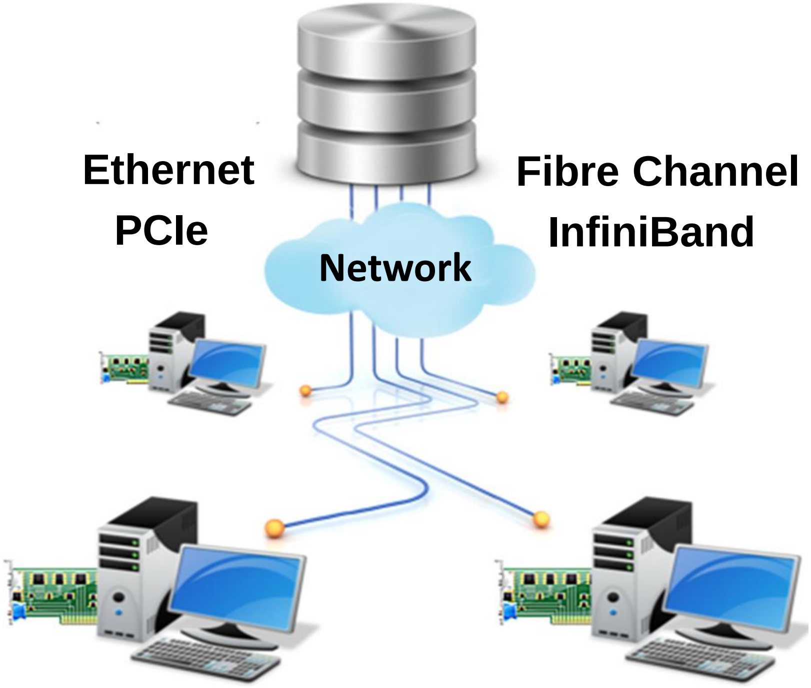
Pre-Conference Seminar D
Flash Storage Networking

Rob Davis, Ilker Cebeli, J Metz, Motti Beck, Curt Beckmann, Peter Onufryk, and Alan Weckel

Why Network Flash Based Storage?

- There are advantages to shared storage
  - Better utilization:
    - capacity, rack space, power
  - Scalability
  - Manageability
  - Fault isolation
- Shared storage requires a Network



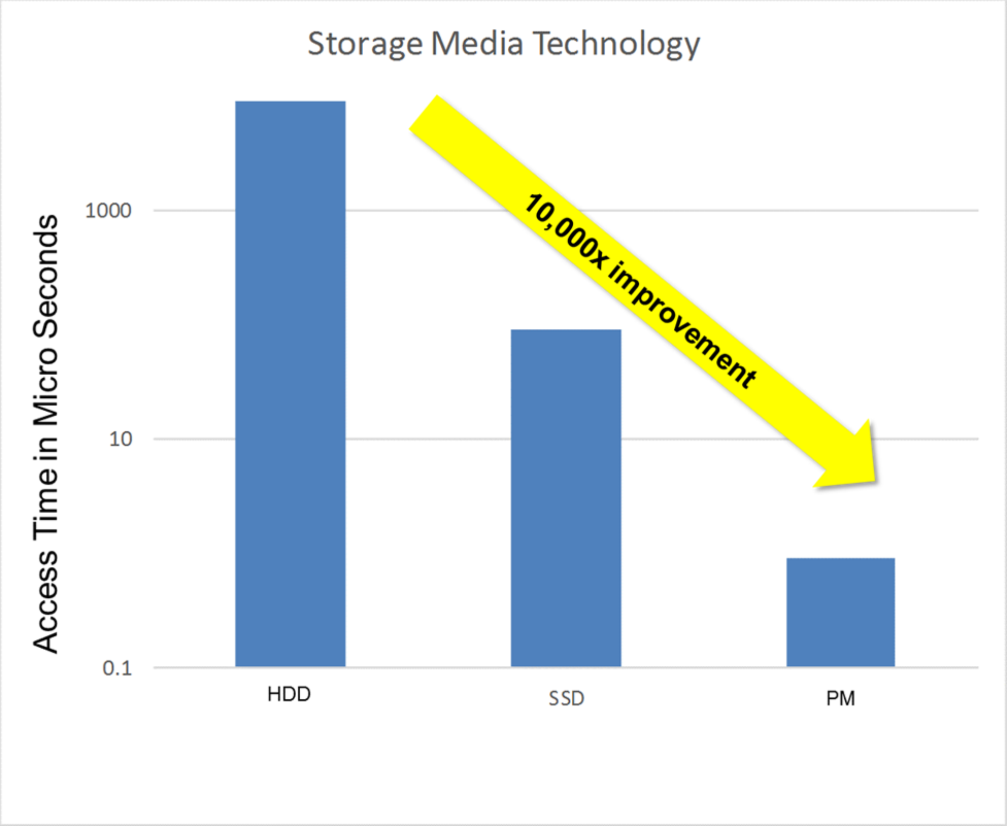
Agenda

- Networked Flash Storage Overview – 8:30 to 8:45
  - Rob Davis, Mellanox, VP Storage Technology
- ***PCIe*** Networked Flash Storage – ~8:45 to 9:05
  - Peter Onufryk, **Microsemi(Microchip)**, NVM Solutions Fellow
- ***InfiniBand*** Networked Flash Storage – ~9:05 to 9:25
  - Motti Beck, **Mellanox**, Sr. Dir. Enterprise Market Development
- ***Fibre Channel*** Networked Flash Storage – ~9:25 to 9:45
  - Curt Beckmann, Principal Product Architect, **Brocade(Broadcom)**
- ***Ethernet*** Networked Flash Storage – ~9:45 to 10:05
  - J Metz, **Cisco**, R&D Engineer, Advanced Storage, Office of the CTO, UCS Systems Group

