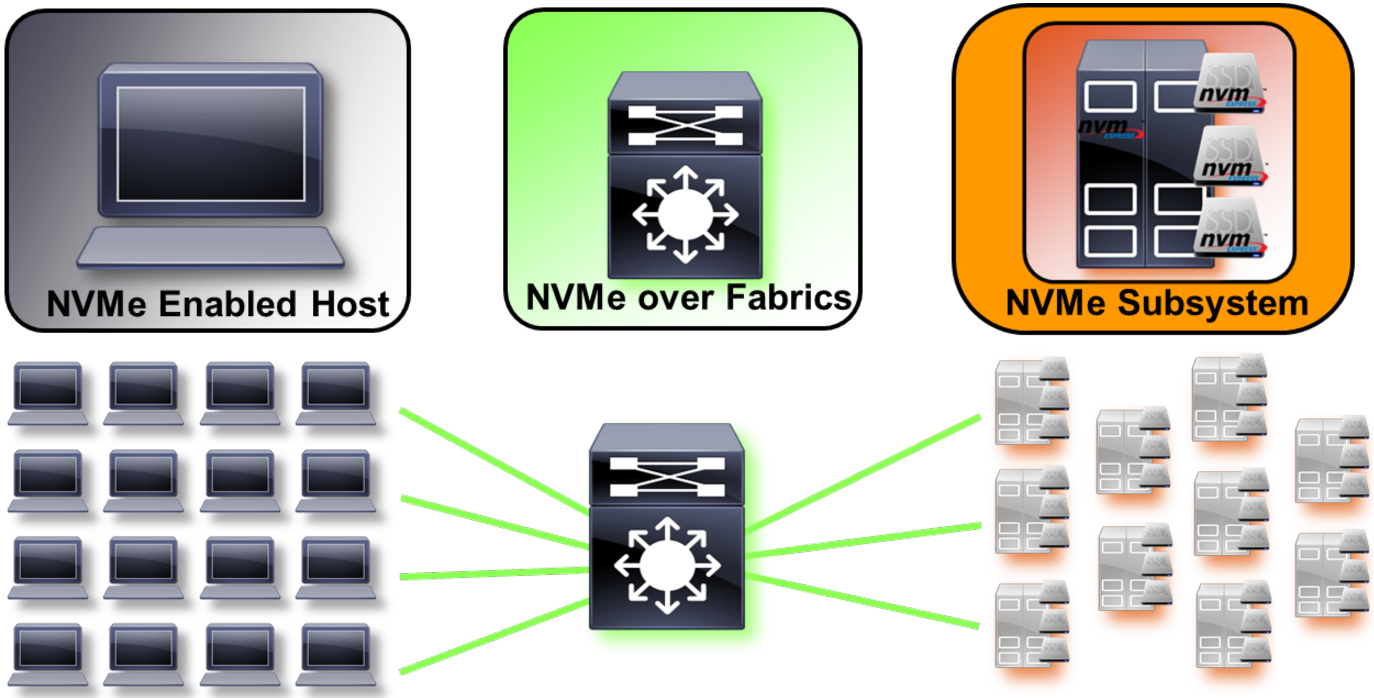
That Said, NVMe/F Expands NVMe to Fabrics

- Adds message-based NVMe operations
  - Leverages common NVMe architecture with additional definitions
  - Allows remote and shared access to NVMe subsystems
- Standardization of NVMe over a range Fabric types
  - Initial fabrics: RDMA (RoCE, iWARP, InfiniBand™) and Fibre Channel
  - First release candidate specification in early 2016
  - NVMe.org Fabrics WG developing Linux host and target drivers

