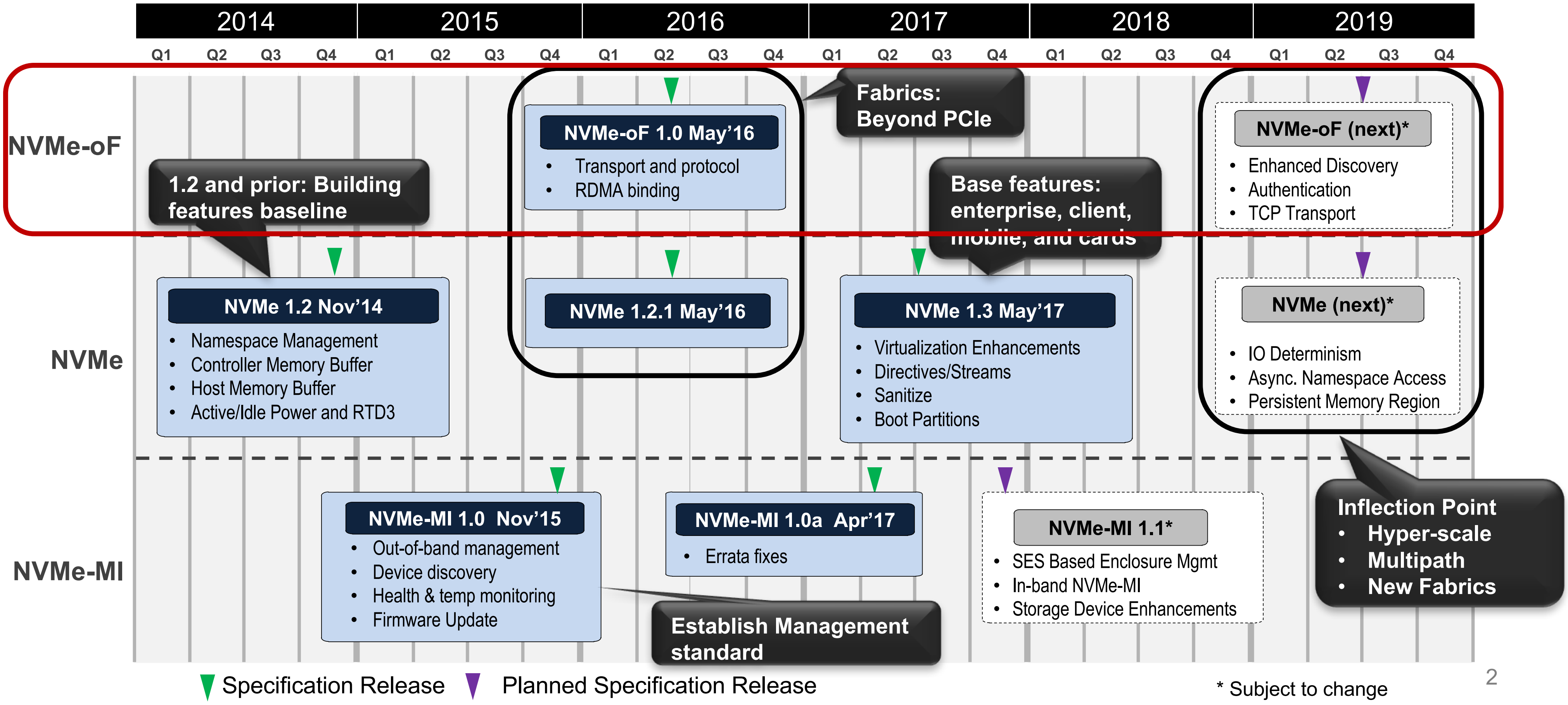
# NVMe over Fabrics
# Session A12 Part B
## 4:55 to 6:00

| | | |
|---|---|---|
| Current/available fabrics Fibre Channel, RoCE, iWarp, and Infiniband | Brandon Hoff<br>Rob Davis<br>Praveen Midha<br>Curt Beckmann<br>Fazil Osman | Software Architect, Broadcom<br>VP of Storage Technology, Mellanox<br>Director, Product Marketing, Cavium<br>Principal Architect, Brocade<br>Distinguished Engineer, Broadcom |
| NVMe-oF Next Frontier – on TCP Layer, et. al. | Dave Minturn | Principal Engineer, Intel |

# NVMe Roadmap



2014 2015 2016 2017 2018 2019

Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4

NVMe-oF

**NVMe-oF 1.0 May'16**
- Transport and protocol
- RDMA binding

**Fabrics: Beyond PCIe**

**NVMe-oF (next)***
- Enhanced Discovery
- Authentication
- TCP Transport

NVMe

**1.2 and prior: Building features baseline**

**NVMe 1.2 Nov'14**
- Namespace Management
- Controller Memory Buffer
- Host Memory Buffer
- Active/Idle Power and RTD3

**NVMe 1.2.1 May'16**

**Base features: enterprise, client, mobile, and cards**

**NVMe 1.3 May'17**
- Virtualization Enhancements
- Directives/Streams
- Sanitize
- Boot Partitions

**NVMe (next)***
- IO Determinism
- Async. Namespace Access
- Persistent Memory Region

NVMe-MI

**NVMe-MI 1.0 Nov'15**
- Out-of-band management
- Device discovery
- Health & temp monitoring
- Firmware Update

**NVMe-MI 1.0a Apr'17**
- Errata fixes

**Establish Management standard**

**NVMe-MI 1.1***
- SES Based Enclosure Mgmt
- In-band NVMe-MI
- Storage Device Enhancements

**Inflection Point**
- **Hyper-scale**
- **Multipath**
- **New Fabrics**

▼ Specification Release ▼ Planned Specification Release

* Subject to change

# NVMe over Fabrics delivers for the External Block Storage Market

All Flash Arrays is a **$6.8B Market** in 2017, growing at a **32% CAGR.**

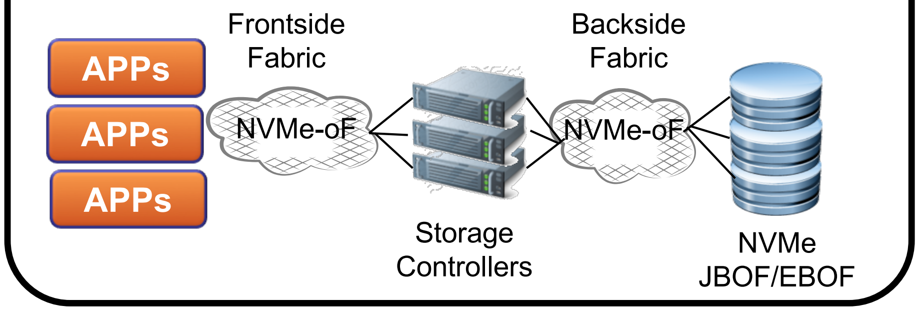
Only 13% of storage capacity shipped is DAS (inside the server), **87% of the total storage capacity shipped is external storage.**

NVMe-oF 1.0 was released in June 2016 and provides support for **RDMA and Fibre Channel,** plus **NVMe-TCP with 1.1**

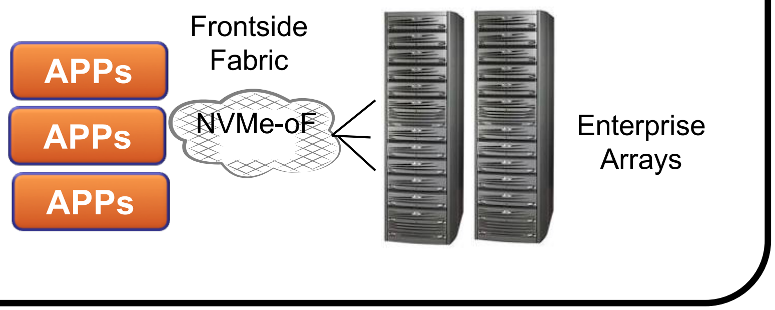
As NVMe becomes adopted, NVMe-oF will enable applications **access to 1000's of NVMe drives with FC, RoCE, iWARP, and TCP** as transport options.