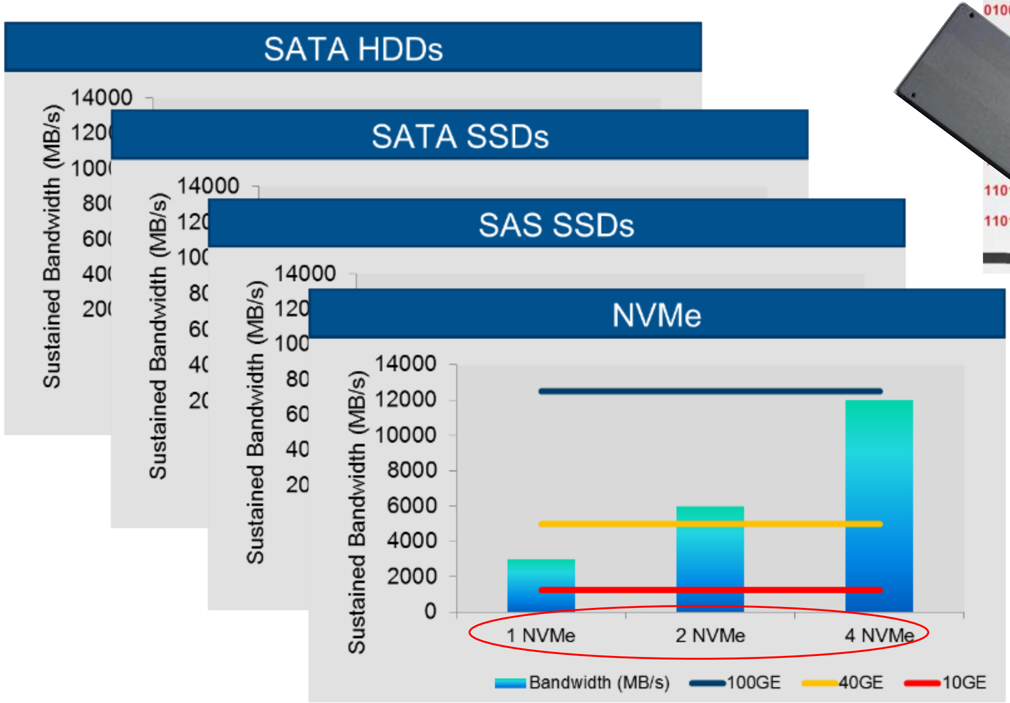
Agenda (cont.)

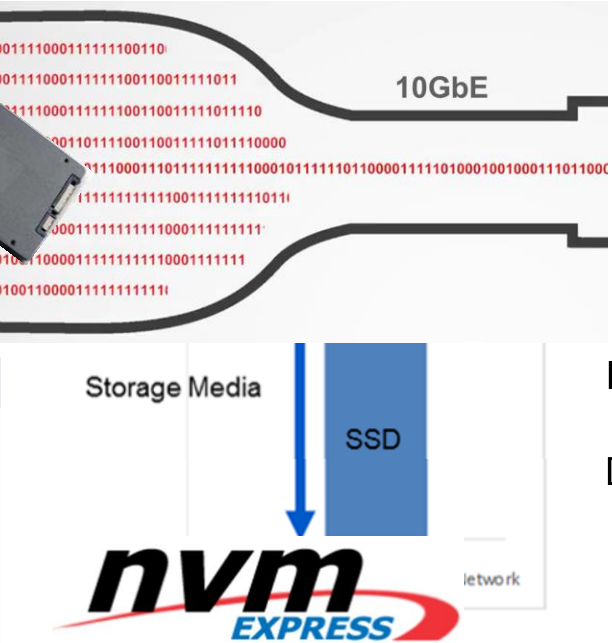
- Conference Break – 10:15 to 10:30
- How Networking Affects Flash Storage Systems – 10:30 to 10:50
  - Ilker Cebeli, **Samsung**, Sr. Dir. Product Planning
- Flash Storage Networking, How the market is evolving – ~10:50 to 11:10
  - Alan Weckel, Technology Analyst/Co-Founder at **650 Group**
- Q/A and Panel Discussion – ~11:10 to 12:00
  - All Presenters