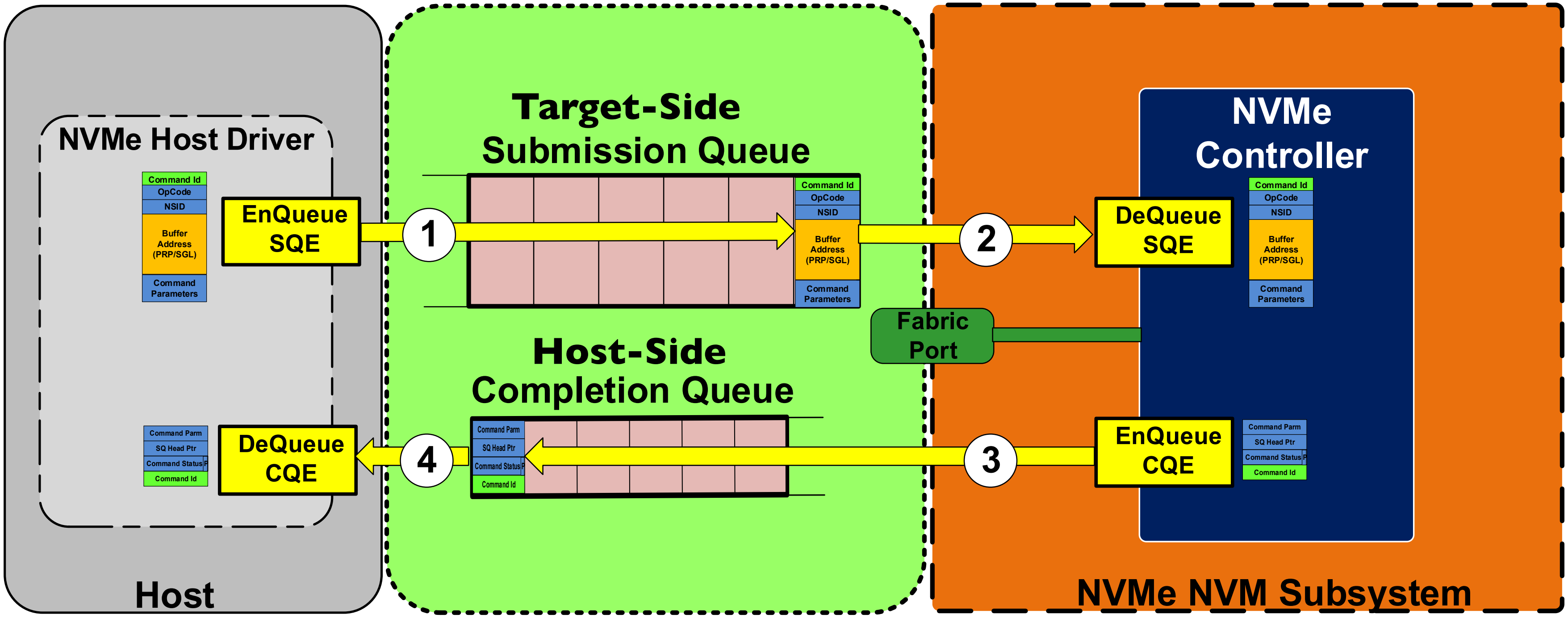


Why NVMe Over Fabrics

- End-to-End NVMe semantics across a range of topologies
  - Retains NVMe efficiency and performance over network fabrics
  - Eliminates unnecessary protocol translations (e.g. SCSI)
  - Enables low-latency and high IOPS remote NVMe storage solutions

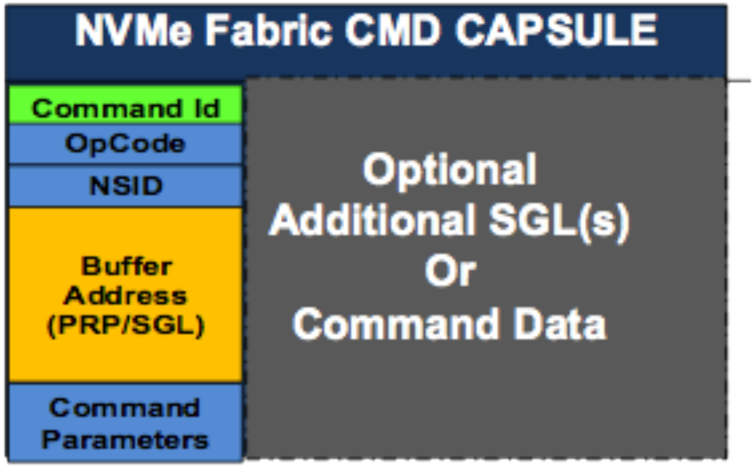


NVMe Queuing Operational Model



- 1. Host Driver enqueues the Submission Queue Entries into the SQ
- 2. NVMe Controller dequeues Submission Queue Entries
- 3. NVMe Controller enqueues Completion Queue Entries into the CQ
- 4. Host Driver dequeues Completion Queue Entries

NVMe Over Fabrics Capsules



NVMe Fabric CMD CAPSULE

Command Id

OpCode

NSID

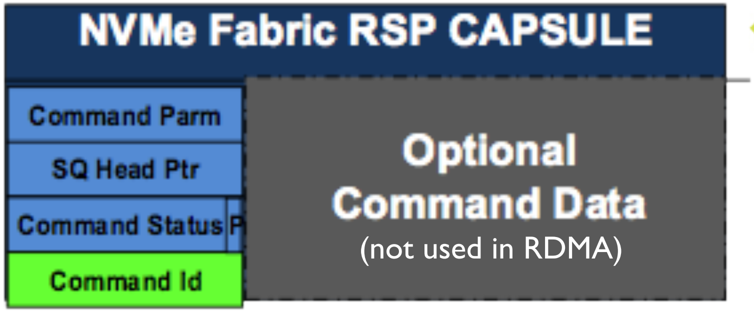
Buffer Address (PRP/SGL)