# NVMe over Fabrics - Use Cases

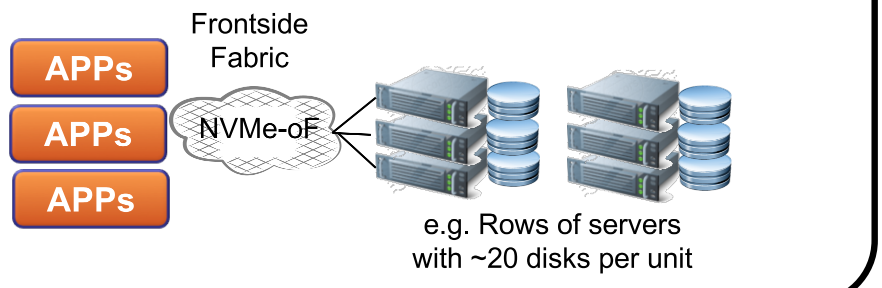
## End to NVMe and NVMe-oF Solutions



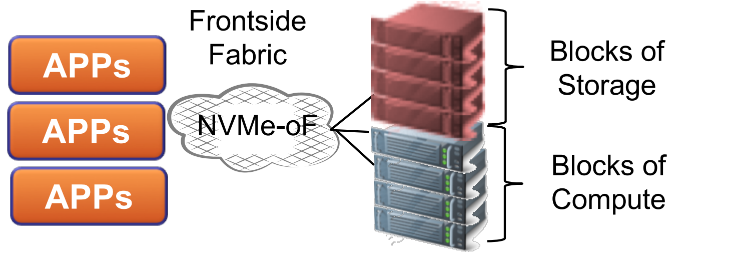
## Traditional SAN



## Server SAN/Disaggregated Storage

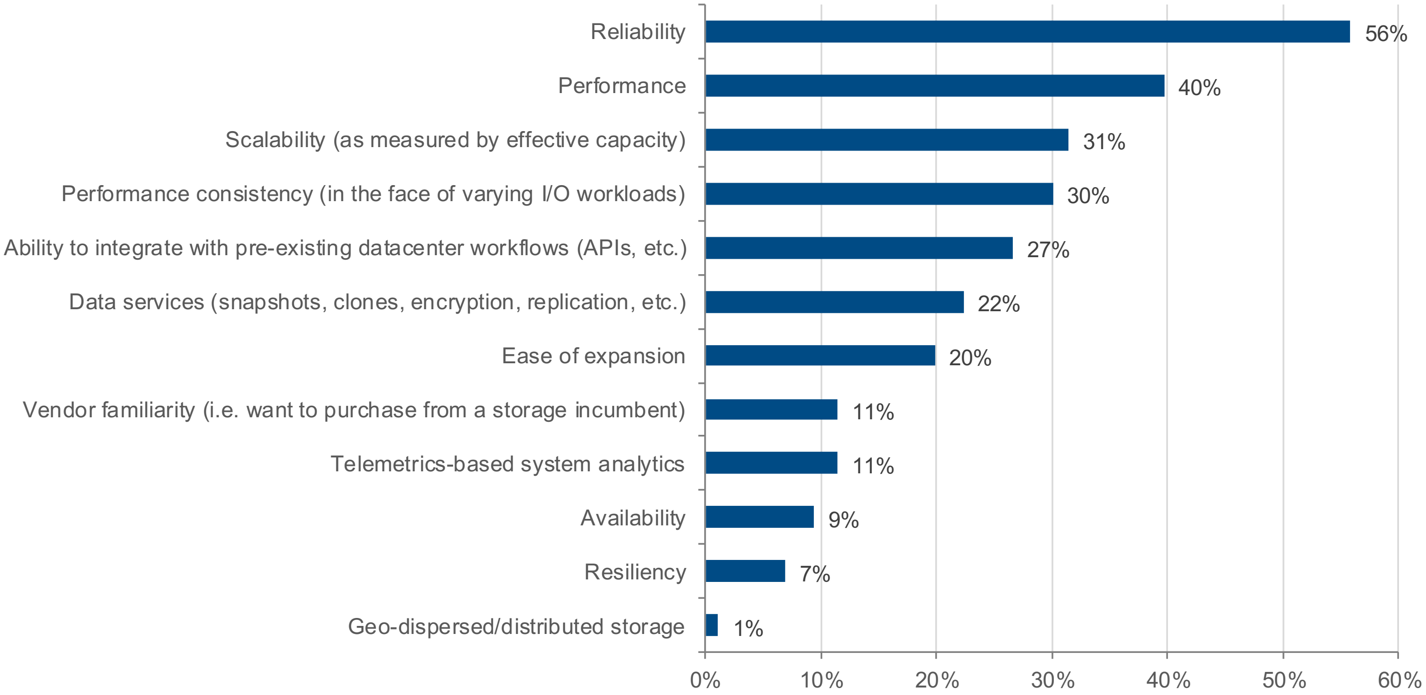


## Rack Scale/Scaleout/HyperScale



# What Drives AFA Purchases?

From the list below, please, select up to three most important criteria when purchasing/considering AFA