Flash Makes Networking More Difficult



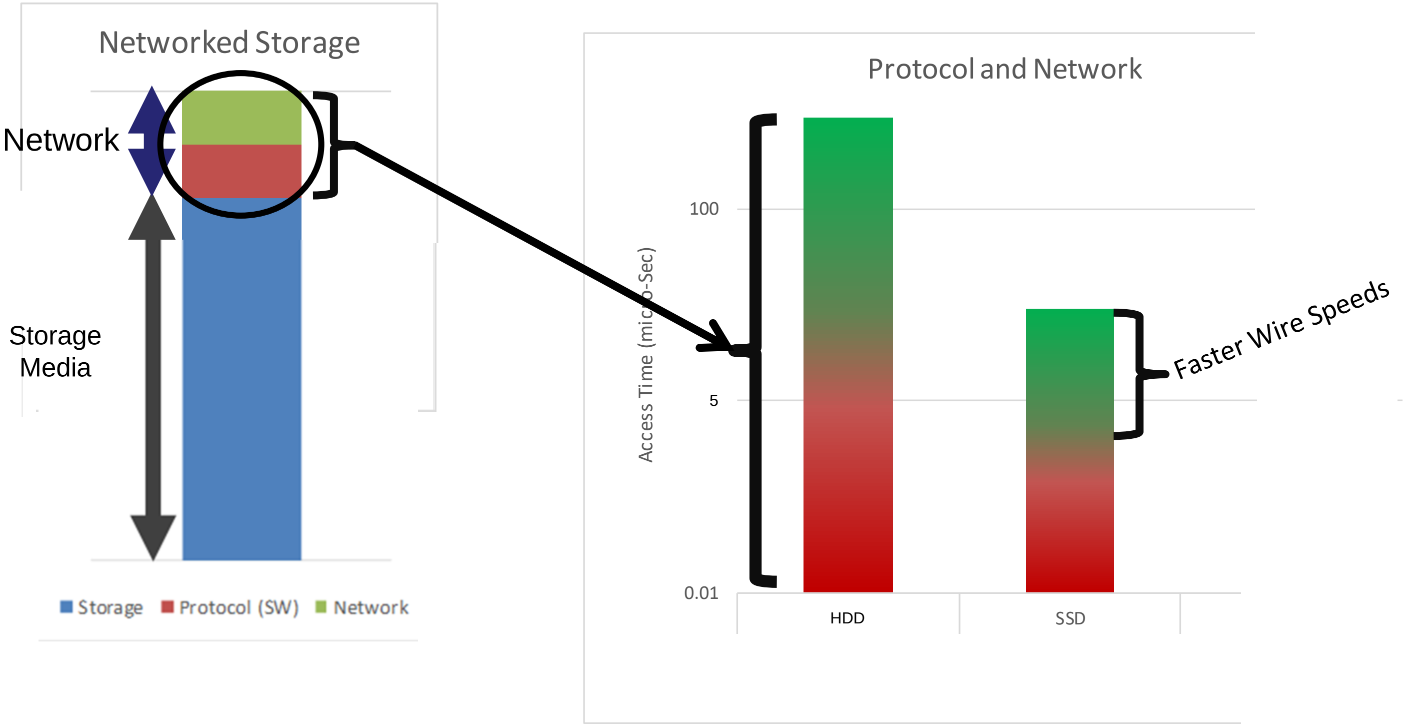
Faster Storage Needs a Faster Network



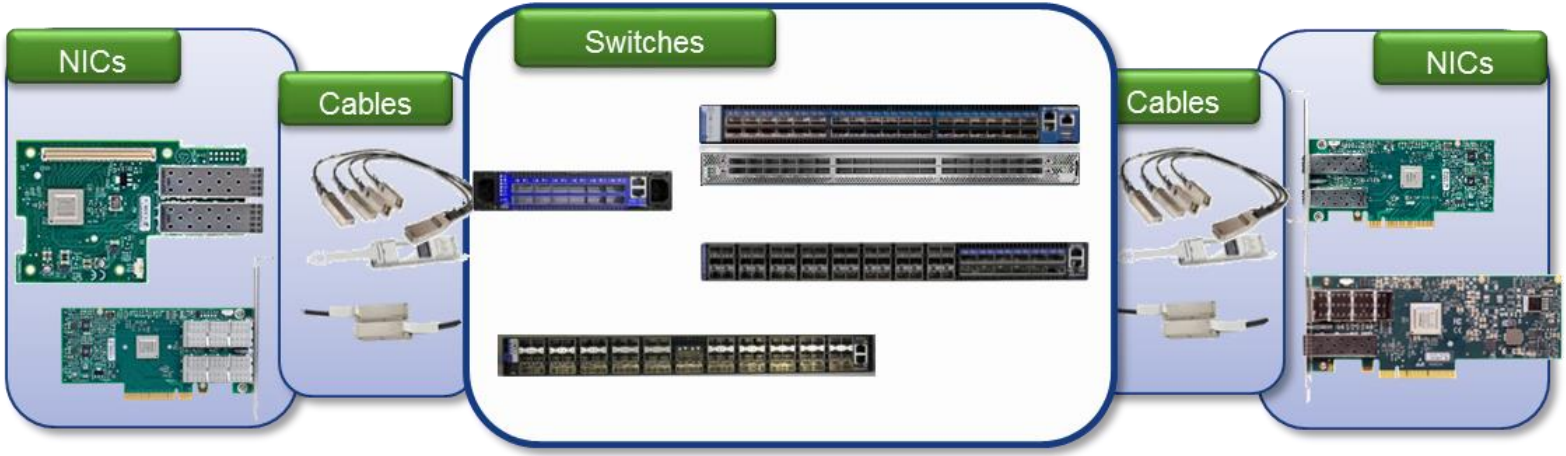


Flash SSDs move the Bottleneck from the Disk to the Network

What is the solution?

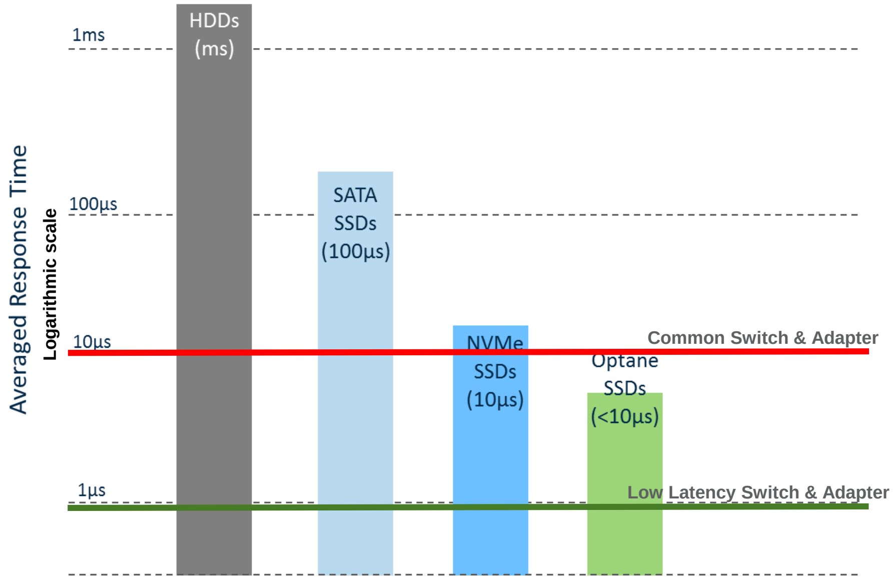


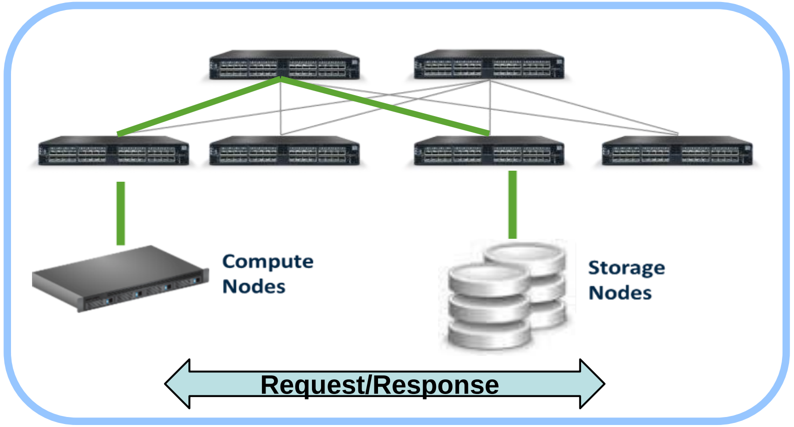
Faster Network Wires are Available



Ethernet & InfiniBand – 100Gb, going to 200 and 400Gb…
PCIe – Gen3(8Gb/lane), going to Gen4(16Gb/lane)…
FC – 32Gb, going to 128Gb…