Command Parameters

Optional Additional SGL(s) Or Command Data

NVMe over Fabric Command Capsule

- **Encapsulated** NVMe SQE Entry
- May contain additional Scatter Gather Lists (SGL) or NVMe Command Data
- Transport agnostic Capsule format



NVMe Fabric RSP CAPSULE

Command Parm

SQ Head Ptr

Command Status P

Command Id

Optional Command Data

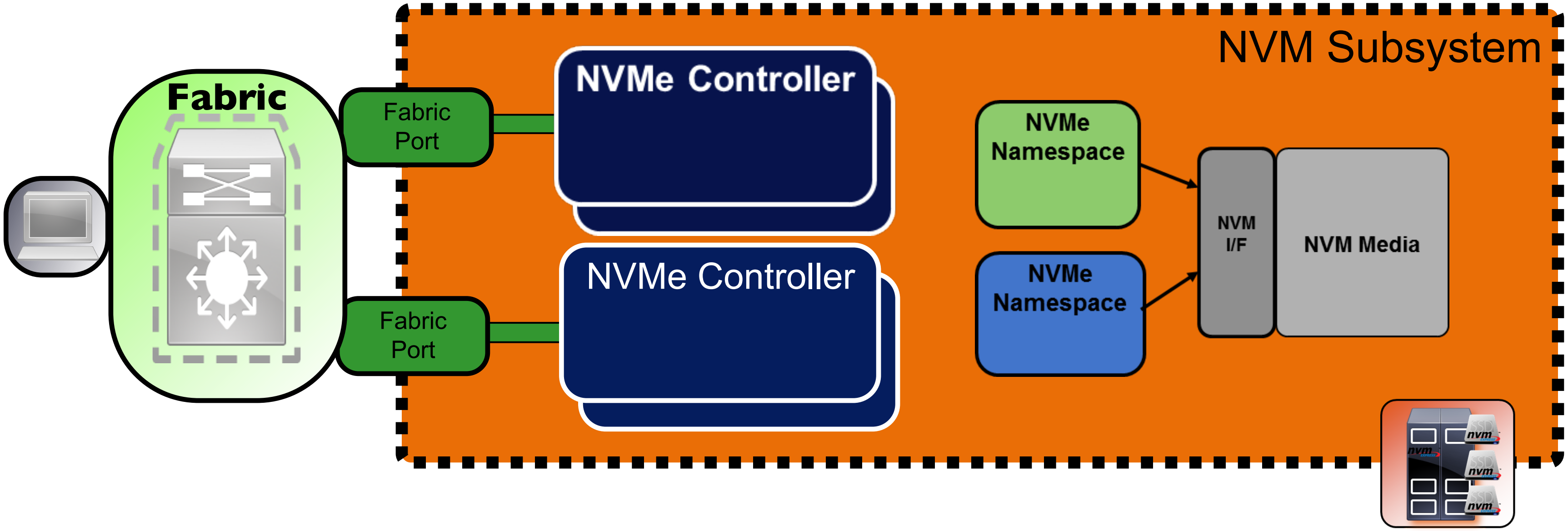
(not used in RDMA)

NVMe over Fabric Response Capsule

- Encapsulated NVMe CQE Entry
- May contain NVMe Command Data
- Transport agnostic Capsule format

Fabric Ports

- Subsystem Ports are associated with Physical Fabric Ports
- Multiple NVMe Controllers may be accessed through a single port
- NVMe Controllers each associated with one port
- Fabric Types; PCIe, RDMA (Ethernet RoCE/iWARP, InfiniBand™), Fibre Channel/FCoE



Key Points About NVMe/F

- NVMe built from the ground up to support a consistent model for NVM interfaces, even across network fabrics
  - Host “sees” networked NVM as if local
  - NVMe commands and structures are transferred end-to-end
  - Maintains the NVMe architecture across a range of fabric types
- Simplicity enables hardware automated I/O Queues – NVMe transport bridge
- No translation to or from another protocol like SCSI (in firmware/software)
- Separation between control traffic (administration) and data I/O traffic
- Inherent parallelism of NVMe multiple I/O Queues exposed to the host

RDMA Explained and Why Chosen for NVMe/F

What Is Remote Direct Memory Access (RDMA)?