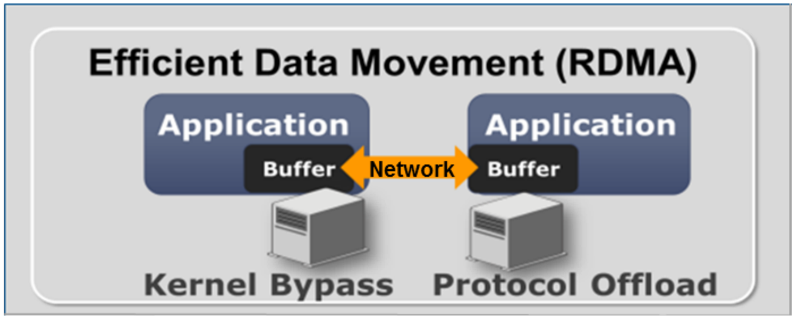
Source: IDC, All-Flash Array Adoption,

# What is RDMA?

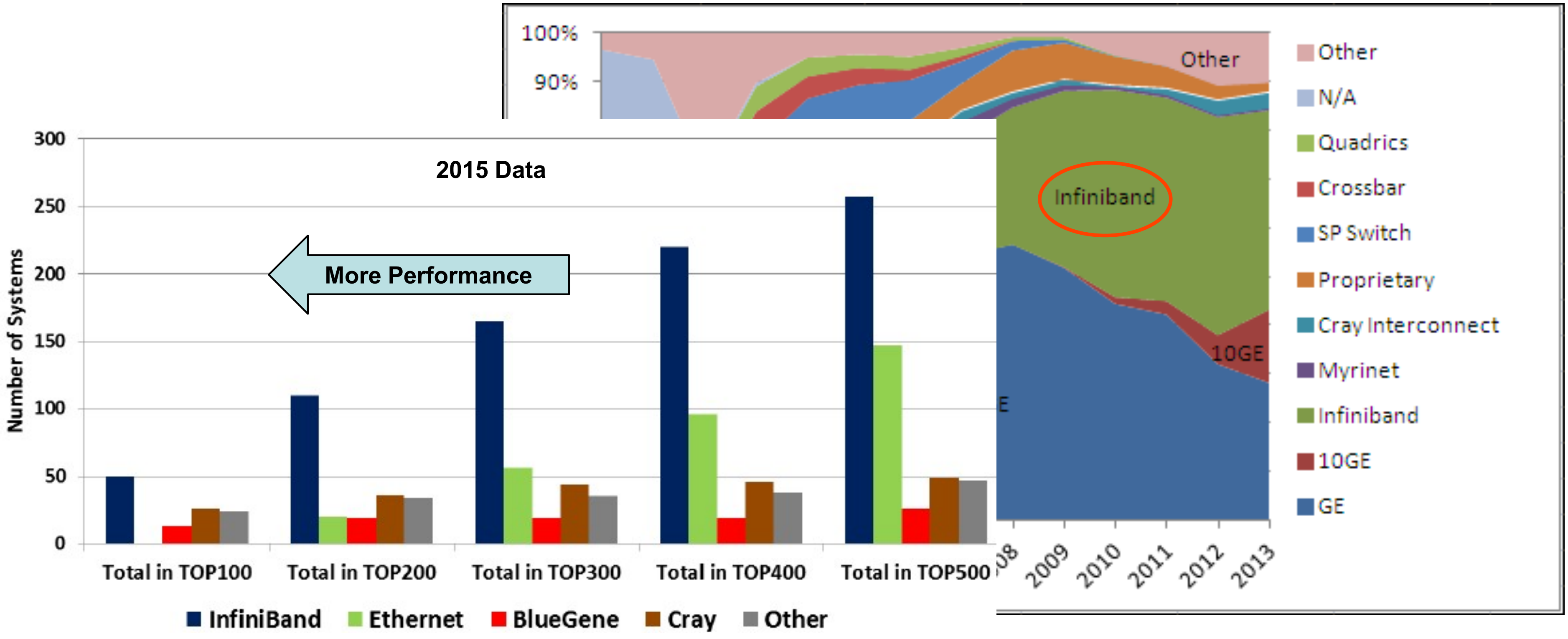
Rob Davis

VP of Storage Technology, Mellanox

# What is RDMA?

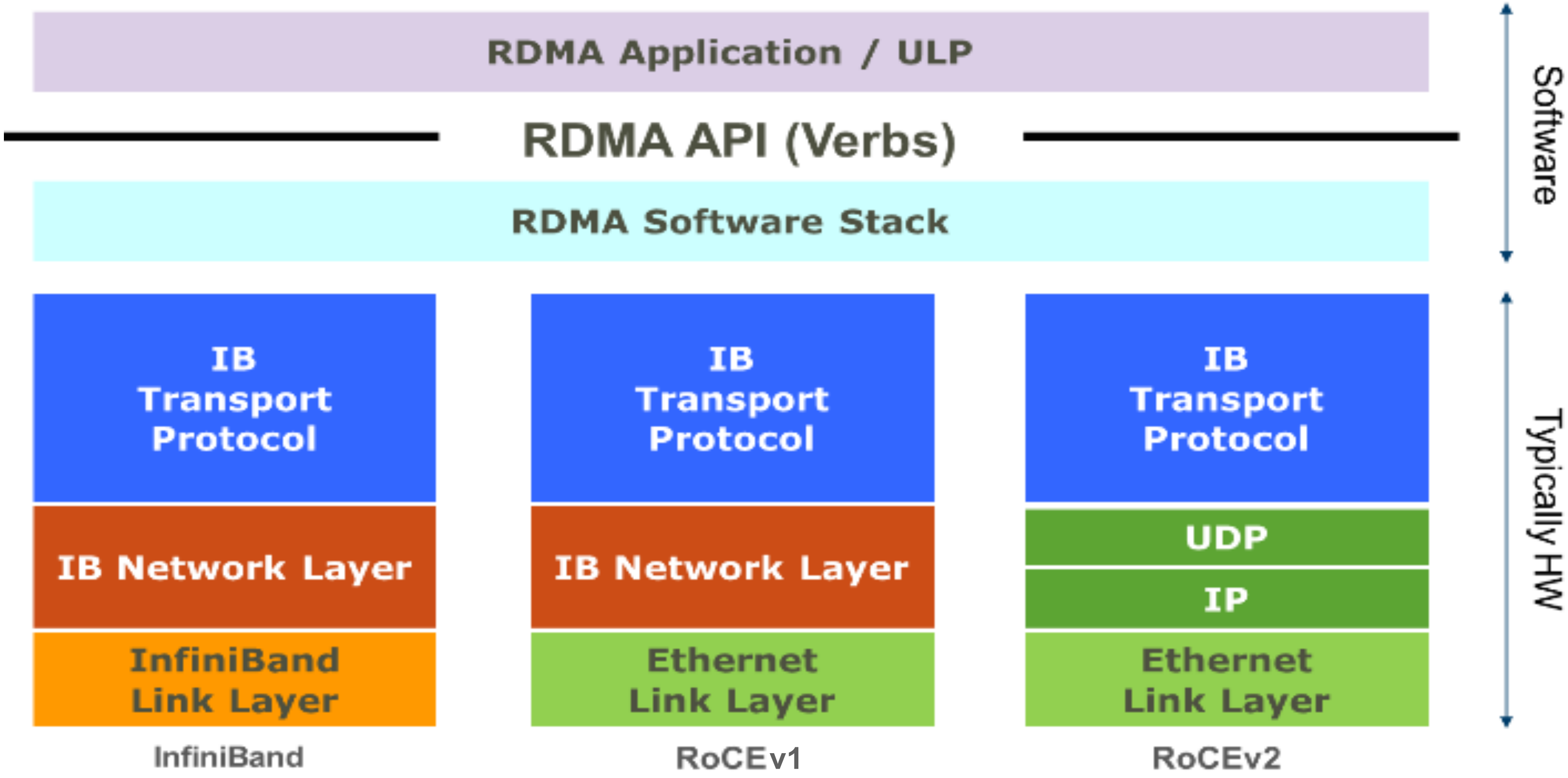

Efficient Data Movement (RDMA)
Application
Buffer
Network
Application
Buffer
Kernel Bypass
Protocol Offload

