REALIZE

The Future of Flash:
NVMe over Fibre Channel

BROCADE

DELL EMC World

Legal Disclaimer

All or some of the products detailed in this presentation may still be under development and certain specifications, including but not limited to, release dates, prices, and product features, may change. The products may not function as intended and a production version of the products may never be released. Even if a production version is released, it may be materially different from the pre-release version discussed in this presentation.

Nothing in this presentation shall be deemed to create a warranty of any kind, either express or implied, statutory or otherwise, including but not limited to, any implied warranties of merchantability, fitness for a particular purpose, or non-infringement of third-party rights with respect to any products and services referenced herein.

ADX, Brocade, Brocade Assurance, the B-wing symbol, DCX, Fabric OS, HyperEdge, ICX, MLX, MyBrocade, OpenScript, The Effortless Network, VCS, VDX, Vplane, and Vyatta are registered trademarks, and Fabric Vision and vADX are trademarks of Brocade Communications Systems, Inc., in the United States and/or in other countries. Other brands, products, or service names mentioned may be trademarks of others.

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

Top Two Exciting Storage Trends

Faster Media and Faster Protocols

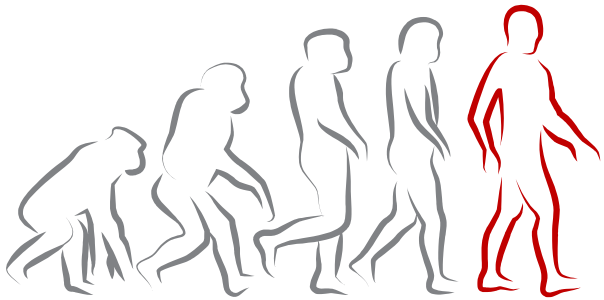
BROCADE

DELL EMC World

Storage is Rapidly Evolving

The next evolution in high performance SSD interfaces is NMVe

- NVMe is a purpose-built protocol for solid state drives
  - Replaces SCSI protocol
  - Dramatically reduces latency through parallelism
  - Originally limited to PCIe direct attached storage
- To provide scalability, the NVMe folks have established “NVMe over Fabrics”
- Network requirements include low latency, fast performance, reliability, and scalability



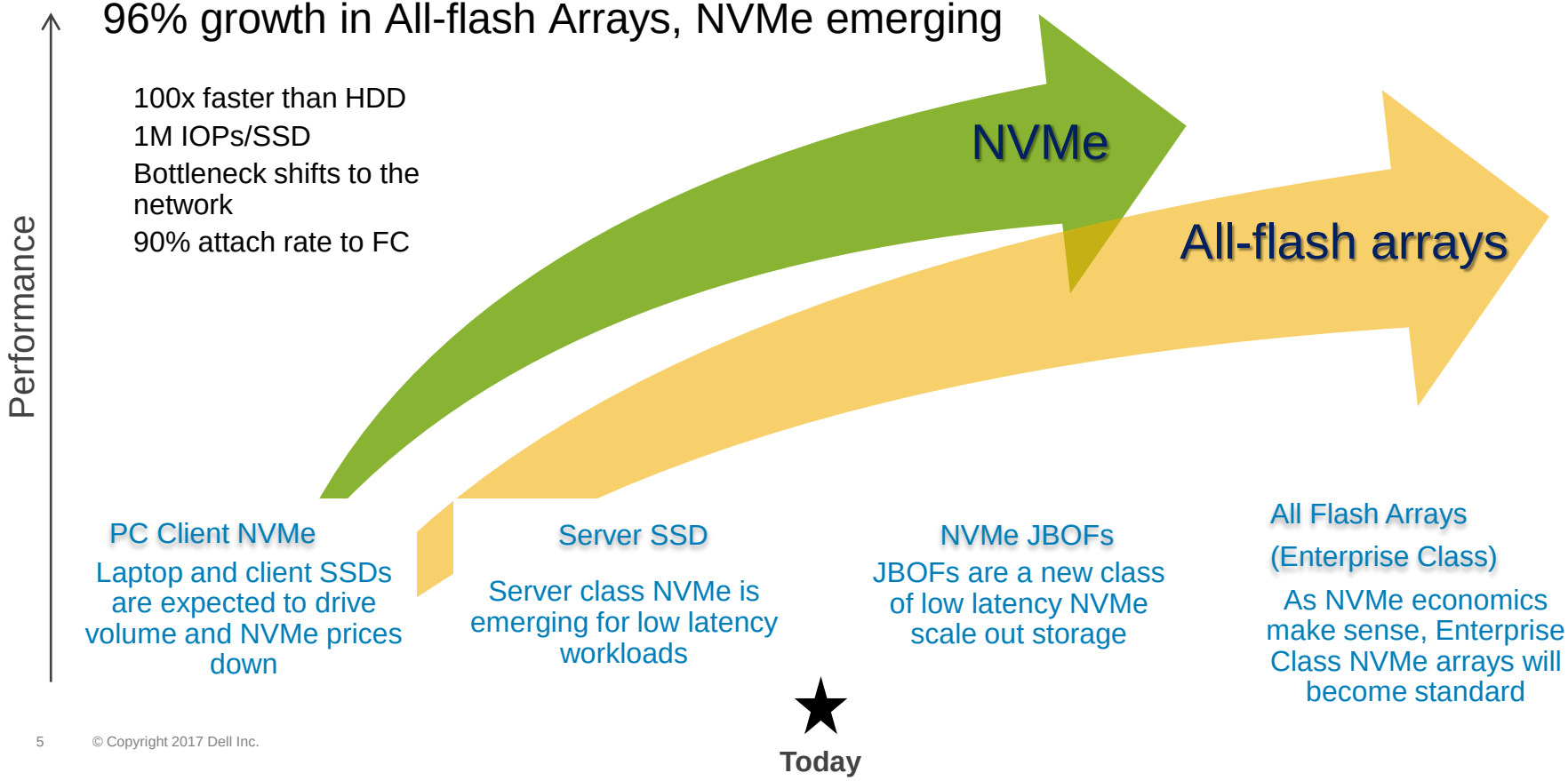
Evolving SSD Technologies:

- 3D XPoint
- 3D Nand

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

A Seismic Shift is Underway in Enterprise Datacenters

96% growth in All-flash Arrays, NVMe emerging



© Copyright 2017 Dell Inc.