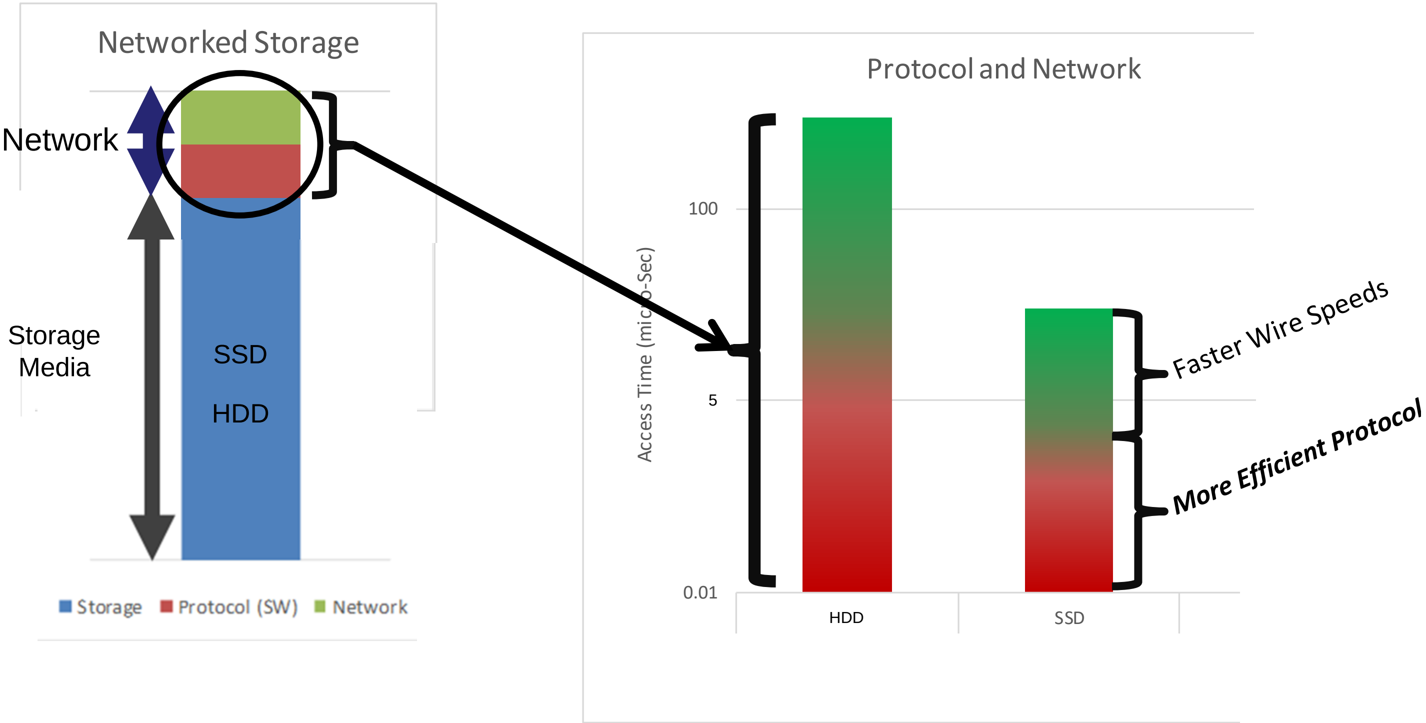
Importance of Network Latency





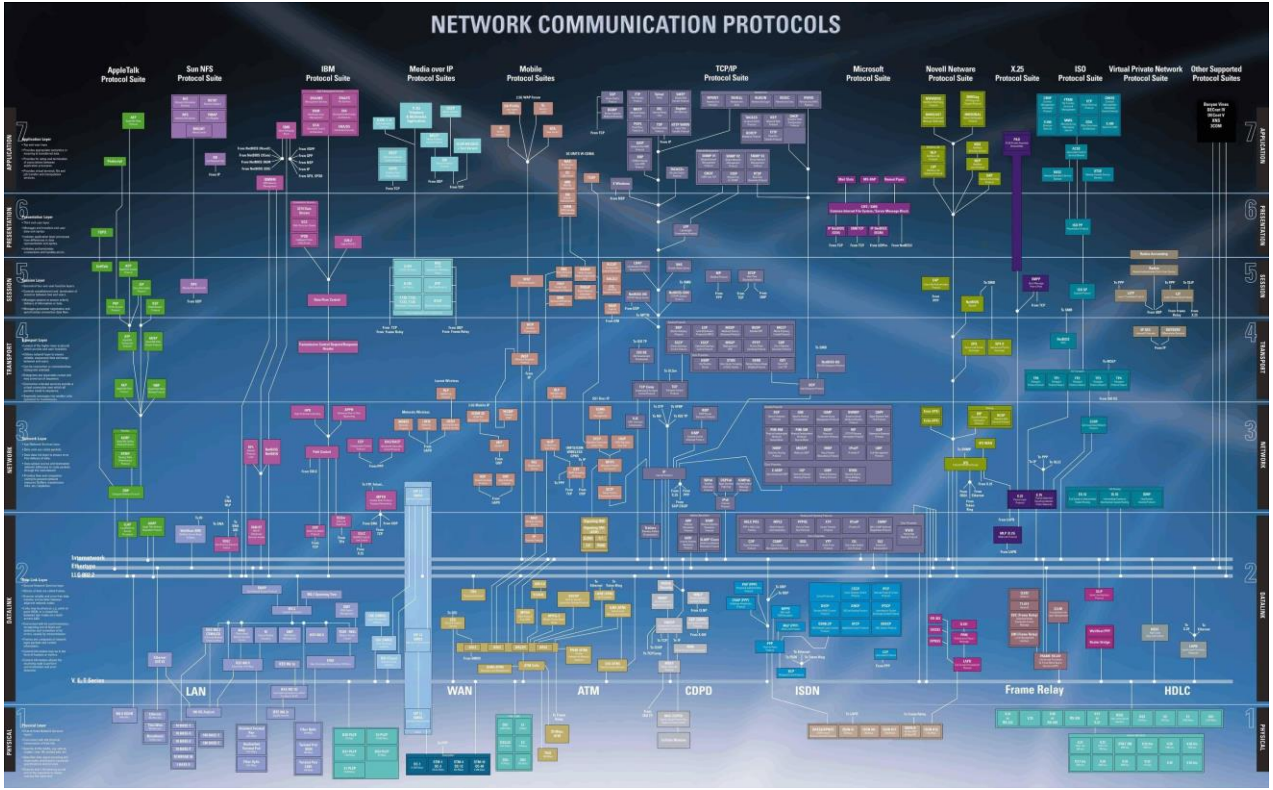
Network hops multiply latency

Faster Network Components Solves Some of the Problem…