# RDMA barrowed from HPC

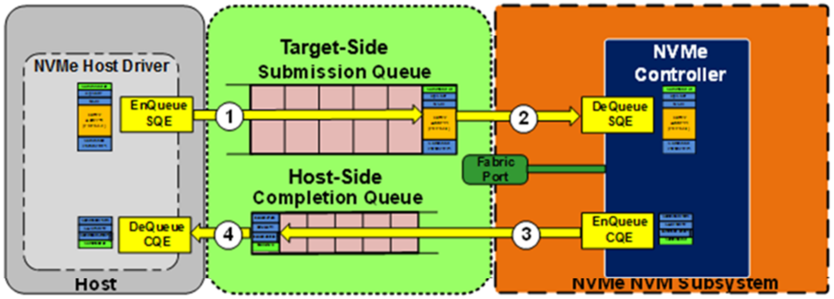


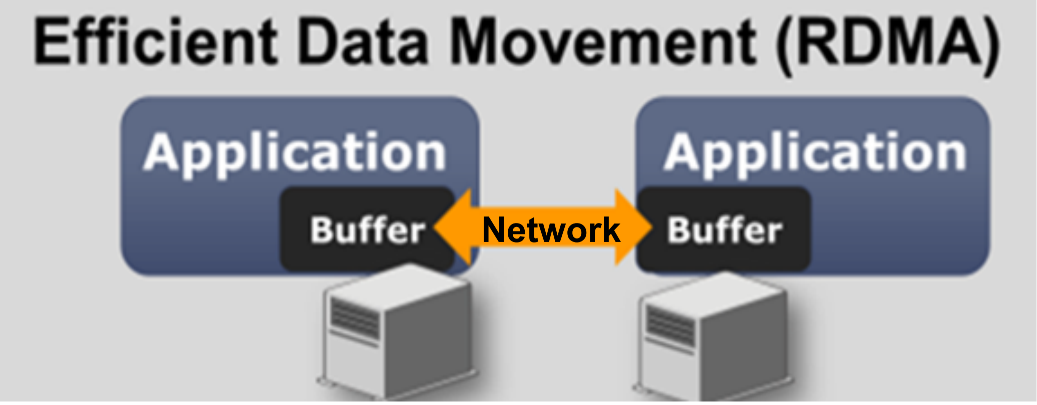
Flash Memory Summit

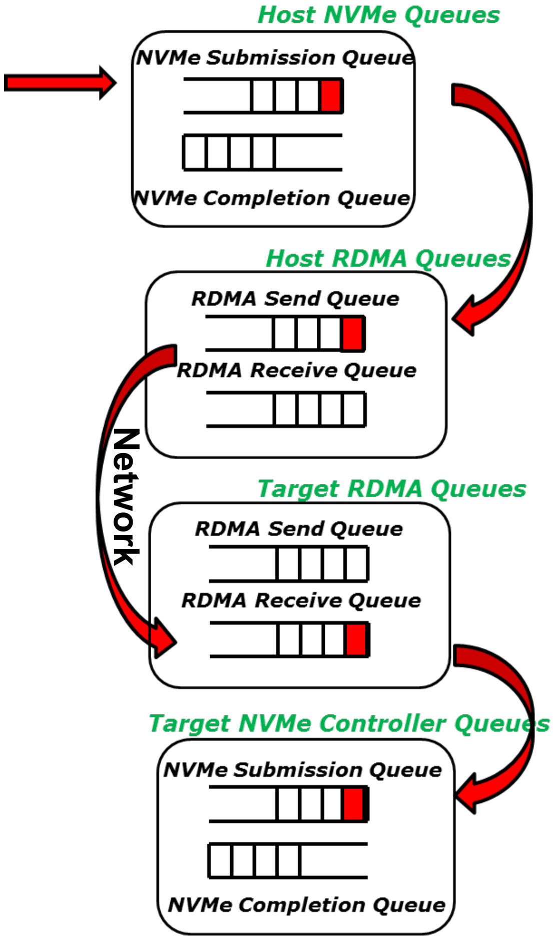
# RoCE and IB Protocol



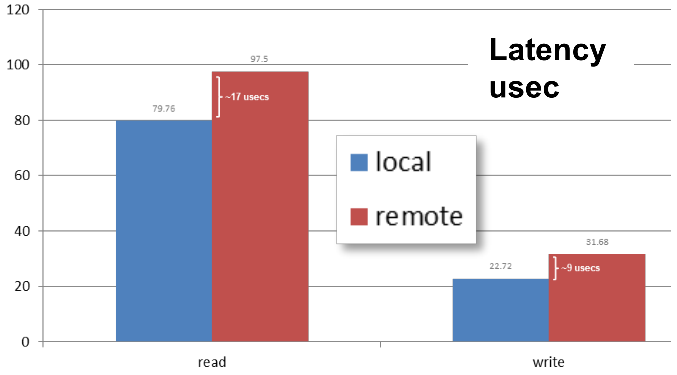
# RDMA for NVMe-oF

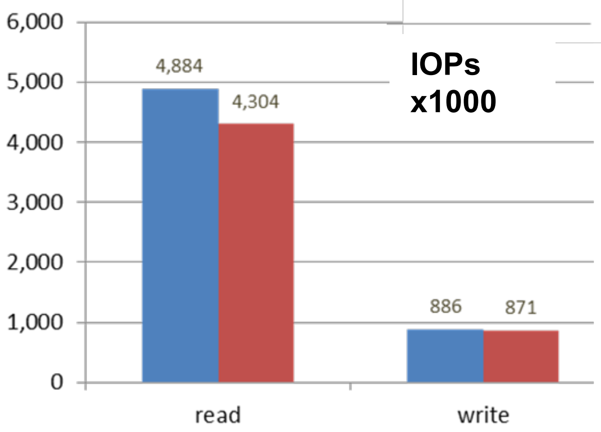


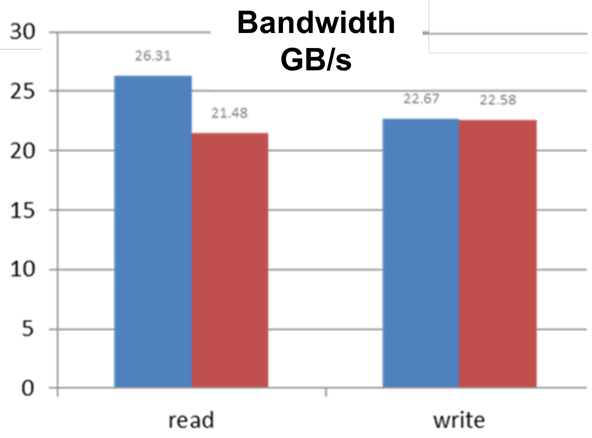




# NVMe-oF RoCE Performance







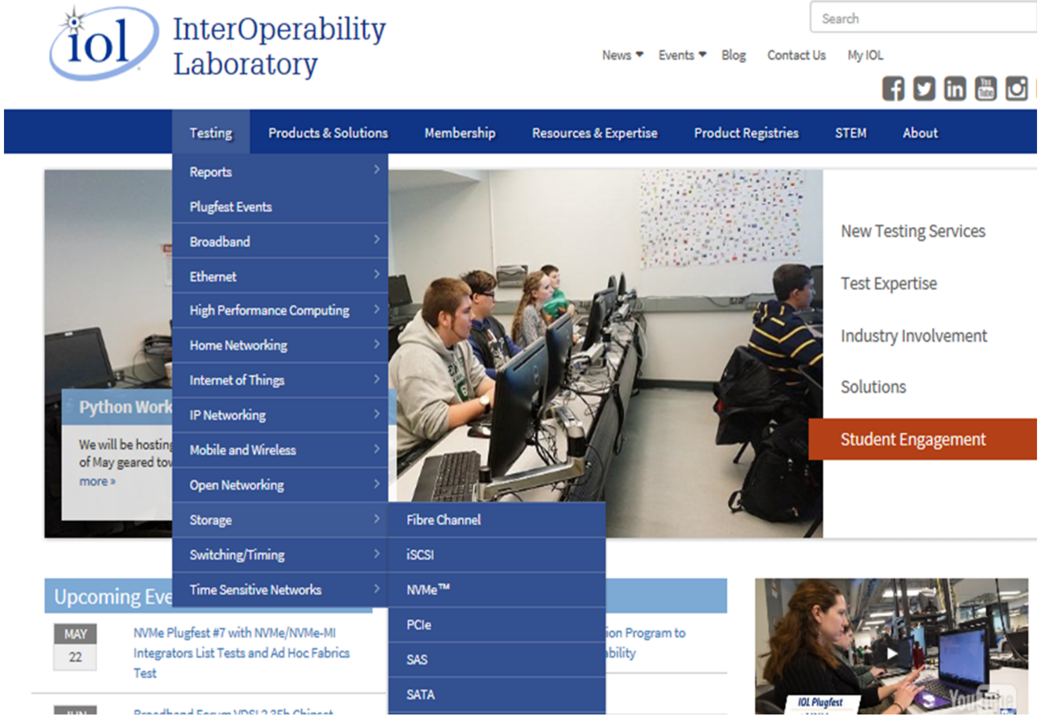
| Performance Delta | | 1-drive | 24-drive |
|---|---|---|---|
| Latency | Read | 11% | 15% |
| | Write | On par | On par |
| IOPS | Read | 10% | 12% |
| | Write | On par | 2% |
| Throughput | Read | On par | 18% |
| | Write | On par | On par |

# RoCE Demos & Production

- FMS Demos
  - E8, Micron, Celestica, Toshiba, Samsung, Mellanox, IBM, Kaminario, Excelero, MicroSemi, Newisys/Sanmina, Seagate/AIC, others
- Announced or Shipping Products
  - Huawei, Pure, Supermicro, Micron, AIC, Echostream, Inventec, E8, Liqid, Excelero, Newisys , Pavilion, others
- Reference Designs
  - Samsung, Seagate, Micron, Toshiba, others

# UNH IOL Multivendor RoCE NVMe-oF Interoperability Test

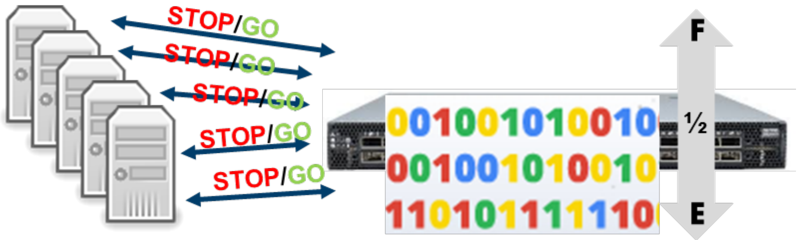
- UNH-IOL provides a neutral environment for multi-vendor interoperability and conformance to standards testing since 1988
- This May hosted the first test for NVMe-oF
- Test was organized to coincide with the regularly scheduled bi-yearly NVMe testing to leverage the SSD expertize already on site
- Test plan called for participating vendors to mix and match their NICs in both Target and Initiator positions
- Testing was completely successful with near line rate performance at 25Gb/s also achieved



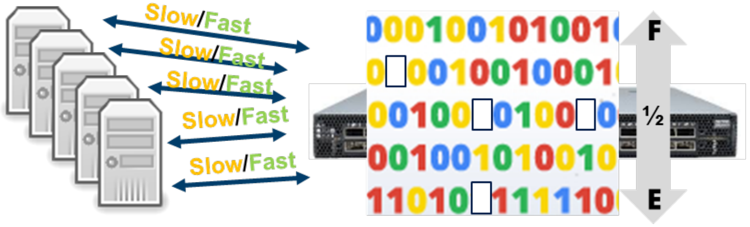
# Congestion and Network Performance Management

- Attention to congestion and data path quality are essential to maintain peak performance with RDMA on Ethernet
- Some of today's RoCE products require a lossless network implemented through PFC(IEEE Priority Flow Control)
- Some can also use ECN(IETF Explicit Congestion Notification) or both

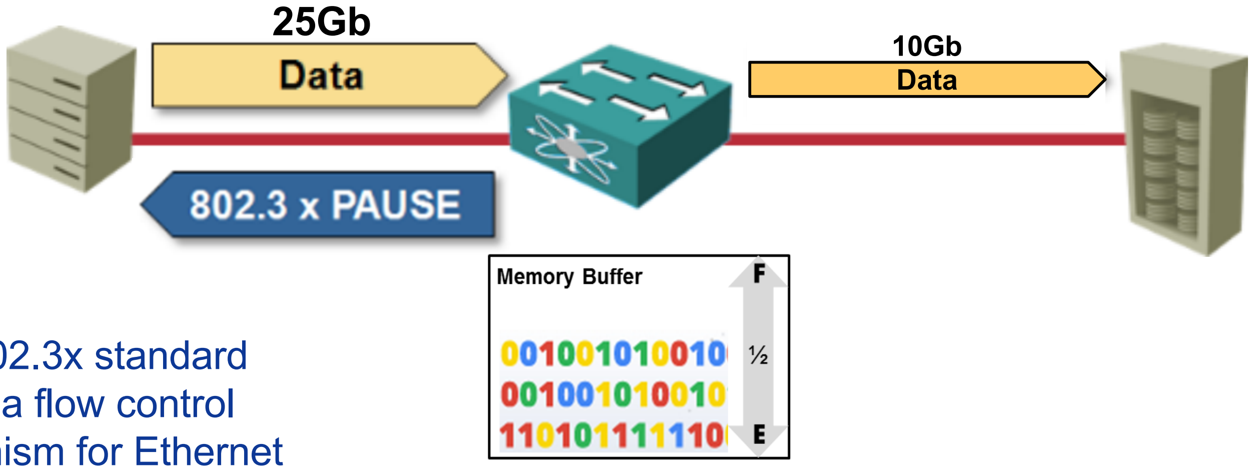
Lossless PFC



Lossy and/or Congested with ECN



# Pause Frame

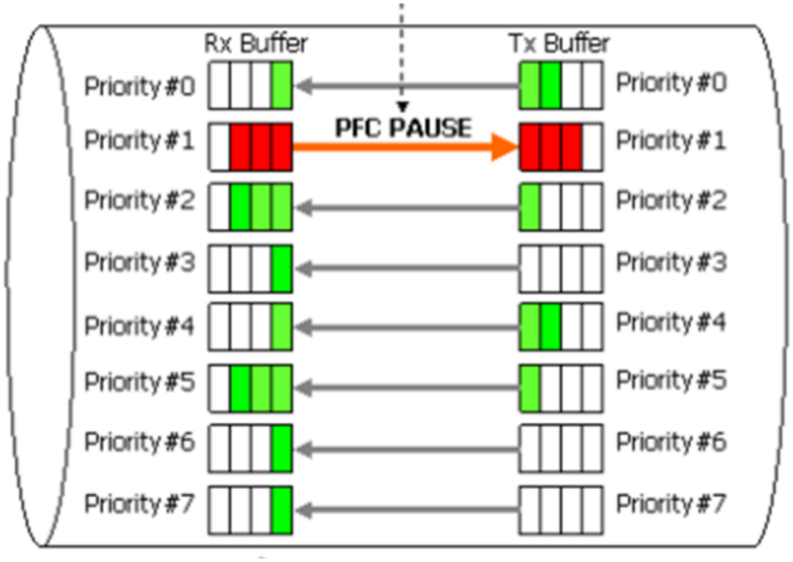


IEEE 802.3x standard defines a flow control mechanism for Ethernet called the pause frame

# Priority Flow Control

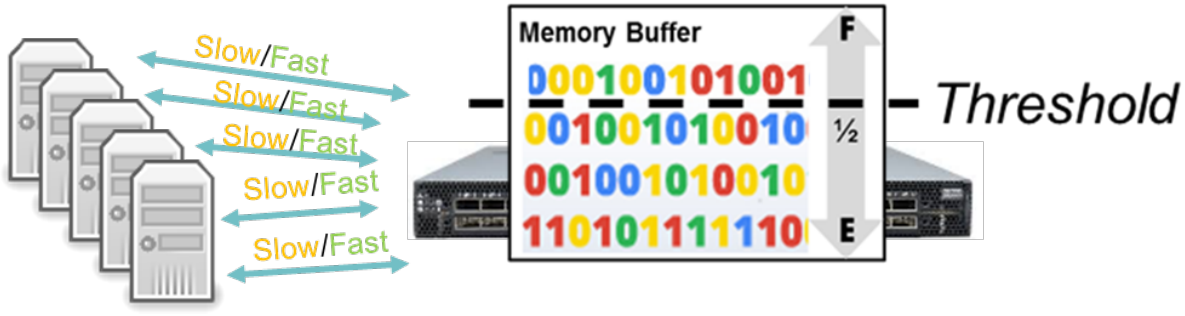
Priority Flow Control (PFC) is similar to 802.3x Pause, except seven priority levels are added. When the data in any of the eight buffers gets to a certain level a pause is sent causing the upstream device to stop sending data only for that priority level for a specified amount of time.