NVMe over PCIe

A Faster Software Interface for Server-attached Flash

BROCADE

DELL EMC World

NVM Express™ Ecosystem

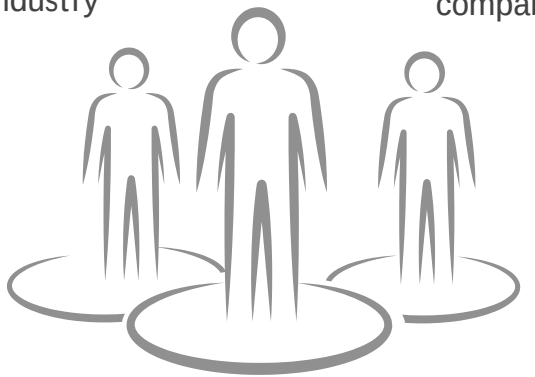
Drivers for Windows®, Linux*, VMware*, Solaris*, FreeBSD*, and UEFI

NVM Express, Inc.

Includes more than 75 firms from across the industry

Promoter Group

Led by 13 elected companies



Windows Server 2012 R2 Certified

Windows Server 2012 Certified

Certified for Windows Server 2008 R2

Windows 8 Compatible

Windows 7

Windows 8.1

suse

ubuntu

FreeBSD

redhat

chrome

solaris

Microsoft

NetApp

ORACLE

PMC

SAMSUNG

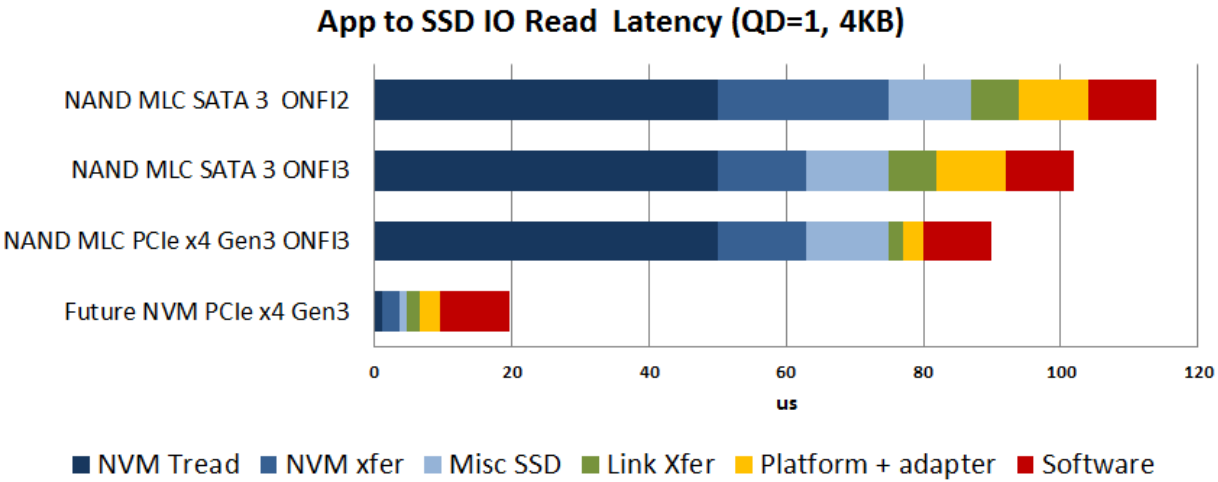
SanDisk

Seagate

© Copyright 2017 Dell Inc.

© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

NVMe advantage over SCSI as protocol



- NVM Express* (NVMe) latency may be < 10 µs with next generation NVM
- Using a SCSI-based protocol for remote NVMe adds over 100 µs in latency

Concern: Low latency of next gen NVM lost in (SCSI) translation.

SOURCE: INTEL MEASUREMENTS.

© Copyright 2017 Dell Inc

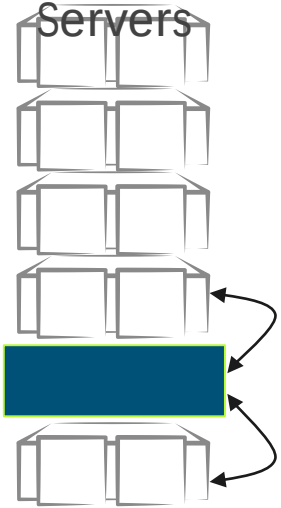
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC