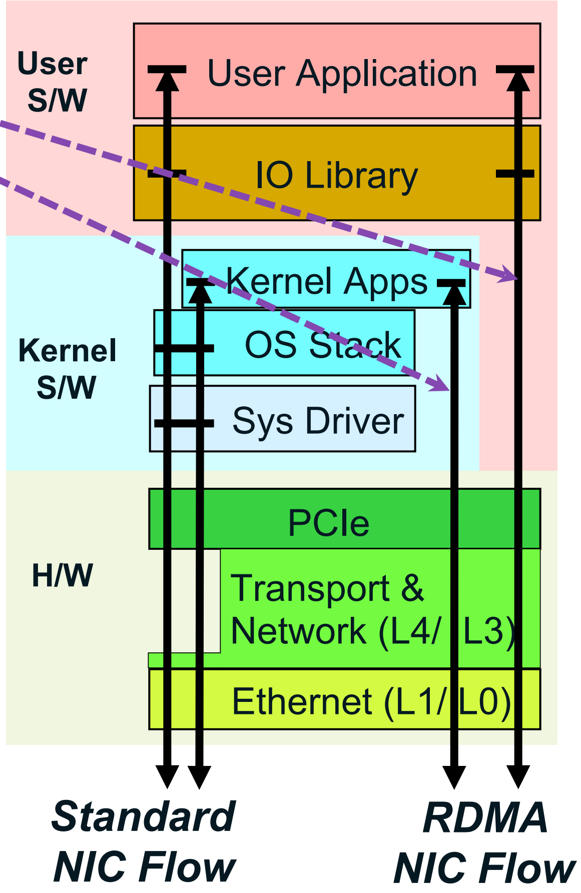
- RDMA is a host-offload, host-bypass technology that allows an application (including storage) to make data transfers directly to/from another application's memory space

- The RDMA-capable Ethernet NICs (RNICs) – not the host – manage reliable connections between source and destination

- Applications communicate with the RDMA NIC using dedicated Queue Pairs (QPs) and Completion Queues (CQs)
  - Each application can have many QPs and CQs
  - Each QP has a Send Queue (SQ) and Receive Queue (RQ)
  - Each CQ can be associated with multiple SQs or RQs

Benefits of Remote Direct Memory Access

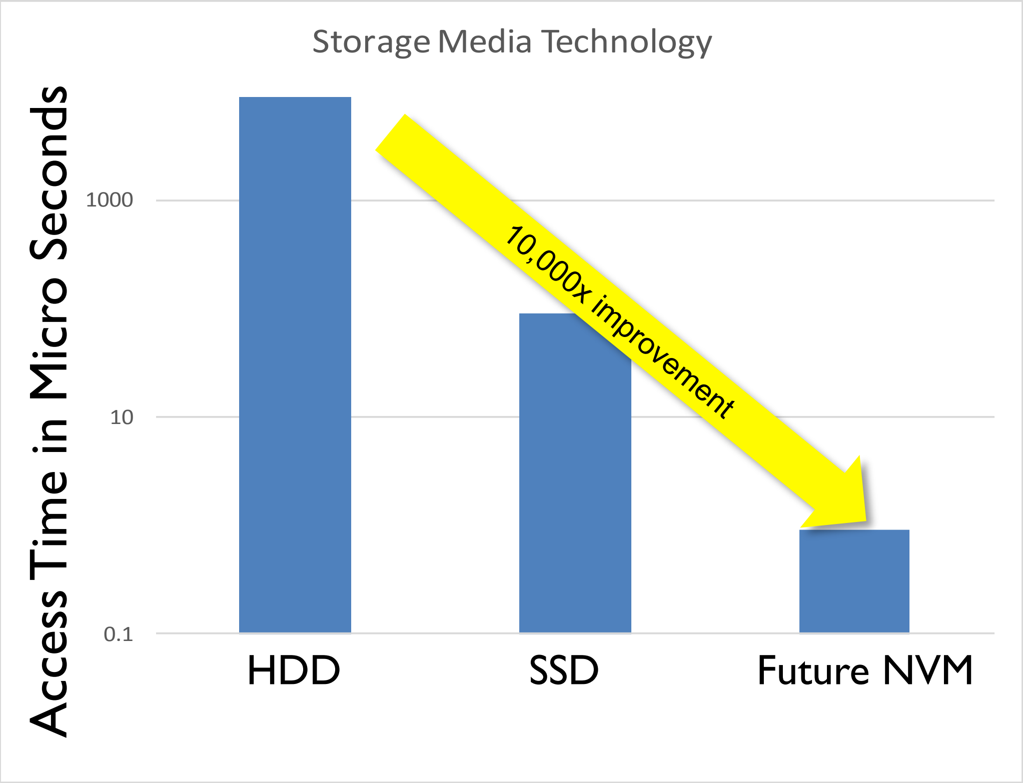
- Bypass of system software stack components that processes network traffic
  - For user applications (outer rails), RDMA bypasses the kernel altogether
  - For kernel applications (inner rails), RDMA bypasses the OS stack and the system drivers
- Direct data placement of data from one machine (real or virtual) to another machine – without copies
- Increased bandwidth while lowering latency, jitter, and CPU utilization
  - Great for networked storage!