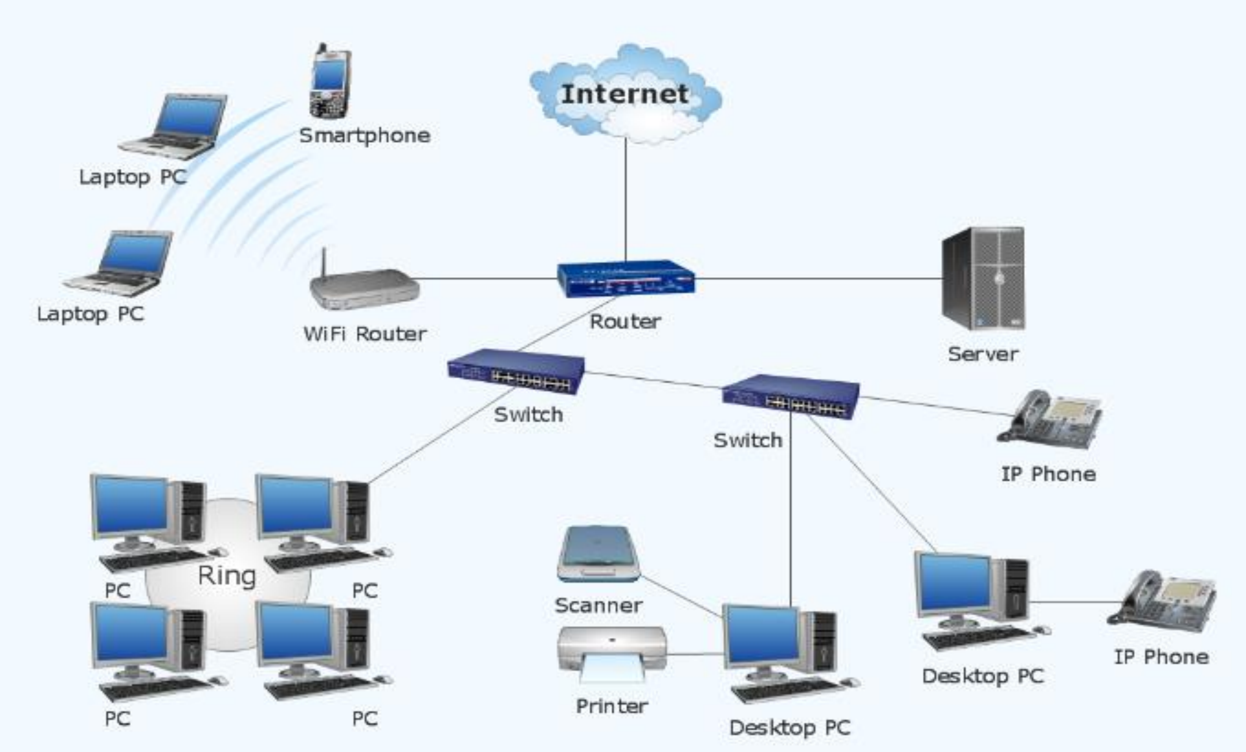


Faster Protocols

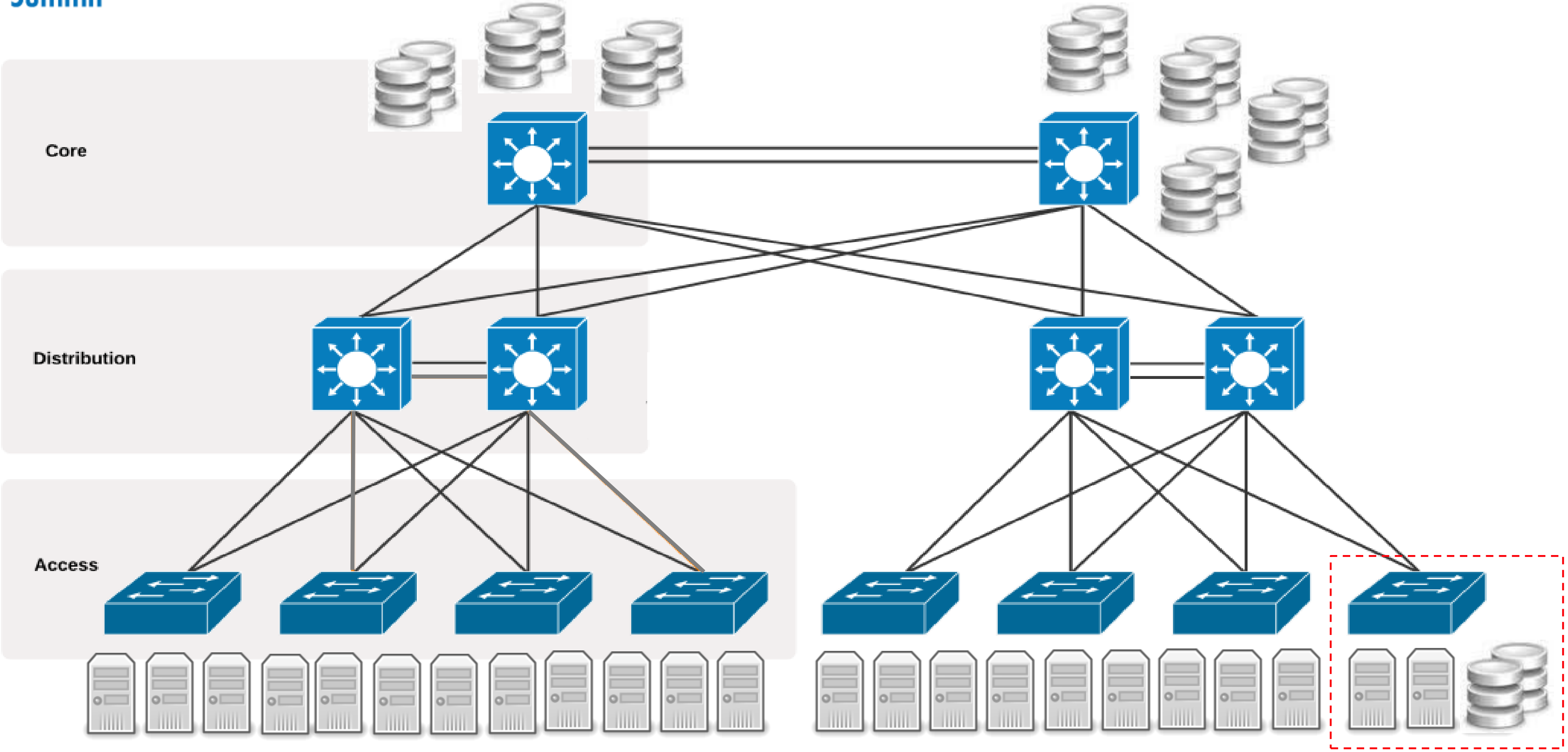
- NVMe-oF
  - RDMA(RoCE, IB)
  - Fibre Channel
  - PCIe
  - Coming soon TCP
- RDMA
  - SMB Direct
  - iSER



Where best to plug in?