8

Basic NVMe Picture

The scale of NVMe is one Server, or (with PCIe Expansion) a Sub-Rack

Rack of Servers



NVMe flash modules can plug directly into servers

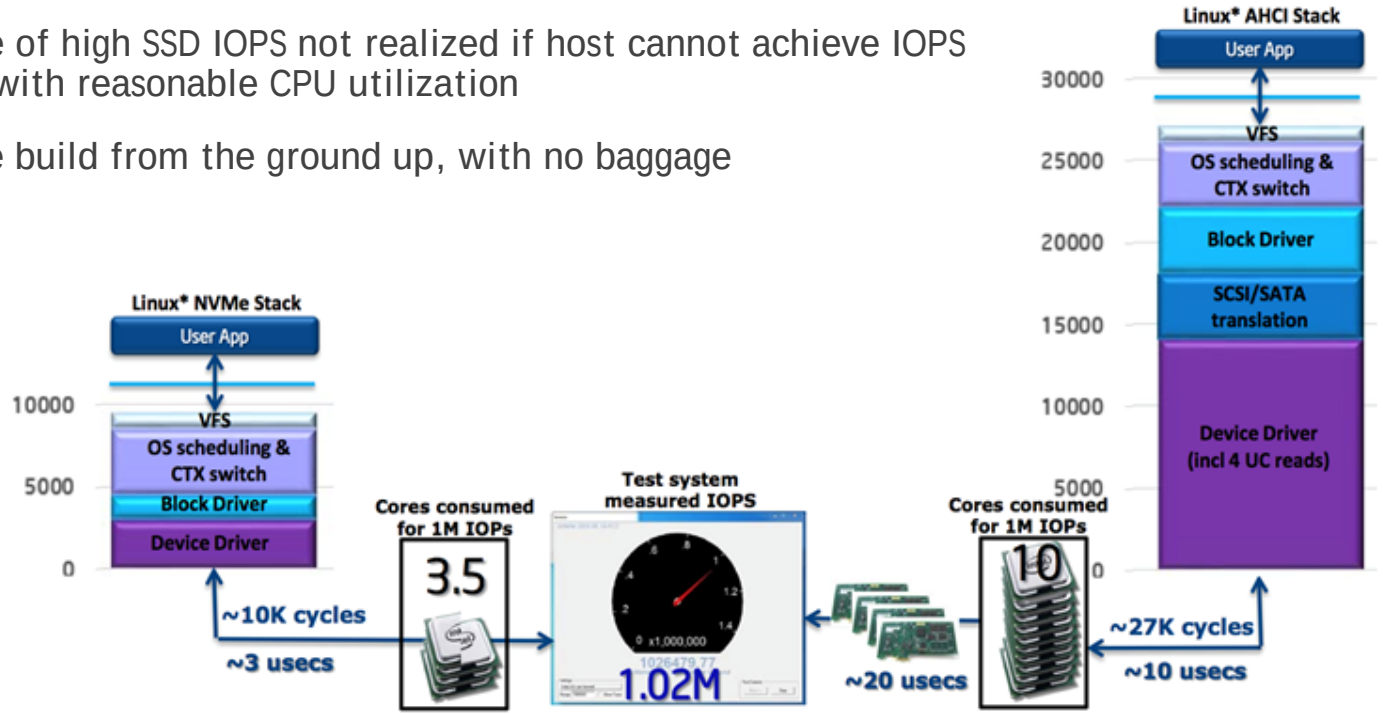
Option of using PCIe Expander

Or they can plug into PCIe expansion slots

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC

NVMe stack efficiency

- Value of high SSD IOPS not realized if host cannot achieve IOPS rate with reasonable CPU utilization
- NVMe build from the ground up, with no baggage

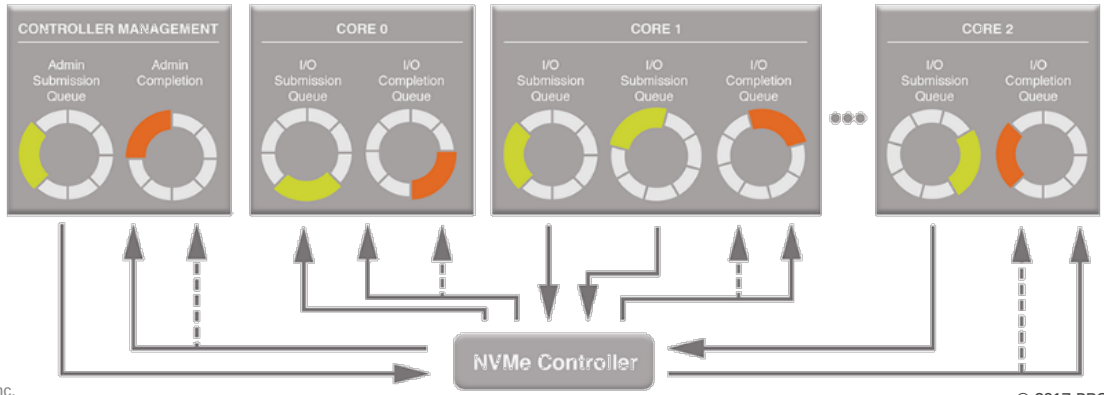


MEASUREMENTS TAKEN ON INTEL CORE , I5-2500K 3.3GHZ 6MB L3 CACHE QUAD-CORE DESKTOP PROCESSOR USING LINUX REDHAT EL.6.0 2.6.32-71 KERNEL USING FIO WITH RAW IO. TESTING AND MEASUREMENT BY INTEL

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

NVMe in More Details

- Supports deep queues (64K commands per queue, up to 64K queues)
- Supports MSI-X and interrupt steering
- Streamlined and simple command set (13 required commands)
- Optional features to address target segment
  - Data Center: Reservations, etc. Client: Power features, etc.



© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC

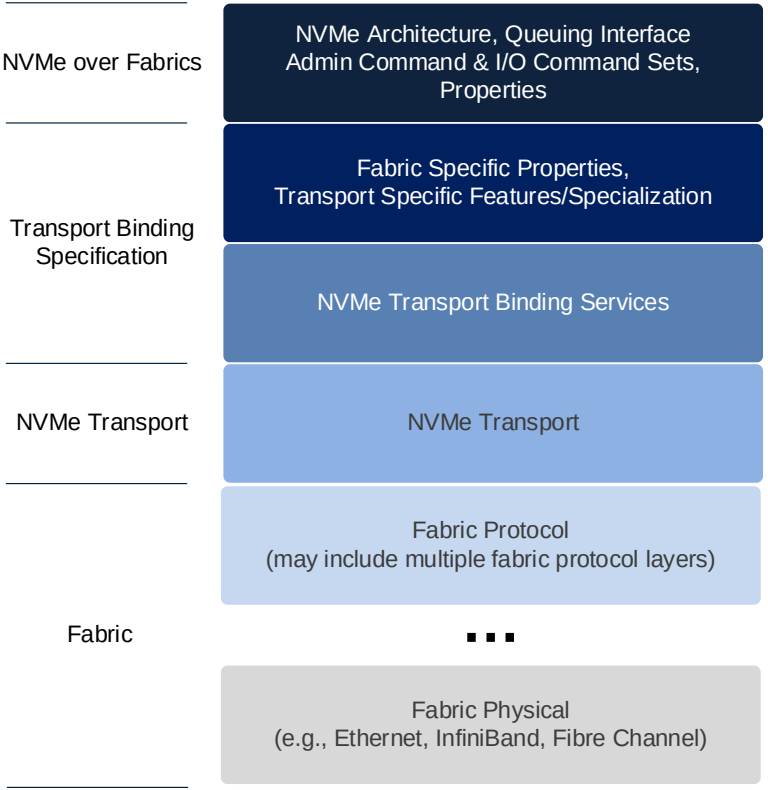
NVMe over Fabrics

Scaling NVMe Beyond the Rack

BROCADE

DELL EMC World

NVMe over Fabrics Stack Architecture



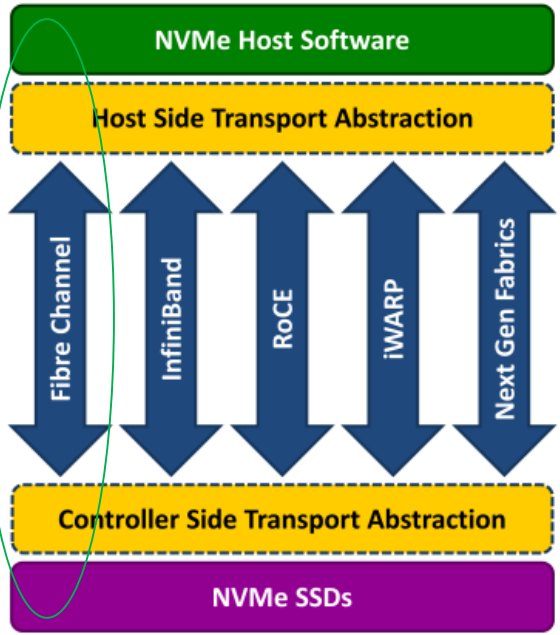
© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

http://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf

NVMe over Fabrics

NVM Express over Fabrics defines a common architecture that supports a range of storage networking fabrics for NVMe block storage protocol over a storage networking fabric. This includes enabling a front-side interface into storage systems, scaling out to large numbers of NVMe devices and extending the distance within a datacenter over which NVMe devices and NVMe subsystems can be accessed.

Work on the NVMe over Fabrics specification began in 2014 with the goal of extending NVMe onto fabrics such as Ethernet, Fibre Channel and InfiniBand®. NVMe over Fabrics is designed to work with any suitable storage fabric technology. This specification was published in June 2016.



Two types of fabric transports for NVMe are currently under development:

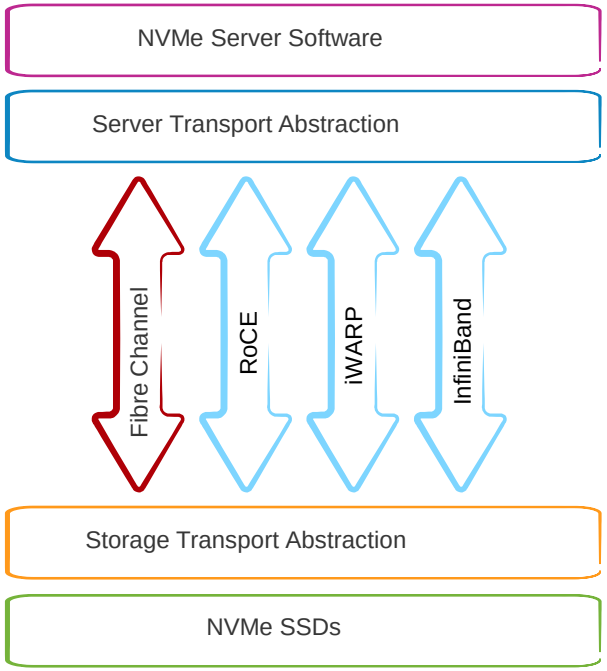
- NVMe over Fabrics using RDMA
- NVMe over Fabrics using Fibre Channel (FC-NVMe)

Using RDMA with NVMe over Fabrics includes any of the RDMA technologies, including InfiniBand, RoCE and iWARP. The development of NVMe over Fabrics with RDMA is defined by a technical sub-group of the NVM Express organization.

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

NVMe over Fabrics

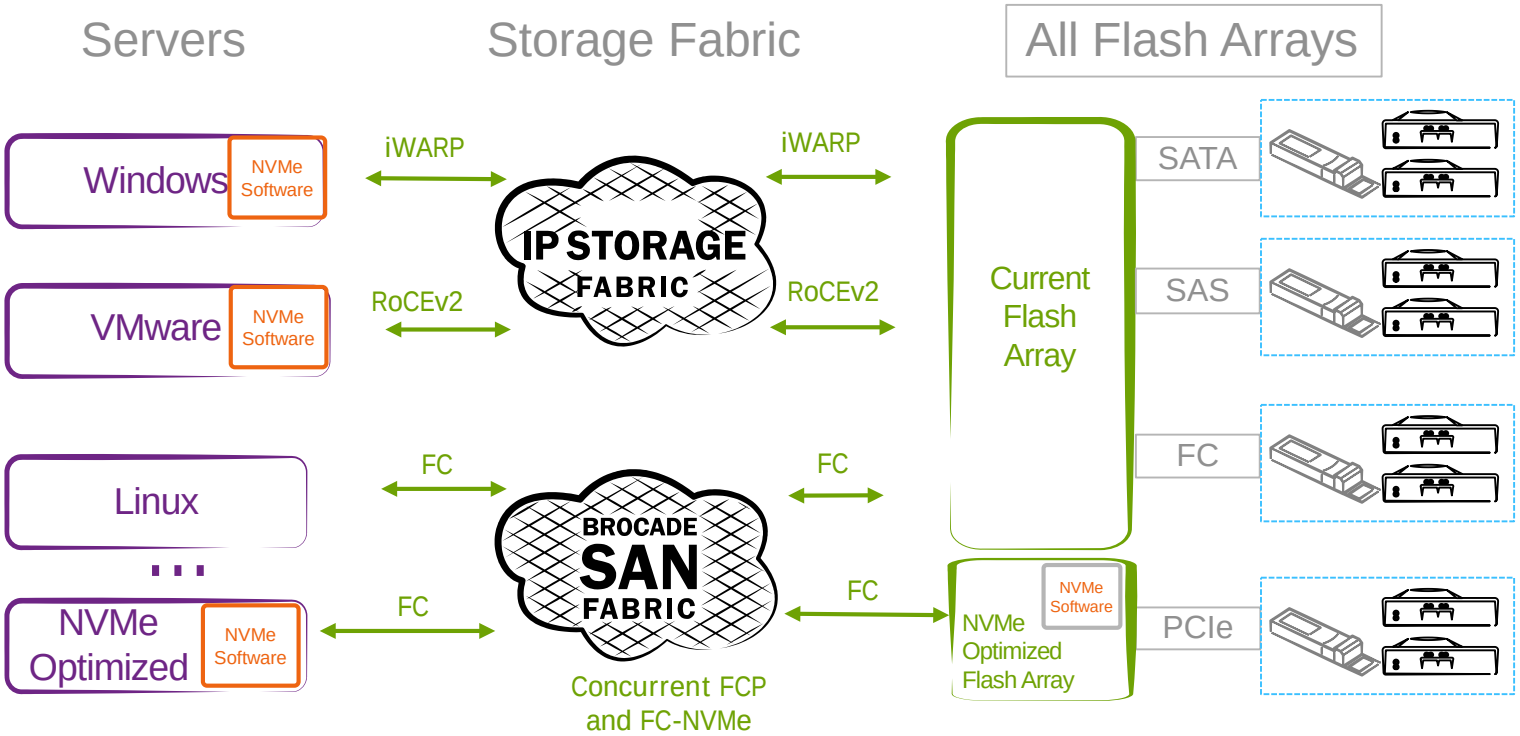
- Direct (PCIe) attached SSD doesn't scale
- Networked storage is a must for large customers needing scalability, availability, and reliability
- Management makes clustering, VM mobility, etc. a reality i.e. requires networked storage
- Fibre Channel is key, as the vast majority of today's all-flash storage is deployed on FC
- Ethernet is also available, but RoCE(v2) needs DCB ("lossless") & iWARP is slow and not widely used
  - In addition, the two are not compatible with each other



© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

The NVMe over Fabrics Picture

NVMe over Fabrics scales to the Datacenter (Rows of Server Racks)



© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC

NVMe over Fibre Channel

The Path Forward for NVMe-based Enterprise Storage

BROCADE

DELL EMC World

NVMe over Fibre Channel

Deploy next-general storage with confidence

Gen 5 and Gen 6 support FC-NVMe

- Seamless Integration with Gen 5 and 6 without a rip-and-replace
- NVMe and SCSI can coexist in the same server and on the same FC SAN.
- Gen 6 supports speeds up to 128Gb with additional monitoring statistics
- Low risk based on proven storage networks

© Copyright 2017 Dell Inc.

© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

Again: http://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf

NVMe over Fabrics Technical Characteristics

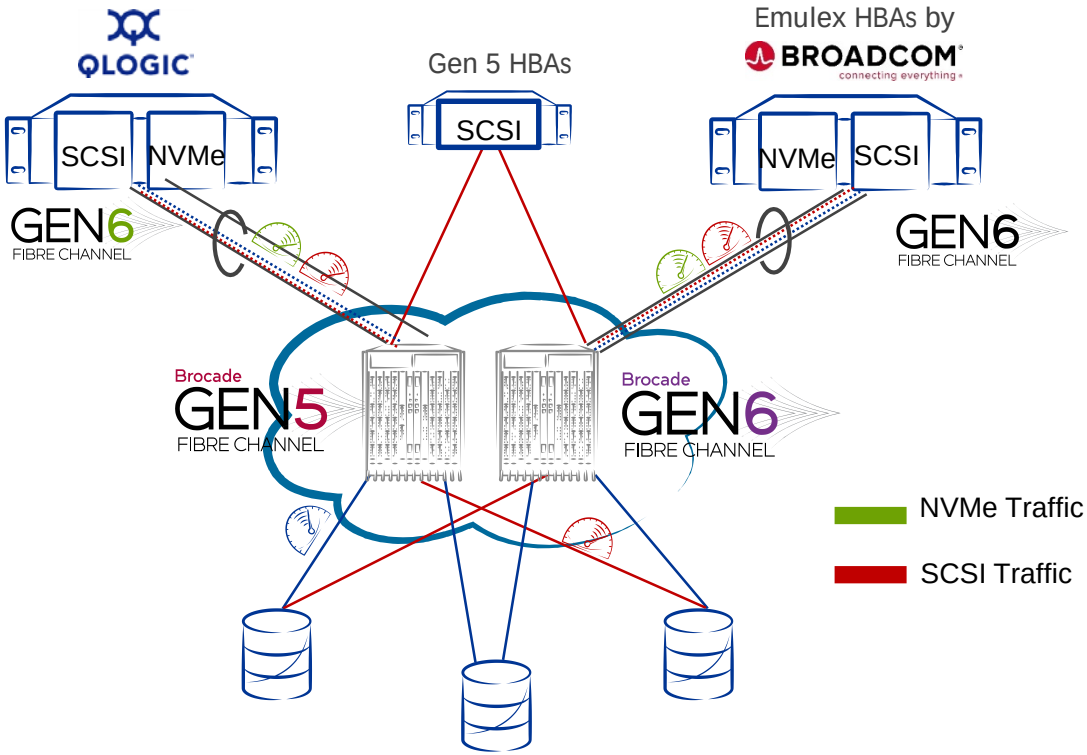
Obviously, transporting NVMe commands across a network requires special considerations over and above those that are determined for local, in-storage memory. For instance, in order to transmit NVMe protocol over a distance, the ideal underlying network or fabric technology will have the following characteristics:

- **Reliable, credit-based flow control and delivery mechanisms.** This type of flow control allows the network or fabric to be self-throttling, providing a reliable connection that can guarantee delivery at the hardware level without the need to drop frames or packets due to congestion. Credit-based flow control is native to Fibre Channel, InfiniBand and PCI Express® transports.
- **An NVMe-optimized client.** The client software should be able to send and receive native NVMe commands directly to and from the fabric without having to use a translation layer such as SCSI.