Details on RDMA Performance Benefits

RDMA Technique	Benefit		
	CPU Util.	Latency	Mem bw
Offload network transport (e.g. TCP/IP) from Host	✓	✓	
Eliminate receive memory copies with tagged buffers		✓	✓
Reduce context switching with OS bypass (map NIC hardware resources into user space)		✓	
Define an asynchronous "verbs" API (sockets is synchronous)	✓		✓
Preserve message boundaries to enable application (e.g. SCSI) header/data separation	✓		✓
Message-level (not packet-level) interrupt coalescing	✓		

Low NVMe Latency “Exposes” Network Latencies

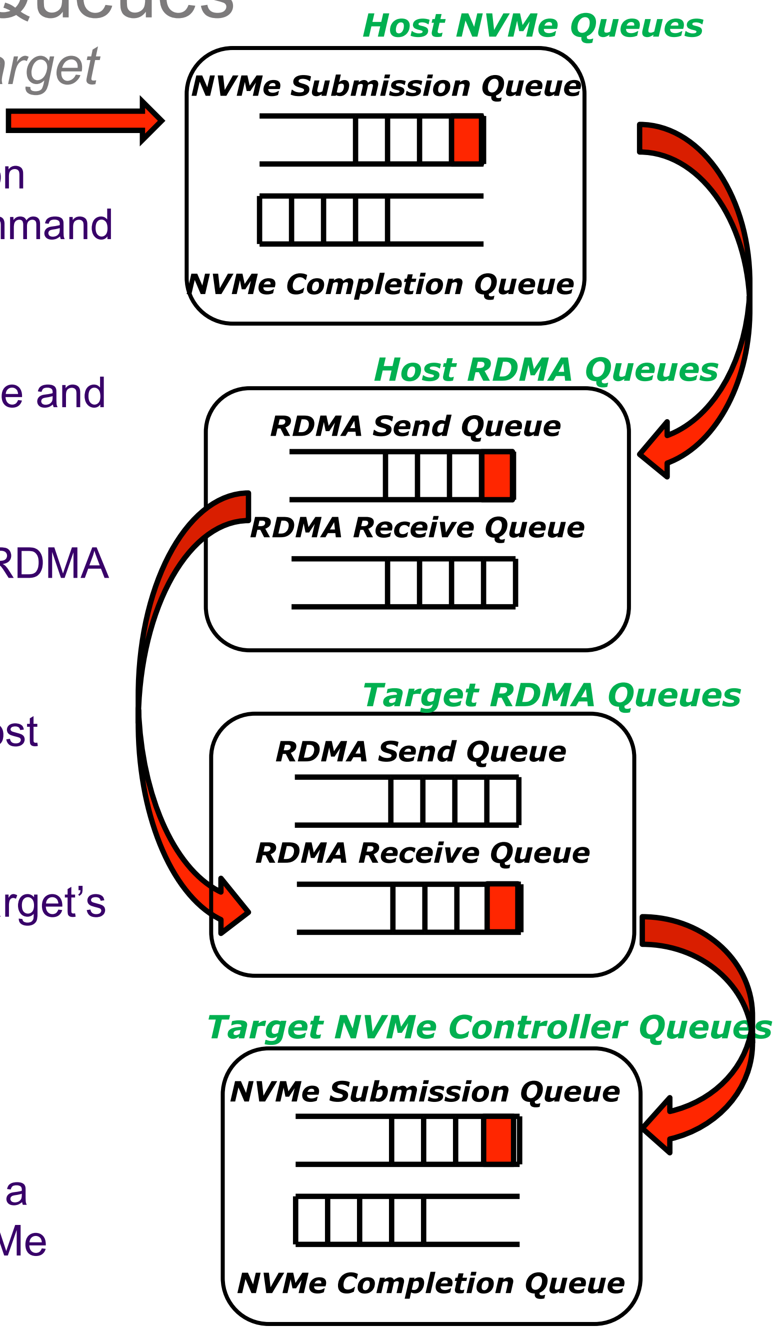


- As storage latency drops, network latency becomes important
  - Both the physical network and the network software stack add latency
  - CPU interrupts and utilization also matter
  - Faster storage requires faster networks