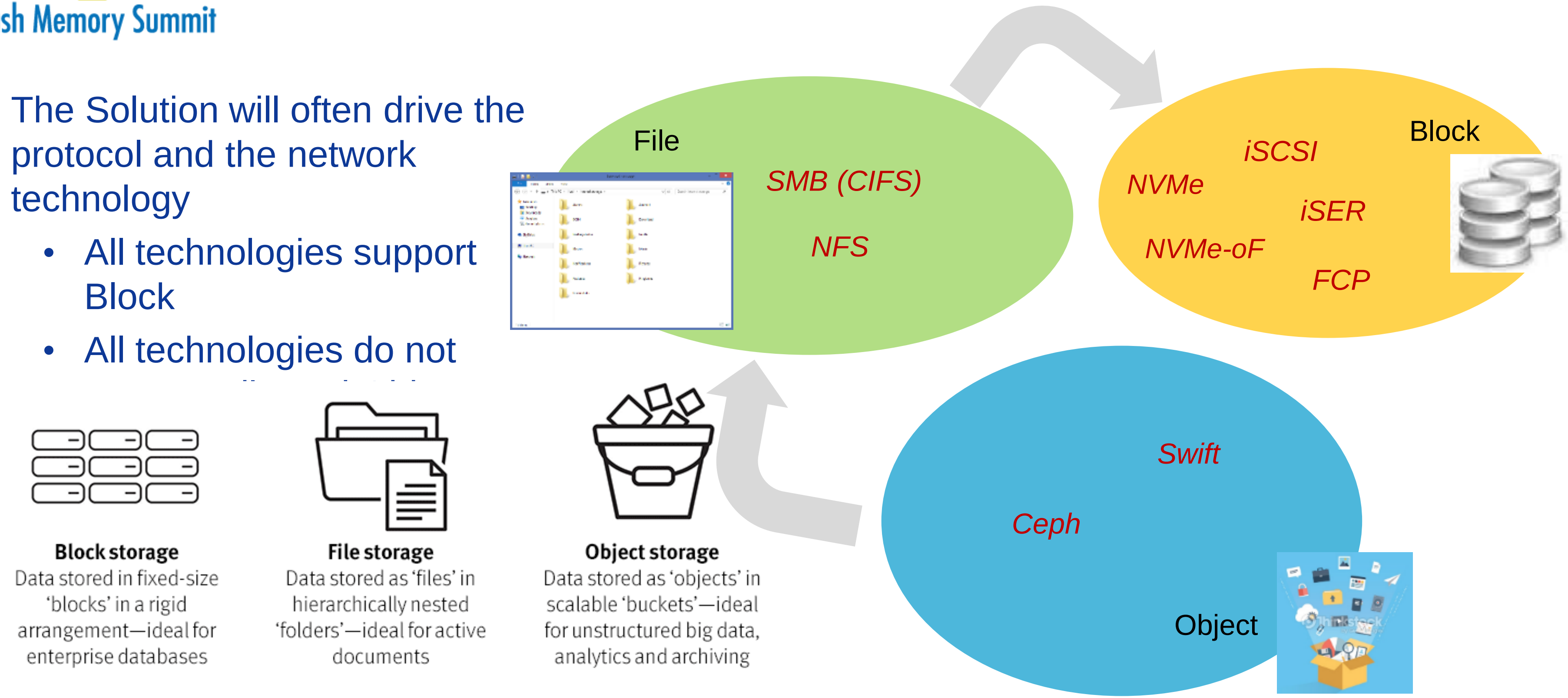


Flash Storage – Closer to Servers



Match the Network to the Solution

- The Solution will often drive the protocol and the network technology
  - All technologies support Block
  - All technologies do not



Block storage
Data stored in fixed-size 'blocks' in a rigid arrangement—ideal for enterprise databases

File storage
Data stored as 'files' in hierarchically nested 'folders'—ideal for active documents

Object storage
Data stored as 'objects' in scalable 'buckets'—ideal for unstructured big data, analytics and archiving

Conclusions

- There are tried and true reasons for networking your storage
- Networking flash requires special considerations
  - Faster Storage needs Faster Networks!
  - And protocols
- For the next few hours this team will present the different options and trade offs
- Then you get to question us

Peter Onufryk

- Peter is a Fellow in the Data Center Solutions Business Unit. where he is responsible for architecture and validation of storage products. He received a Ph.D. in Electrical and Computer Engineering from Rutgers University, has been granted over 40 patents

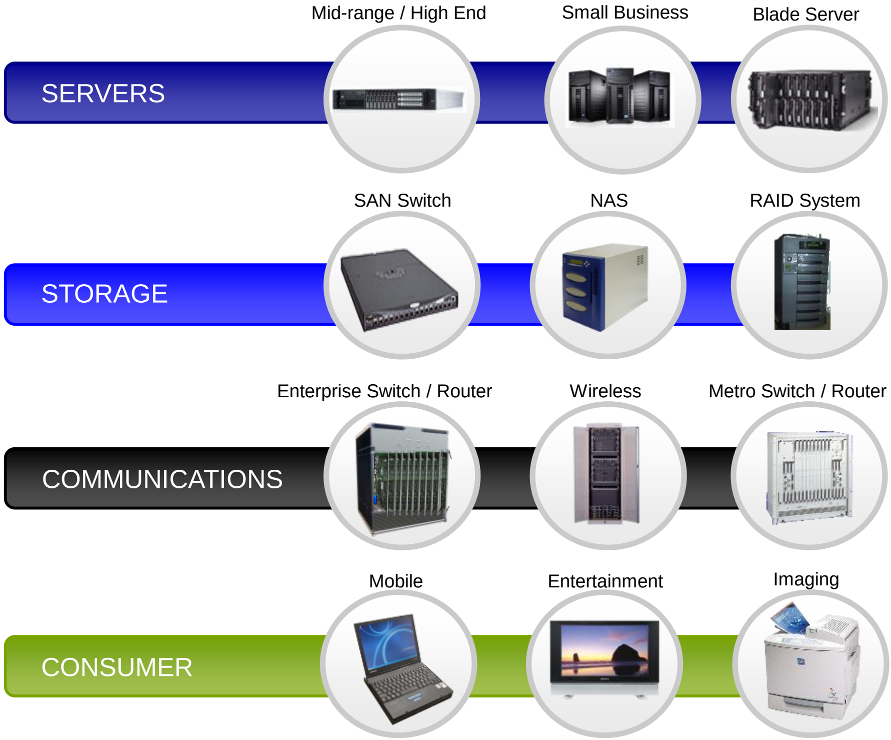
NVM
PCIe® Networked ~~Flash~~ Storage

Peter Onufryk

Microsemi Corporation

PCI Express® (PCIe®)