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

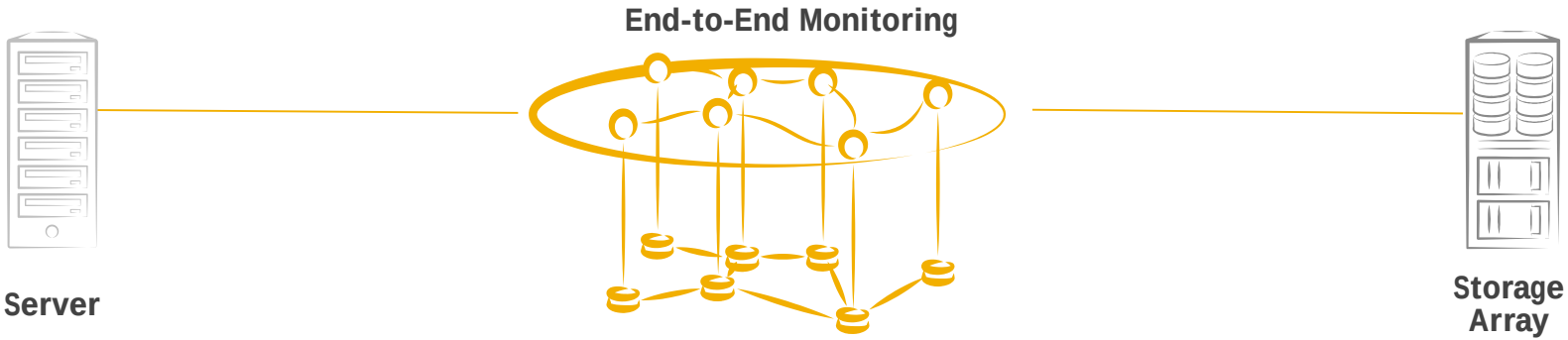
NVMe and SCSI coexistence is huge

- Enhance the performance of your existing SAN; avoid expensive and disruptive duplication or replacement
- NVMe and SCSI will coexist for years to come; separate infrastructures mean years of avoidable stranded capacity and planning pain
- Concurrent FC offers full fabric awareness, visibility and manageability with existing Brocade Fabric Vision technology



© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

NVMe over FC Integrated Network Sensors



Infrastructure Health

Infrastructure Performance

Infrastructure Availability

infrastructure utilization

Application Performance

Application baselining

© Copyright 2017 Dell Inc.

© 2016 BROCADE COMMUNICATIONS SYSTEMS, INC.

Comparing Fabric Options

The Details Matter

BROCADE

DELL EMC World

NVMe over Fabrics – Enterprise Options

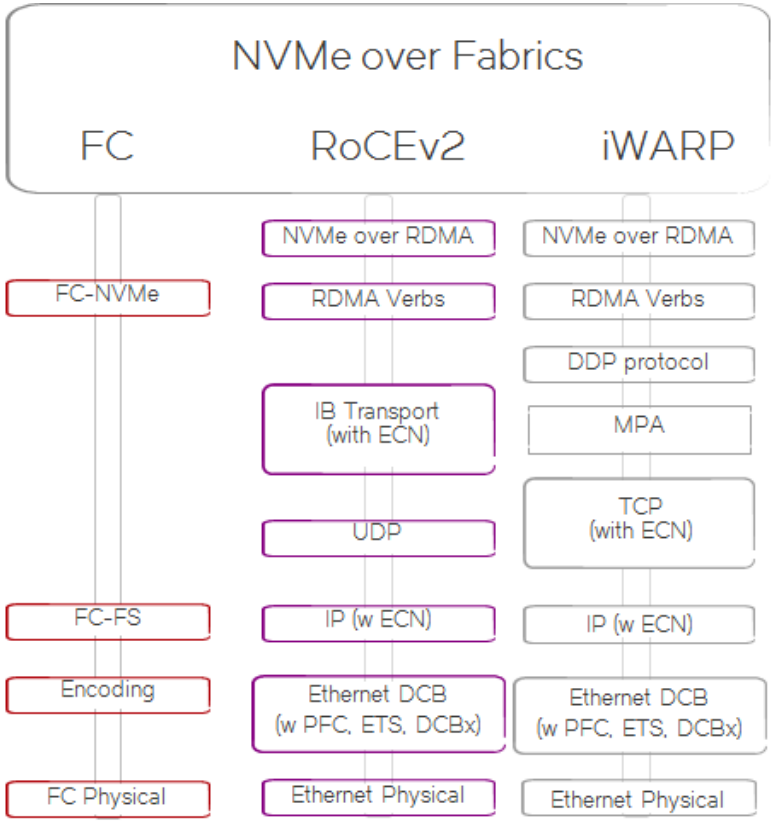
Fabric Type	Latency	Lossless & Deterministic	Bandwidth Link & x4	Scalable	Enterprise Footprint
Fibre Channel	**Lower**	**Yes**	**32Gb/128Gb**	**Yes**	**Dominant Storage Fabric**
Ethernet (RoCE v2)	Low	No	25Gb/100Gb	Yes	Negligible
Infiniband/ OmniPath	Very Low	Yes	56Gb/100Gb	Limited	Limited-scientific (OmniPath has no footprint)
iWARP	Med	No	25Gb/100Gb	Yes	Negligible

NVMe over Fibre Channel – Today; An evolutionary enterprise transition

© Copyright 2017 Dell Inc.

NVMe over different Fabrics

- NVMe's benefit is a simpler stack than SCSI
- NVMe over Fibre Channel extends that simplicity
- But NVMe over RDMA fabrics means complexity
  - RDMA was originally defined for InfiniBand
  - Making RDMA work over IP adds extra network layers
  - Extra network layers complicate the protocol stack
- Using RDMA contradicts the NVMe benefit
  - Even though special RNICs or TOEs manage most of the special protocol layers, they still translate to extra work for configuration, management and troubleshooting



© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

Pain Points with Ethernet/IP (#1)

- Ethernet/IP features (and skills) are focused on best-effort, campus and WAN
  - Dominant NVMe use case is for premium, lossless, datacenter usage
- Ethernet (TRILL) fabrics never took off
  - So fabric must be implemented at IP layer, adding complexity
- Ethernet/IP/TCP/UDP layers and algorithms → complexity
  - Required for internet scale, but not for DC scale
  - Ethernet flow control not credit-based, doesn't propagate through IP routers
  - Flow control with Explicit Congestion Notification (ECN) encourages bad end node behavior
- Lossless Ethernet (DCB) is "recommended"
  - But DCB is not default behavior, poorly interoperable, adds complexity

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC. INTERNAL USE ONLY

Pain Points with Ethernet/IP (#2)