Queues, Capsules, and More Queues

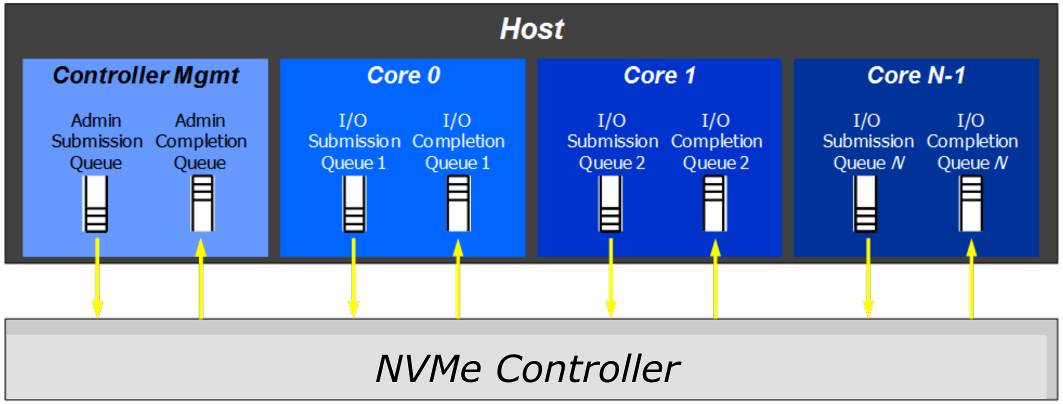
Example of Host Write To Remote Target

- NVMe Host Driver encapsulates the NVMe Submission Queue Entry (including data) into a fabric-neutral Command Capsule and passes it to the NVMe RDMA Transport
- Capsules are placed in Host RNIC RDMA Send Queue and become an RDMA_SEND payload
- Target RNIC at a Fabric Port receives Capsule in an RDMA Receive Queue
- RNIC places the Capsule SQE and data into target host memory
- RNIC signals the RDMA Receive Completion to the target's NVMe RDMA Transport
- Target processes NVMe Command and Data
- Target encapsulates the NVMe Completion Entry into a fabric-neutral Response Capsule and passes it to NVMe RDMA Transport



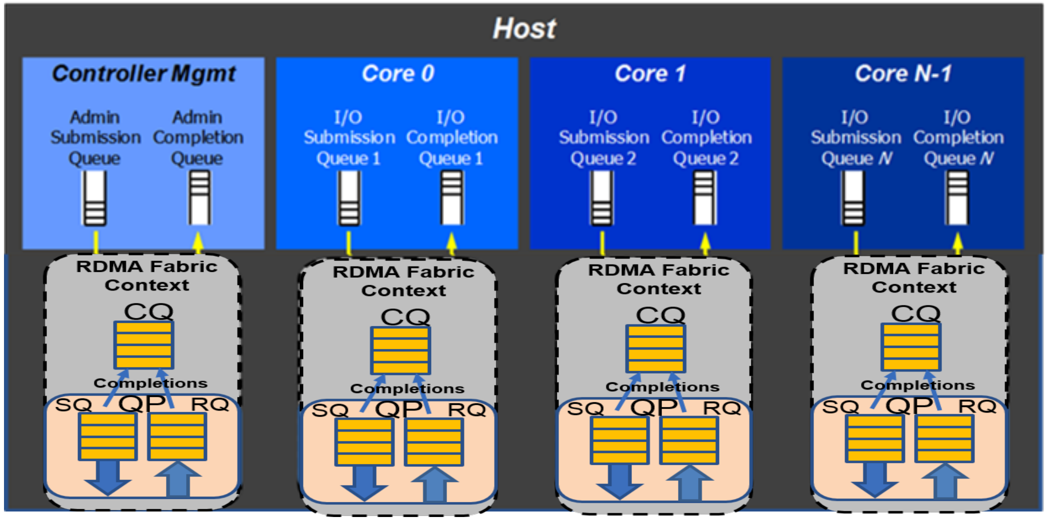
NVMe Multi-Queue Host Interface Maps Neatly to the RDMA Queue-Pair Model

Standard (local) NVMe



- NVMe Submission and Completion Queues are aligned to CPU cores
- No inter-CPU software locks
- Per CQ MSI-X interrupts enable source core interrupt steering

NVMe Over RDMA Fabric



- Retains NVMe SQ/CQ CPU alignment
- No inter-CPU software locks
- Source core interrupt steering retained by using RDMA Event Queue MSI-X interrupts

Verbs, the *lingua franca* of RDMA

The Application RDMA Programming Model Is Defined By “Verbs” (IETF draft[1] and InfiniBand spec[2])

- Verbs are the common standardized basis of the different RDMA system software APIs
  - Verbs also provide a behavioral model for RNICs

- Requires new programming model – not “sockets”