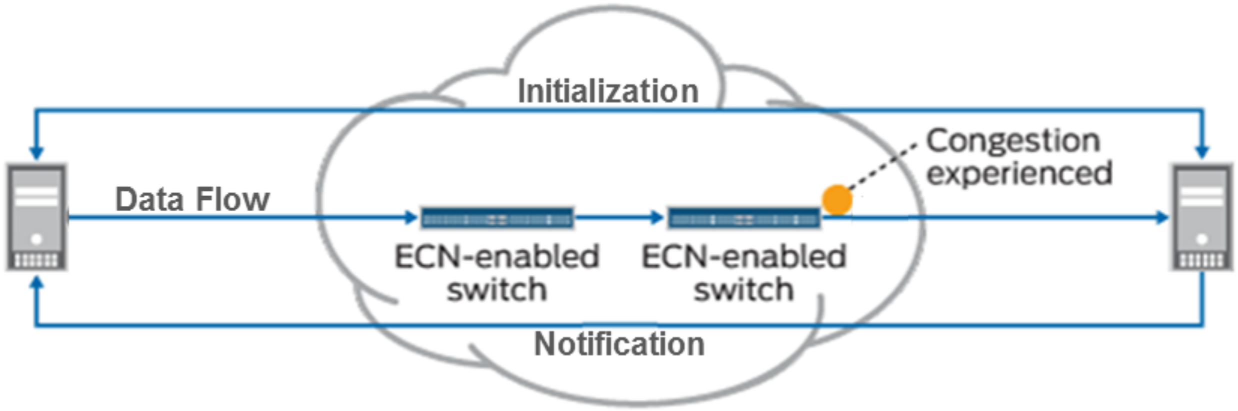
802.1Qbb - Priority-based Flow Control



# Explicit Congestion Notification

RFC 3168 Explicit Congestion Notification (ECN) slows down a explicit device's data rate that is believed to be overflowing another devices buffer.





# NVMe-oF Update

Praveen Midha

Cavium

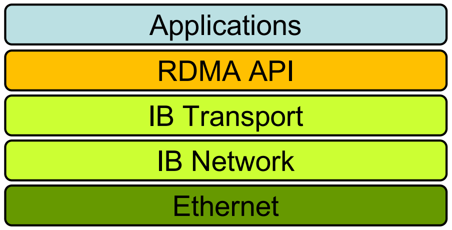
# RDMA Scalability Comparison

**RoCE**

"Neighborhood Scale"

**RDMA over IB (IBoE)**

- Not routable
- Requires DCB
  - P2P Flow control



Applications

RDMA API

IB Transport

IB Network

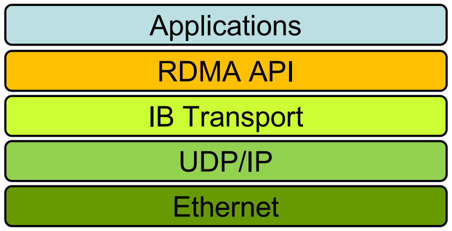
Ethernet

**RoCEv2**

"Subdivision Scale"

**RDMA over IB over UDP**

- Adds routability
- Requires DCB
  - P2P Flow Control
- **DCQCN capable**
  - Congestion mgmt.
  - Requires PFC



Applications

RDMA API

IB Transport

UDP/IP

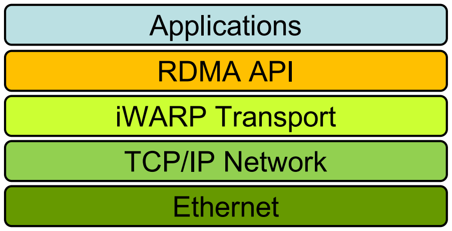
Ethernet

**iWARP**

"Metropolitan Scale

**RDMA over TCP/IP**

- Fully routable
- E2E Flow Control with TCP
  - DCB not required
- Congestion *Avoidance*



Applications

RDMA API

iWARP Transport

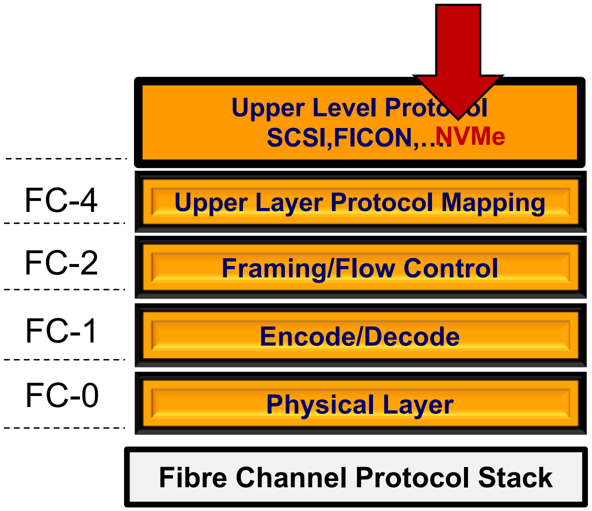
TCP/IP Network

Ethernet

# RDMA – What to Choose When?

- Ecosystem readiness
  - SW - Majority of OSs and applications support both iWARP & RoCE
  - HW – RoCE: BRCM/CAVM/MLNX; iWARP: CAVM, INTC, Chelsio
- iWARP leads in ease of deployment
  - RDMA traffic can span large-scale networks w/o special configuration
  - Packet loss has the potential to cause increase in latency
- RoCE delivers superior performance when properly deployed
  - Lossless Ethernet network results in deterministic latency
  - Bounded latency delivers maximum performance for storage applications
  - But requires network admin to configure switches for VLANs and PFC
  - Best suited to small-scale environments

# FC-NVMe update



- FC-NVMe standard (T11) progressing well
  - Spec in letter ballot – Rev 1.0 ETA Aug 2017
  - Enhanced error recovery in follow-on spec
- Linux community update:
  - FC-NVMe transport support now available in Linux 4.12 kernel
  - Host & Target drivers in various stages of upstream submission
- End-to-End FC-NVMe POC
  - Pre-GA software available - Initiator and Target mode
  - FC Switch support available

# NVMe over Fibre Channel

Curt Beckmann
Principal Architect
Brocade

# Presentation Topics

- FC-NVMe Spec and Interoperability Update
- Dual Protocol SANs boost NVMe adoption
- Enterprise Storage Vendor Demo!

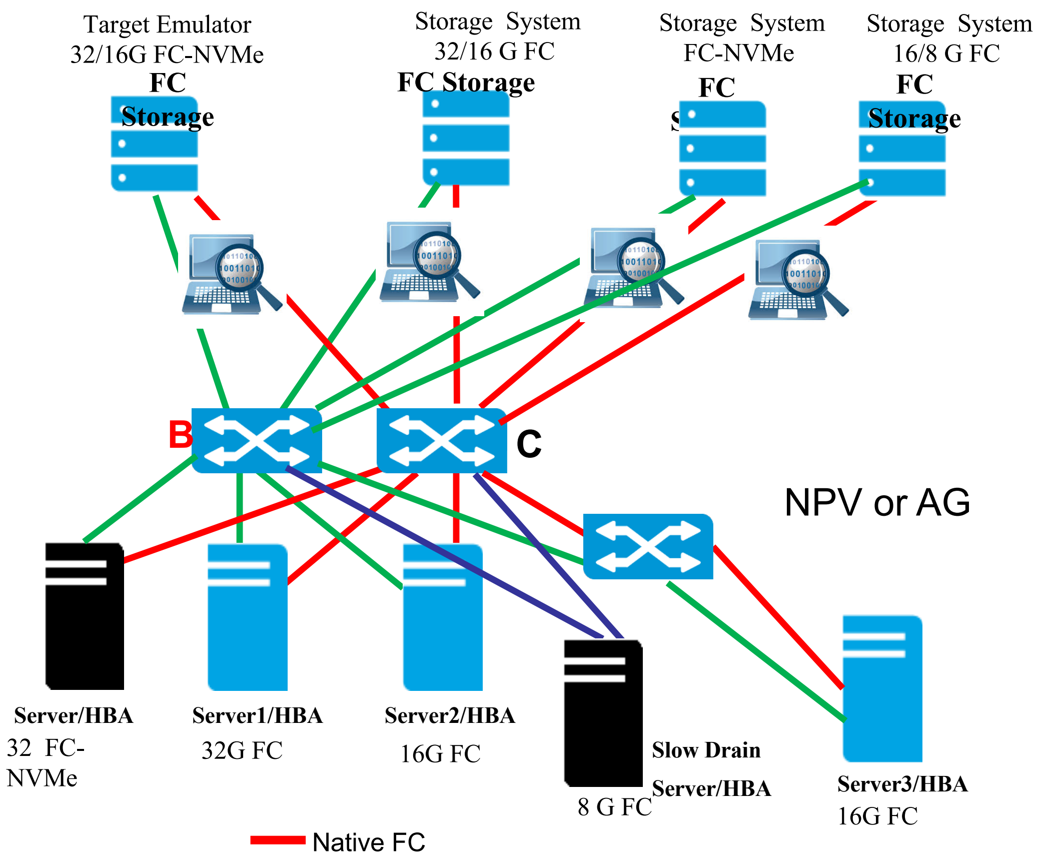
# Presentation Topics

- FC-NVMe Spec and Interoperability Update
- Dual Protocol SANs boost NVMe adoption
- Enterprise Storage Vendor Demo!