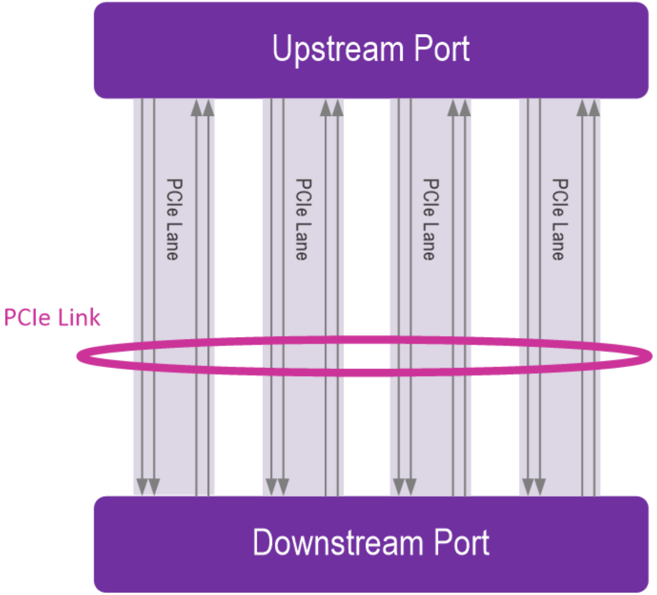
- Specification defined by PCI-SIG®
  - www.pcisig.com
- Packet-based protocol over serial links
  - Software compatible with PCI and PCI-X
  - Reliable, in-order packet transfer
- High performance and scalable from consumer to Enterprise
  - Scalable link speed (2.5 GT/s, 5.0 GT/s, 8.0 GT/s, 16 GT/s, and 32 GT/s)
    - Gen5 (32 GT/s) is still being standardized
  - Scalable link width (x1, x2, x4, …. x32)
- Primary application is as an I/O interconnect



PCIe Characteristics

- Scalable speed
  - Encoding
    - 8b10b: 2.5 GT/s (Gen 1) and 5 GT/s (Gen 2)
    - 128b/130b: 8 GT/s (Gen 3), 16 GT/s (Gen4) and 32 GT/s (Gen5)
- Scalable width: x1, x2, x4, x8, x12, x16, x32

Generation	Raw Bit Rate	Bandwidth Per Lane Each Direction	Total x16 Link Bandwidth
Gen 1*	2.5 GT/s	~ 250 MB/s	~ 8 GB/s
Gen 2*	5.0 GT/s	~500 MB/s	~16 GB/s
Gen 3*	8 GT/s	~ 1 GB/s	~ 32 GB/s
Gen 4	16 GT/s	~ 2 GB/s	~ 64 GB/s
Gen 5	32 GT/s	~4 GB/s	~128 GB/s

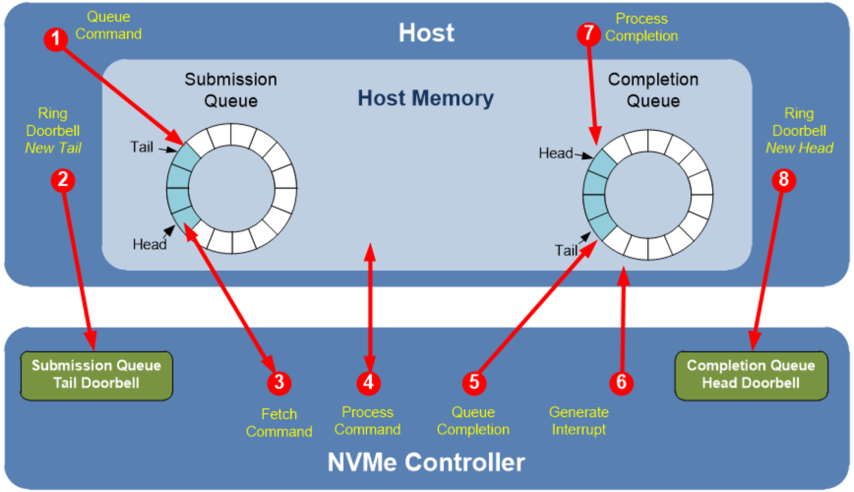


Note

** Source – PCI-SIG PCI Express 3.0 FAQ*

NVM Express^{TM} (NVMe^{TM})

- Two specifications
  1. NVM Express (PCIe)
  2. NVM Express over Fabrics (RDMA and Fibre Channel)
- Architected from the ground up for NVM
  - Simple optimized command set
  - Fixed size 64 B commands and 16 B completions
  - Supports many-core processors without locking
  - No practical limit on the number of outstanding requests
  - Supports out-of-order data deliver