- SMB Direct, iSER, and NFSoRDMA storage protocols take advantage of verbs ***in system software***
  - This makes RDMA transparent to applications

- **NVMe/F adopts similar approach and generates the necessary verbs to drive the fabric**
  - No applications changes or rewrites required!
  - Remote NVMe devices just look local to the host

1. http://tools.ietf.org/html/draft-hilland-rddp-verbs-00
2. https://cw.infinibandta.org/document/dl/7859 , Chapter 11

More On Verbs

- A few of the most common Verbs:
  - PostSQ Work Request (WR): transmit data (or a read request) to remote peer
  - PostRQ WR: provide the RDMA NIC with empty buffers to fill with untagged (unsolicited) messages from remote peer
  - Poll for Completion: Obtain a Work Completion from RDMA NIC
  - A SQ WR completes when the RDMA NIC guarantees its reliable delivery to remote peer
  - A RQ WR completes when its buffer has been filled by a received message
  - Request Completion Notification: Request an interrupt on issue of a CQ Work Completion

Server OS Support for RDMA Verbs

- Windows Server
  - Network Direct userspace API supported since Windows HPC Server 2008
  - Network Direct Kernel API supported since Windows Server 2012
- Linux
  - Userspace/kernel APIs supported by the OpenFabrics Alliance since 2004
  - Upstream in most popular server distros, including RHEL and SLES
- FreeBSD
  - OpenFabrics userspace/ kernel APIs supported since 2011 (FreeBSD 9.0+)