# FC-NVMe Spec Status

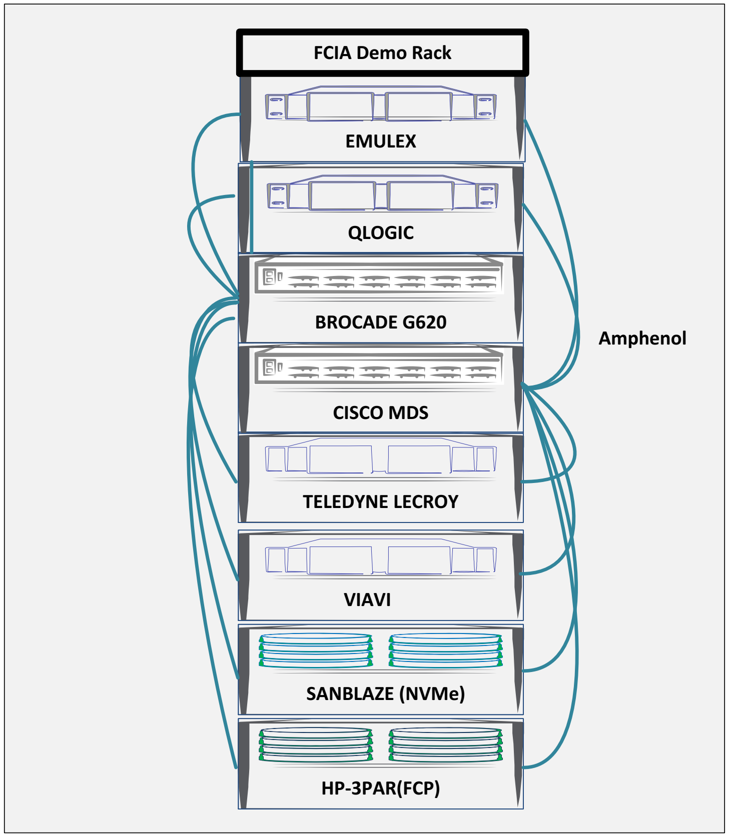
- T11 meeting happening right now
  - Spec stable: T11 to send to INCITS Sept/Oct
- UNH Plugfest in June
  - 12 vendors participated
  - Next UNH plugfest will be in October

# UNH Test Track 32/16/8G FCP & FC-NVMe Redundant Fabric / Availability; Large Fabric – connecting all participating devices



# FCIA FMS FC-NVMe Demo Rack

- 8 Vendors showing interoperability
- Live Demo (hosted remotely at Brocade)



© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

# Presentation Topics

- FC-NVMe Spec and Interoperability Update
- Dual Protocol SANs boost NVMe adoption
- Enterprise Storage Vendor Demo!

# Dual Protocol SANs boost NVMe adoption

- 80% of Flash arrays connect via Fibre Channel
  - These Flash arrays house high-value data assets

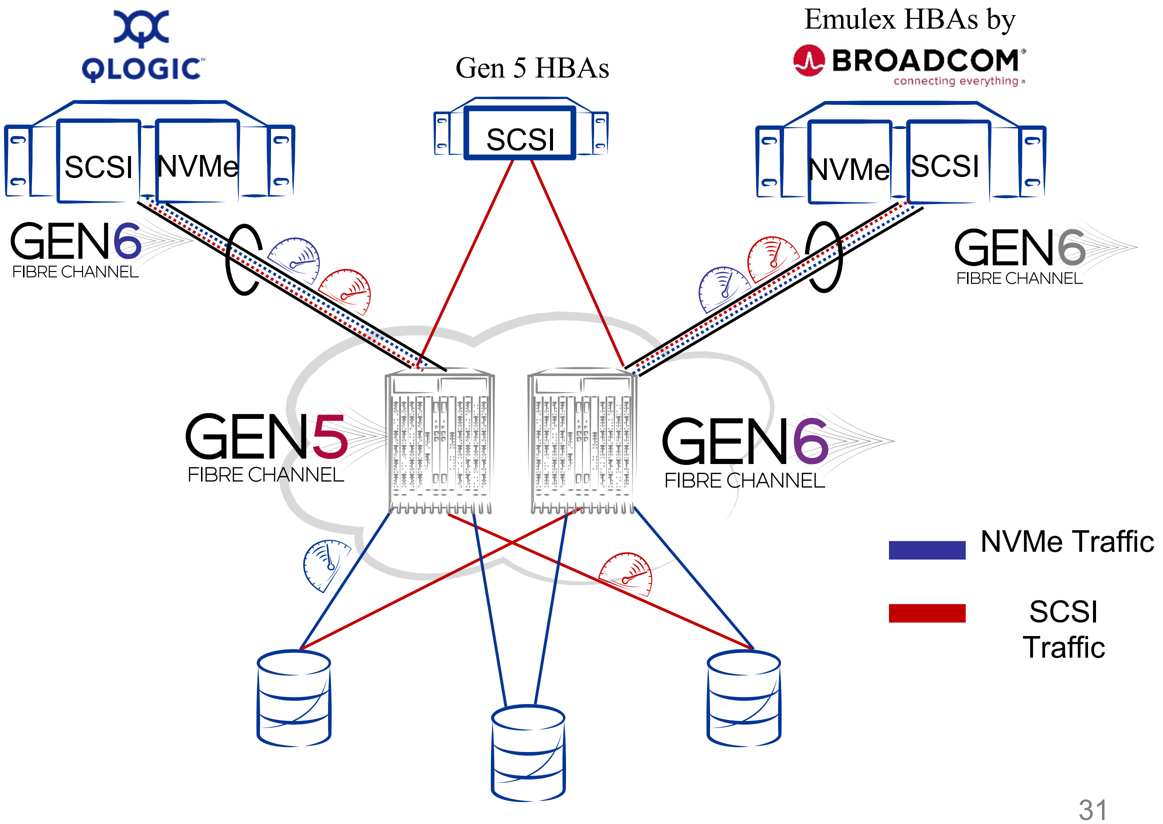
- High-value Assets require protection
  - Storage Teams are naturally risk averse
  - Technology planning ranks risk avoidance highly

# Dual Protocol SANs Reduce Risk

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

# Dual protocol SANs enable low risk NVMe adoption

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

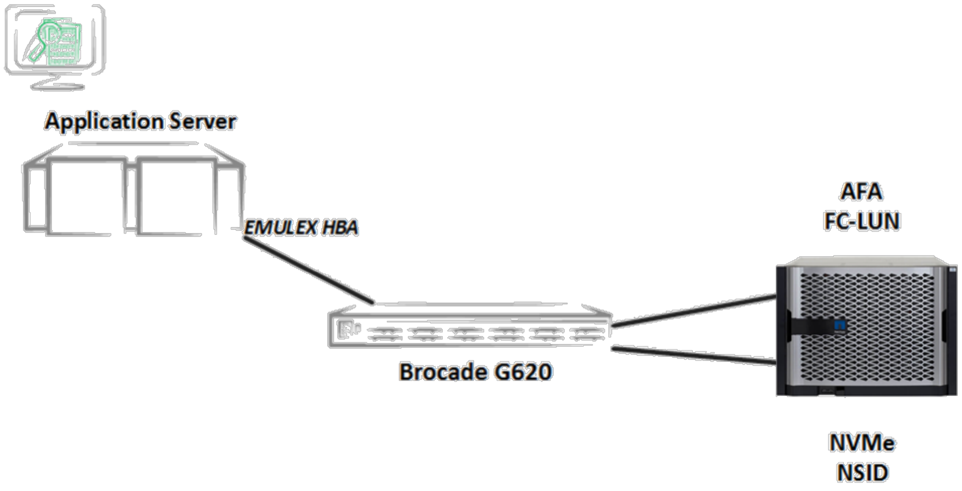


# Presentation Topics

- FC-NVMe Spec and Interoperability Update
- Dual Protocol SANs boost NVMe adoption
- Enterprise Storage Vendor Demo!

# NetApp’s FMS FCP/FC-NVMe Demo

- NetApp Storage Array
  - 32G FC connectivity
  - Presents both NVMe namespace and SCSI LUN to the FC fabric
- Application server
  - Emulex HBA
  - SUSE Linux
  - Can mount and read/write to both namespace and LUN
- Brocade G620
  - Runs FC-NVMe and FCP (i.e. SCSI) traffic simultaneously