- Plug-n-Play Ethernet/IP/TCP can hide misconfigs until traffic spikes, hard to troubleshoot
- Jumbo frames and VXLAN complicate MAXPDU mgmt.
- No time-tested legacy of fabric-resident name services, zoning, analytics features
- Claimed cost savings based on discount vendor or shared rather than dedicated network
- Storage vendor testing and support of Ethernet/IP products covers tiny % of market
- Ethernet/IP fabric not sold by storage vendor means finger pointing when support needed
- Big pain point #1: Will iWARP or RoCE emerge as de facto standard?
- Big pain point #2: Ethernet NVMe fabrics can not access legacy flash arrays

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC

Pain Points with iWARP

- iWARP (wide area RDMA protocol) is a complex stack involving 8 RFCs
  - 5040, 5041, 5042, 5043, 5044, 6580, 6581, 7308

- Based on TCP
  - Common TCP stacks include slow-start behavior, which ramps up the TCP window on start up and after idle periods, not optimal for low latency
  - Low-latency TCP stacks are suboptimal for WAN, not widely available
  - Using a latency-focused stack means managing multiple TCP stacks in DC

- iWARP performance relies on specialized TCP Offload Engines (TOEs)

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

Pain Points with RoCE for NVMe

- RoCE (RDMA over Converged Ethernet) is relatively young and untested
- RoCEv2 uses “Explicit Congestion Notification” & UDP, but ECN specifies TCP
- RoCE defined by InfiniBand Trade Assn (IBTA), not IETF (ECN) or IEEE (DCB)
- RoCEv2 is not protocol compatible with RoCEv1
  - But most RoCEv2 RNICs can be configured to use RoCEv1

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC

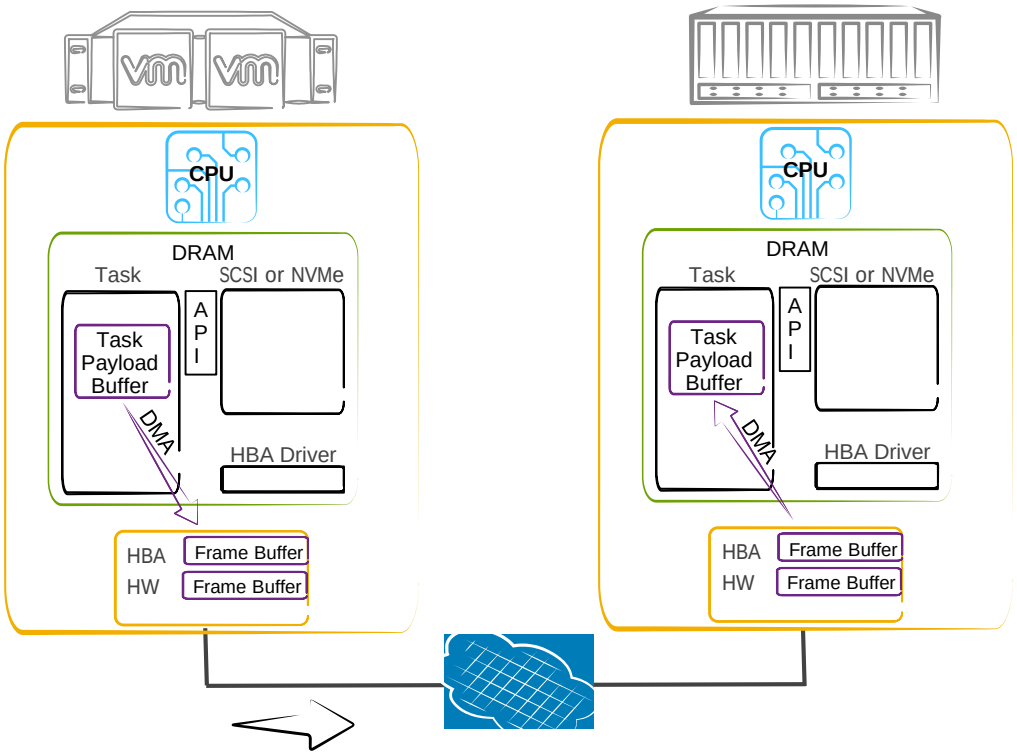
Zero Copy Discussion

Clearing up a Misunderstanding

BROCADE

DELL EMC World

Fibre Channel does Zero Copy

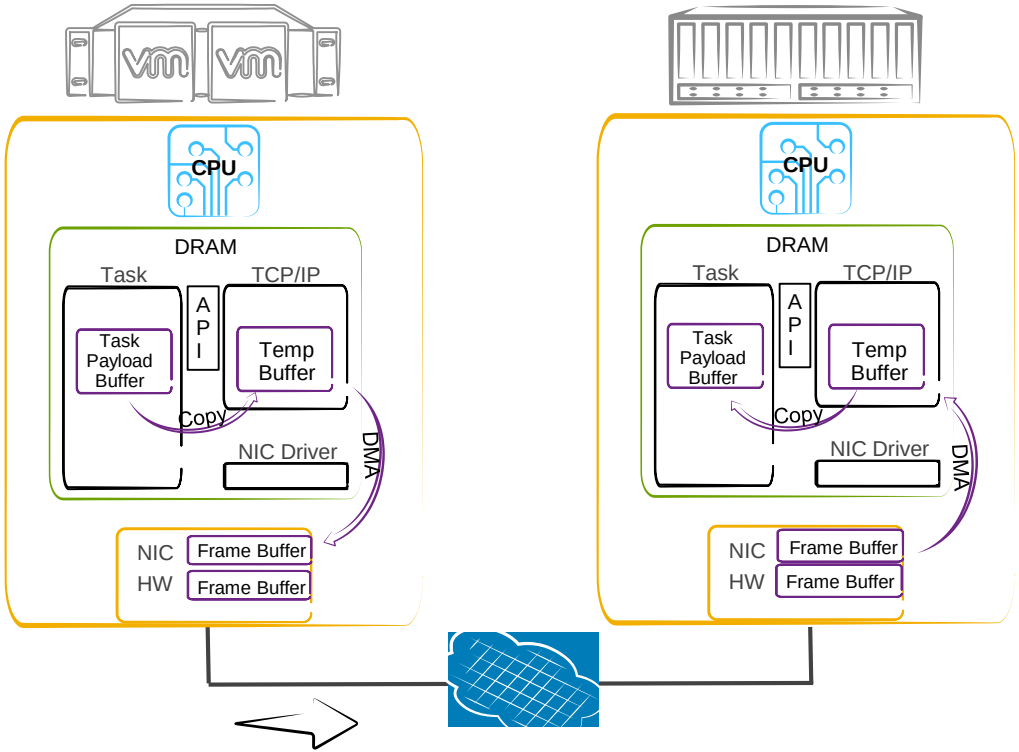


When a task sends using SCSI or NVMe stack and an FC HBA:

- API gives ownership of payload buffer to storage stack; no temp buffer is needed.
- SCSI or NVMe stack processing passes payload buffer address to HBA driver, which passes to HBA hardware.
- HBA hardware DMAs the payload from task buffer to HBA frame buffer for transmission.
- Frame transmitted over network.
- Receiving HBA captures in frame buffer. HBA is storage-centric, so behavior is not protocol dependent.
- HBA DMAs the payload from the frame buffer into task buffer.
- Result: No copies needed on either end.

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

Classic TCP/IP requires DRAM copy

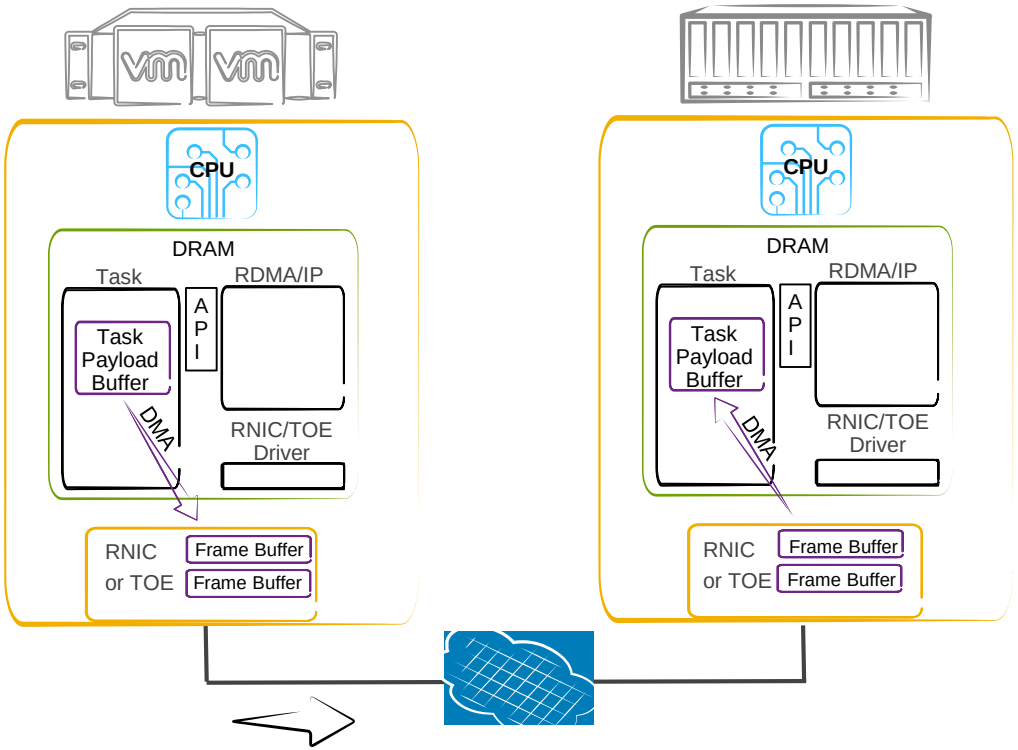


When a task sends using traditional TCP/IP stack:

- API not given ownership of payload buffer, so copies payload to TCP/IP temp buffer.
- TCP/IP stack processing passes temp buffer address to NIC driver, which passes to NIC hardware.
- NIC hardware DMAs from temp buffer to NIC frame buffer for transmission.
- Frame transmitted over network.
- Receiving NIC captures in frame buffer, identifies as TCP/IP.
- NIC DMAs the frame buffer into TCP temp buffer.
- After processing, TCP/IP stack copies temp buffer into task payload buffer.
- Result: DRAM-to-DRAM copies required at each end.

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

RoCE & iWARP caught up to FC zero copy



When a task sends using RDMA/IP stack (works like FC HBA):

- API gives ownership of payload buffer to RMDA stack; no temp buffer needed.
- RDMA/IP stack processing passes payload buffer address to RNIC driver, which passes to RNIC hardware.
- RNIC hardware DMAs the payload from task buffer to RNIC frame buffer for transmission.
- Frame transmitted over the network.
- Receiving RNIC captures in frame buffer, identifies as RDMA/IP.
- RNIC DMAs the payload from the frame buffer into task buffer.
- Result: No copies on either end.

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

Wrapping up

Quick Review: Start Composing Your Questions

BROCADE

DELL EMC World

NVMe - Takeaways

Flash Changes Everything

NVMe offers new type of low latency storage

NVMe will quickly evolve to fabric-based solution

- Initially offered in low latency JBOFs
- Then, integrated into Enterprise-class arrays

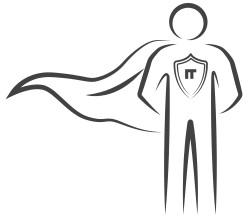
Fibre Channel is the Natural Evolutionary Path

NVMe fabrics require:

- Lossless, Low latency, Secure, Scalable & Proven

Fibre Channel, It Just Works!

Gen 6 is the Right Choice



55% lower latency

Investment protection

- Extends FC technology
- Concurrent SCSI & NVMe

Guaranteed interoperability

Supports the upcoming NVMe over FC T-11 standard

© Copyright 2017 Dell Inc.

© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

Q & A

- Remember these resources available at the Brocade booth
  - “Why Fibre Channel is the NVMe Fabric of Choice” paper
  - “NVMe over Fibre Channel for Dummies” book

© Copyright 2017 Dell Inc.
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.


REALIZE
DELL EMC World