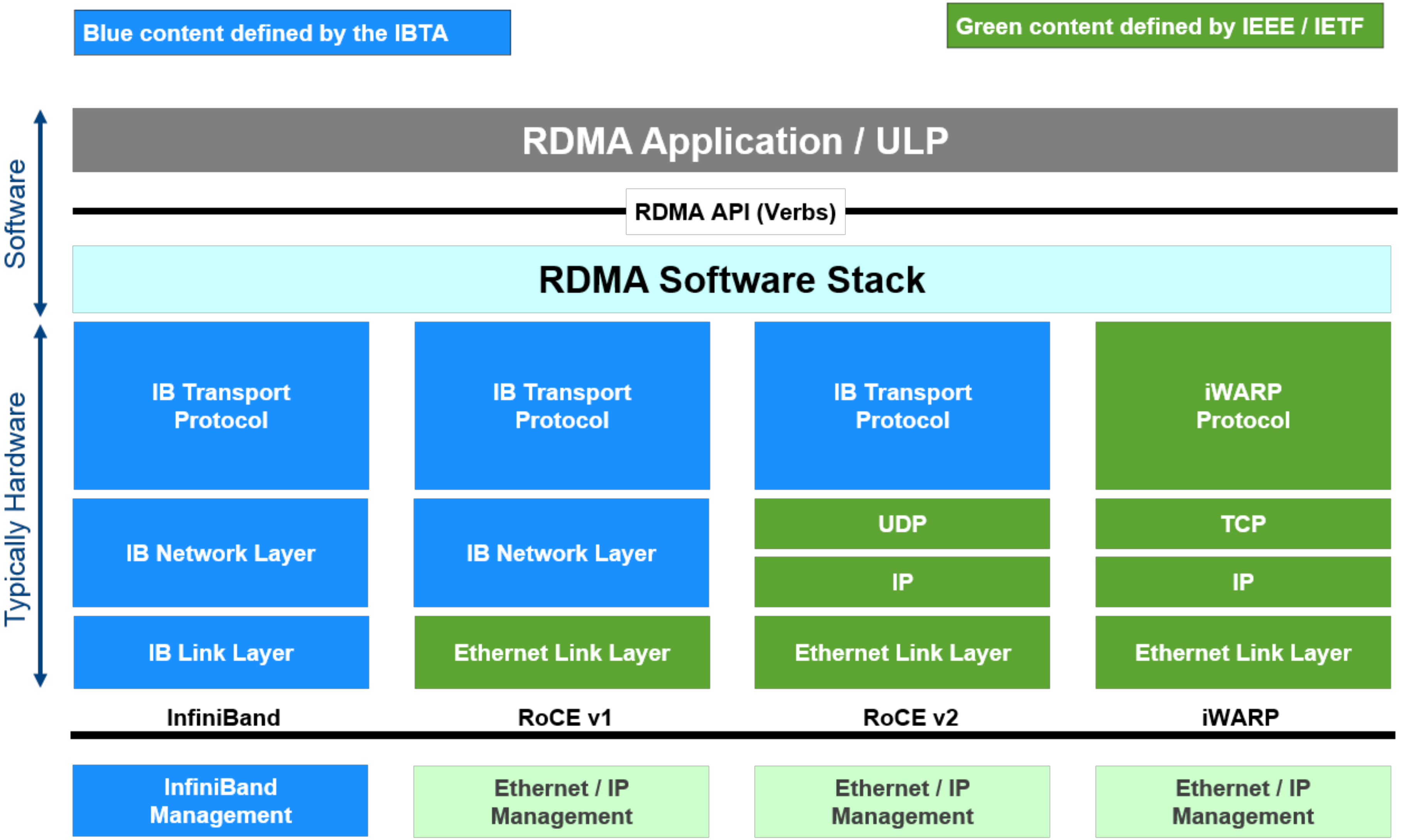
Varieties of Ethernet RDMA explained

SNIA™
Ethernet Storage Forum

Both iWARP and RoCE Provide Ethernet RDMA Services

- RoCE is based on InfiniBand transport over Ethernet
  - RoCEv2 enhances RoCE with a UDP header and Internet routability
    - Uses IP but not TCP
  - RoCEv2 uses InfiniBand transport on top of Ethernet
- iWARP is layered on top of TCP/IP
  - Offloaded TCP/IP flow control and management
- Both iWARP and RoCE (and InfiniBand) support verbs
  - NVMe/F using Verbs can run on top of either transport

Underlying ISO Stacks Of the Flavors of Ethernet RDMA



Blue content defined by the IBTA

Green content defined by IEEE / IETF

Software

Typically Hardware

RDMA Application / ULP

RDMA API (Verbs)

RDMA Software Stack

InfiniBand	RoCE v1	RoCE v2	iWARP
IB Transport Protocol	IB Transport Protocol	IB Transport Protocol	iWARP Protocol
IB Network Layer	IB Network Layer	UDP	TCP
		IP	IP
IB Link Layer	Ethernet Link Layer	Ethernet Link Layer	Ethernet Link Layer
InfiniBand Management	Ethernet / IP Management	Ethernet / IP Management	Ethernet / IP Management

Deployment Considerations for RDMA Enhanced Ethernet

SNIA™
Ethernet Storage Forum

Compatibility Considerations

- iWARP and RoCE are software-compatible if written to the RDMA Verbs

- iWARP and RoCE both require RNICs

- iWARP and RoCE cannot talk RDMA to each other because of L3/L4 differences
  - iWARP adapters can talk RDMA only to iWARP adapters
  - RoCE adapters can talk RDMA only to RoCE adapters

Ethernet RDMA Vendor Ecosystem

- RoCE Supported by IBTA and RoCE Alliance
  - Avago (Emulex), Mellanox
  - Adapter support promised by QLogic, some startups
- iWARP supported by Chelsio and Intel
  - Support from Intel in a future server chipset
  - Adapter support promised by QLogic, some startups
- Both RoCE and iWARP run on all major Ethernet switches (Arista, Cisco, Dell, HPE, Mellanox, etc.)

Network Deployment Considerations

- Data Center Bridging
  - iWARP can benefit from an lossless DCB fabric but does not require DCB because it uses TCP
  - RoCE and RoCEv2 require an lossless DCB fabric
    - Similar to FCoE requirements but across the L2 subnet
      - RoCEv2 is L3 routable
    - Minimum of Priority Flow Control (PFC)
    - All major enterprise switches support DCB

- Congestion management
  - iWARP leverages TCP/IP (e.g., windowing), RFC3168 ECN, and other IETF standards
  - RoCE can use RoCE Congestion Management, which leverages ECN

Summary

- NVMe/F requires the low network latency that RDMA can provide
  - RDMA reduces latency, improves CPU utilization
- NVMe/F supports RDMA verbs transparently
  - No changes to applications required
- NVMe/F maps NVMe queues to RDMA queue pairs
- RoCE and iWARP are software compatible (via Verbs) but do not interoperate because their transports are different
- RoCE and iWARP
  - Different vendors and ecosystem
  - Different network infrastructure requirements

For More Information On RDMA Enabled Ethernet

- For iWARP
  - “iWARP, the Movie”: https://www.youtube.com/watch?v=ksXmfZxqMBQ
  - Chelsio Communications white papers: http://www.chelsio.com/white-papers/
- For RoCE
  - RoCE Initiative: http://www.roceinitiative.org/
  - InfiniBand Trade Association: http://www.infinibandta.org/
  - Mellanox: http://www.mellanox.com
  - Avago (Emulex): http://www.emulex.com/

After This Webcast

- Please rate this Webcast and provide us with feedback
- This Webcast and a PDF of the slides will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
- http://www.snia.org/forums/esf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog
- http://sniaesfblog.org/
- Follow us on Twitter @SNIAESF

Thank you!

SNIA™

Ethernet Storage Forum