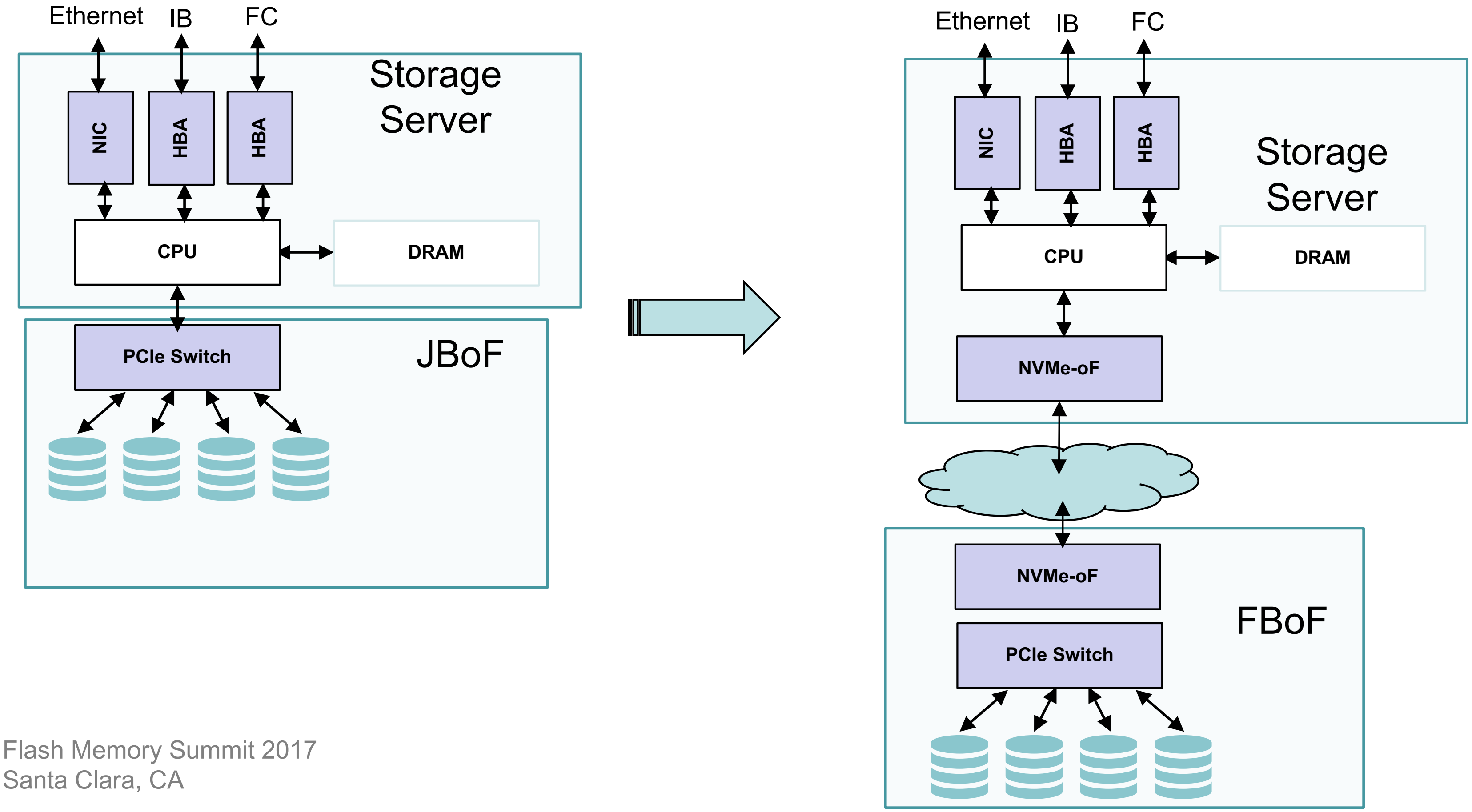
© 2017 BROCADE COMMUNICATIONS SYSTEMS, INC.

# SSD Disaggregation and Scaling with NVMe-OF

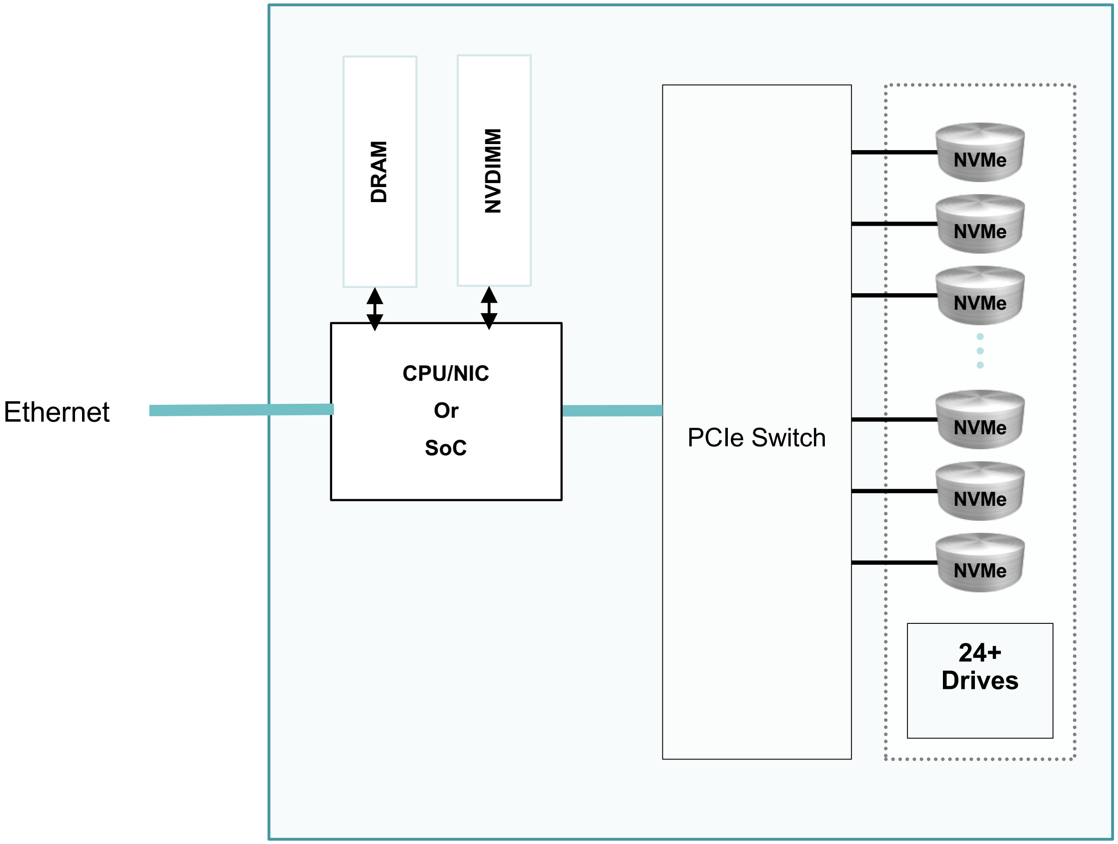
Fazil Osman

Broadcom Limited

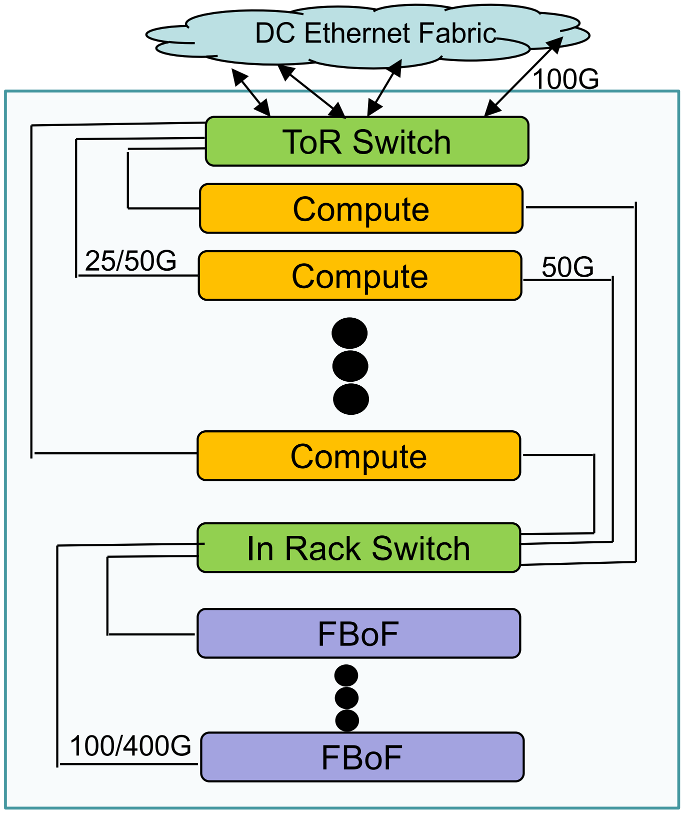
# NVMe-oF decouples SSDs from Server



# Typical FBoF Design



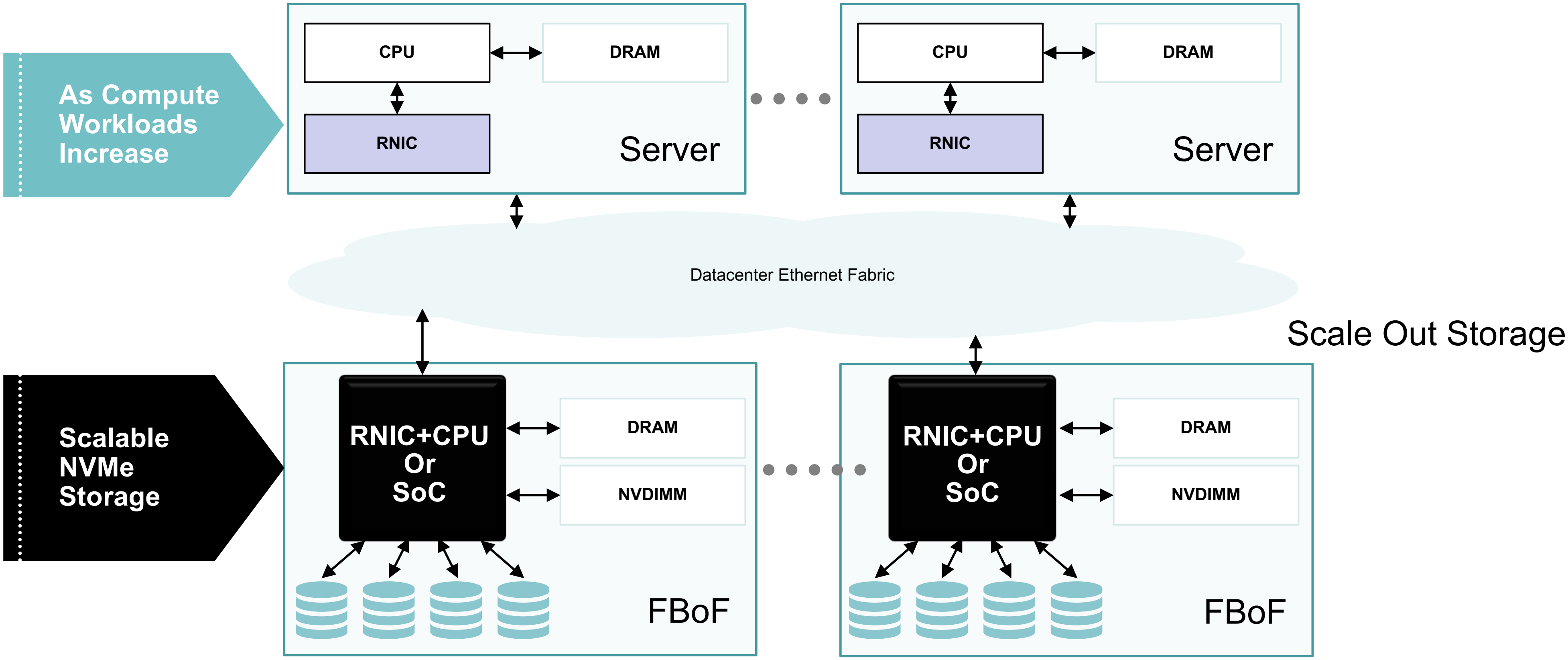
# Rack level disaggregation



- NVMe drives in FBoFs can be provisioned to compute nodes to match needs:
  - Capacity
  - IOPs
  - Bandwidth

Flash Memory Summit

# Datacenter wide disaggregation



# NVMe-oF™ NVMe-TCP Transport

Dave Minturn

Intel Principal Engineer

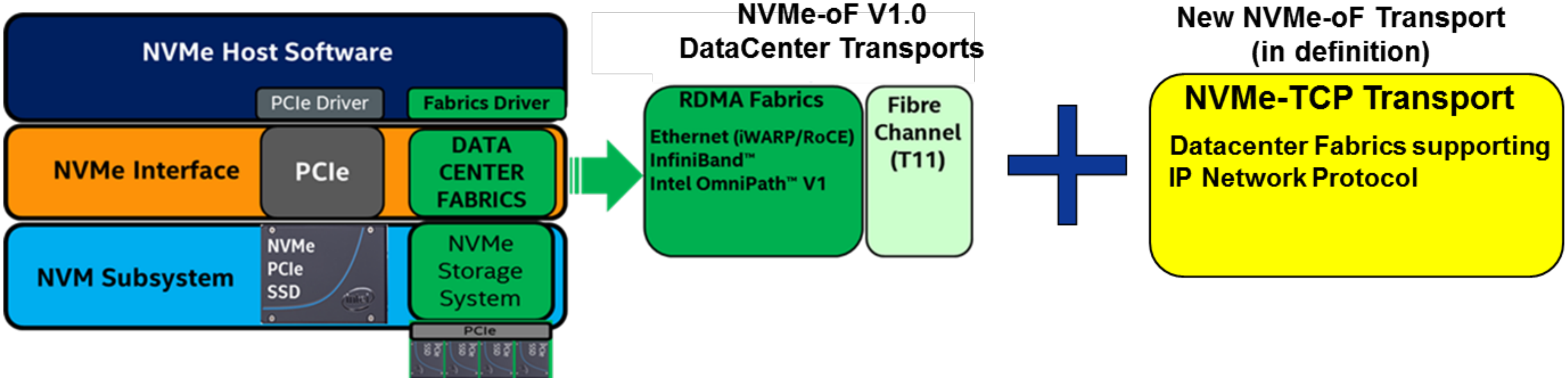
# NVMe-oF coming to a network near you

NVMe-oF V1.0 enabled efficient end-2-end NVMe over RDMA and Fibre Channel networks

- RDMA because of it's high efficiency and similar architecture characteristics
- FC because of it's reliable credit based flow control and delivery mechanism

What about existing IP network infrastructures?

# NVMe-TCP Transport



NVMe Host Software

PCIe Driver

Fabrics Driver

NVMe Interface

PCIe

DATA CENTER FABRICS

NVM Subsystem

NVMe PCIe SSD

NVMe Storage System

PCIe

NVMe-oF V1.0 DataCenter Transports

RDMA Fabrics

Ethernet (iWARP/RoCE)
InfiniBand™
Intel OmniPath™ V1

Fibre Channel (T11)

New NVMe-oF Transport (in definition)

NVMe-TCP Transport

Datacenter Fabrics supporting IP Network Protocol

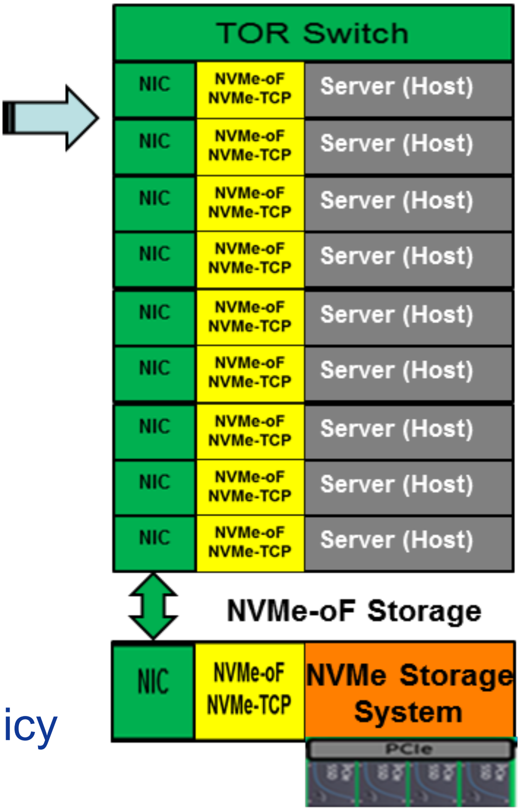
- Enables the use of NVMe-oF over existing Datacenter IP networks
- Supports all of the NVMe-oF and NVMe Architecture features
- Layered over standard IETF TCP transport to allow software-only and/or hardware (accelerated/offloaded) implementations

# 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 eliminating SCSI to NVMe translations
- Co-exists with other NVMe-oF transports
  - Transport selection may be based on h/w support and/or policy

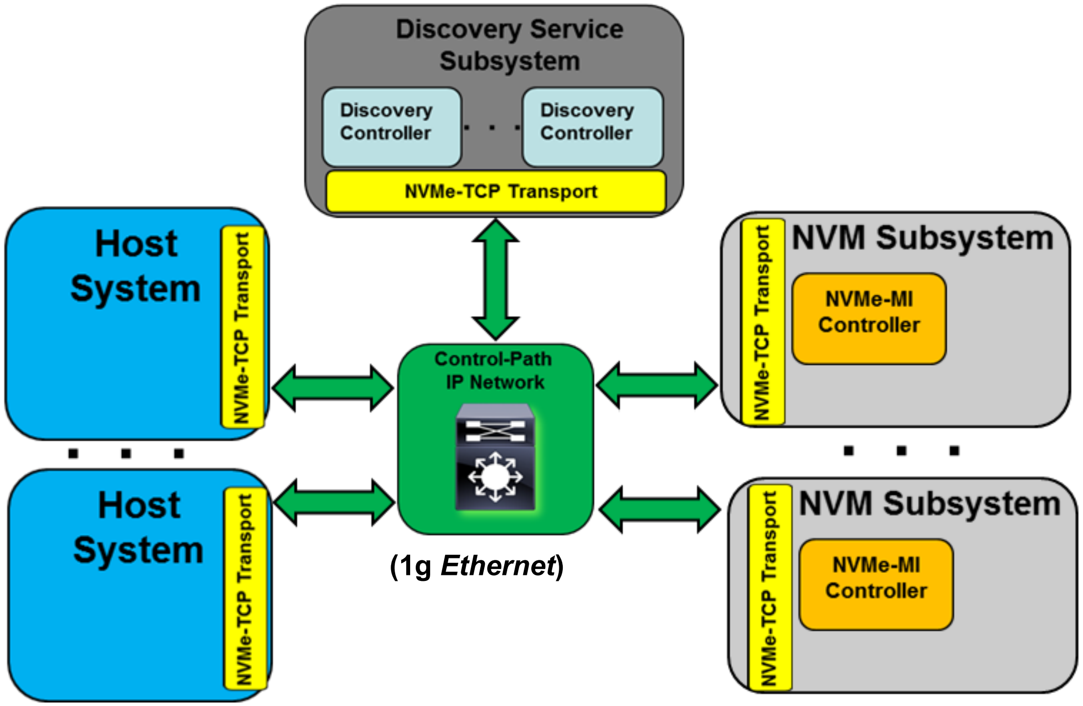
Existing Datacenter

Existing Rack H/W
(NVMe Host Driver with NVMe-TCP)



# 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



# NVMe-TCP Status

- Currently in definition within the NVMe.org Technical Working Group
- Linux Host and Target Drivers being developed in the NVMe.org Fabric Driver Working Group
- Plan to co-release specification and tested Linux drivers as part of NVMe-oF(next) release

nvm EXPRESS®

Architected for Performance

# Speaker Bios

- Add

Flash Memory Summit

# BACKUP