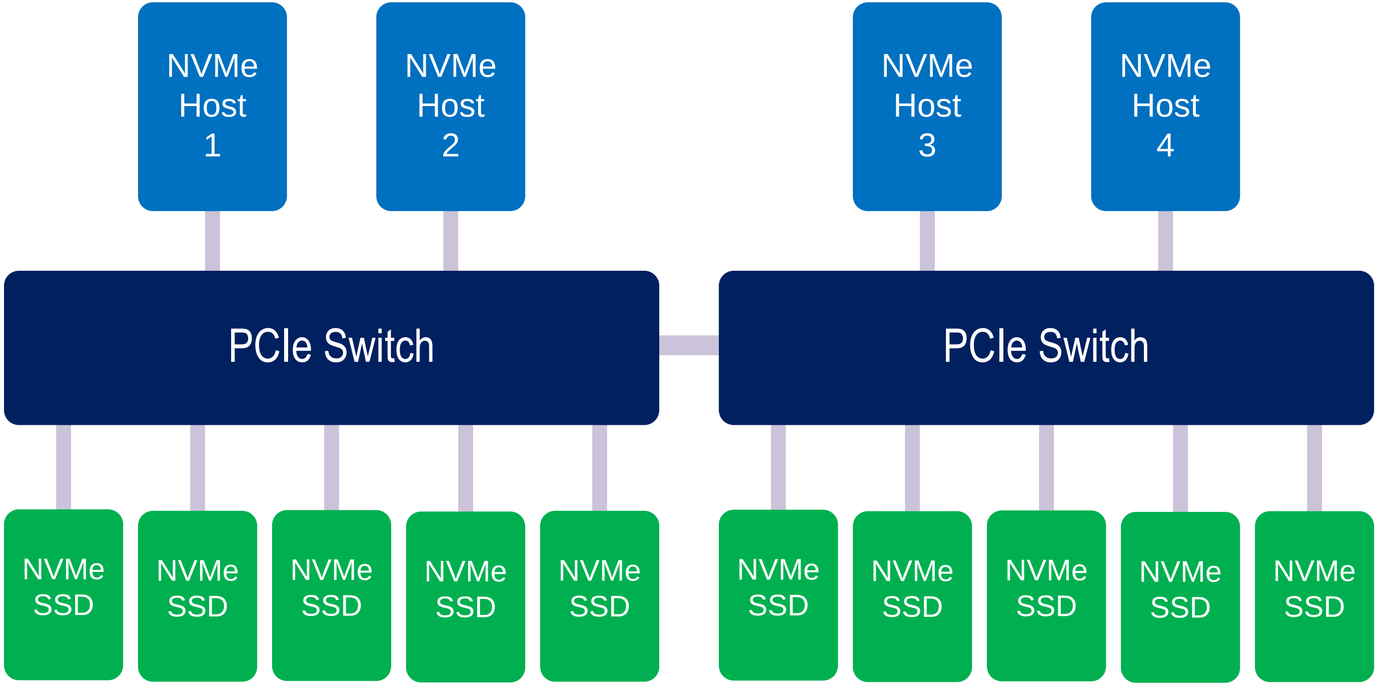


PCIe SSD = NVMe SSD

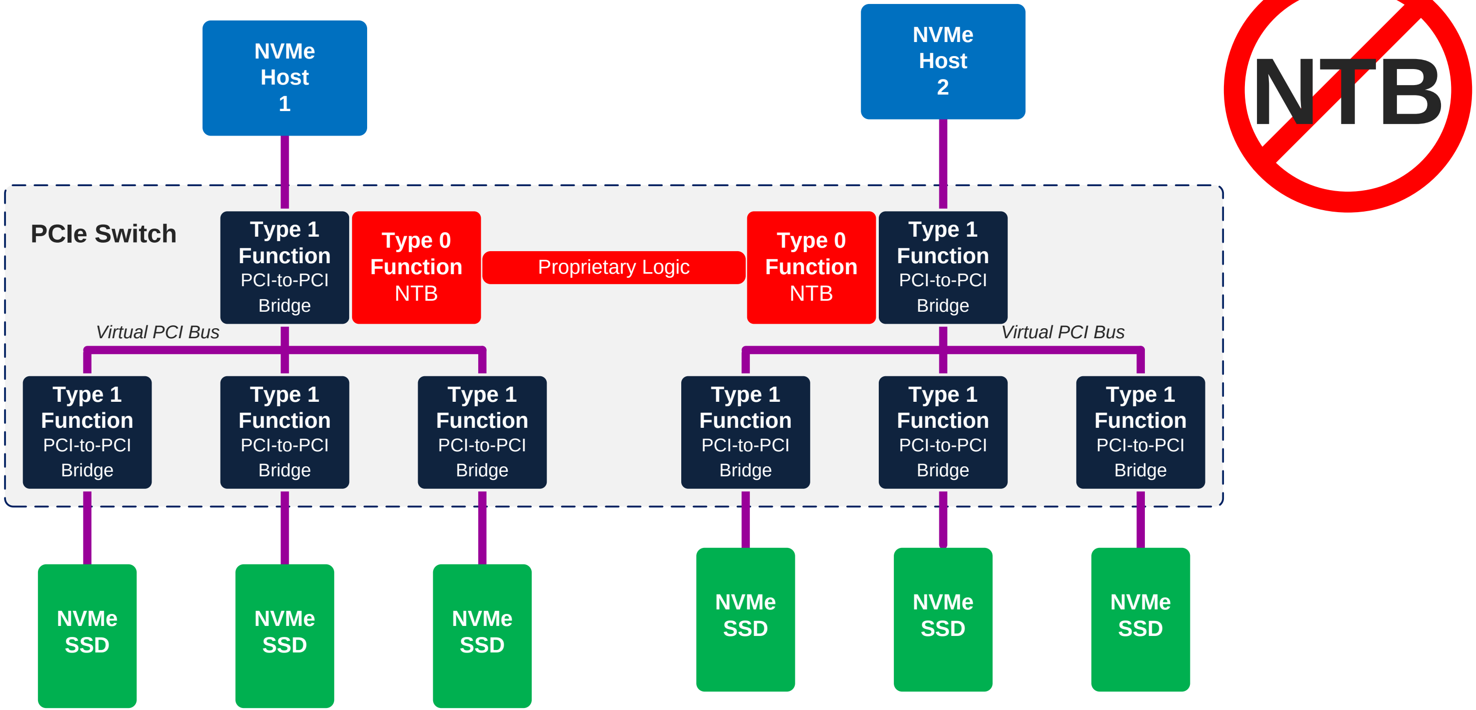
Ideal NVM Fabric

Property	Ideal Characteristic
Cost	Free
Complexity	Low
Performance	High
Power consumption	None
Standards-based	Yes
Scalability	Infinite

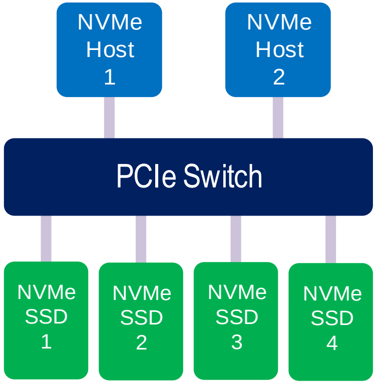
PCIe Fabric


NVMe Host 1
NVMe Host 2
NVMe Host 3
NVMe Host 4
PCIe Switch
PCIe Switch
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD

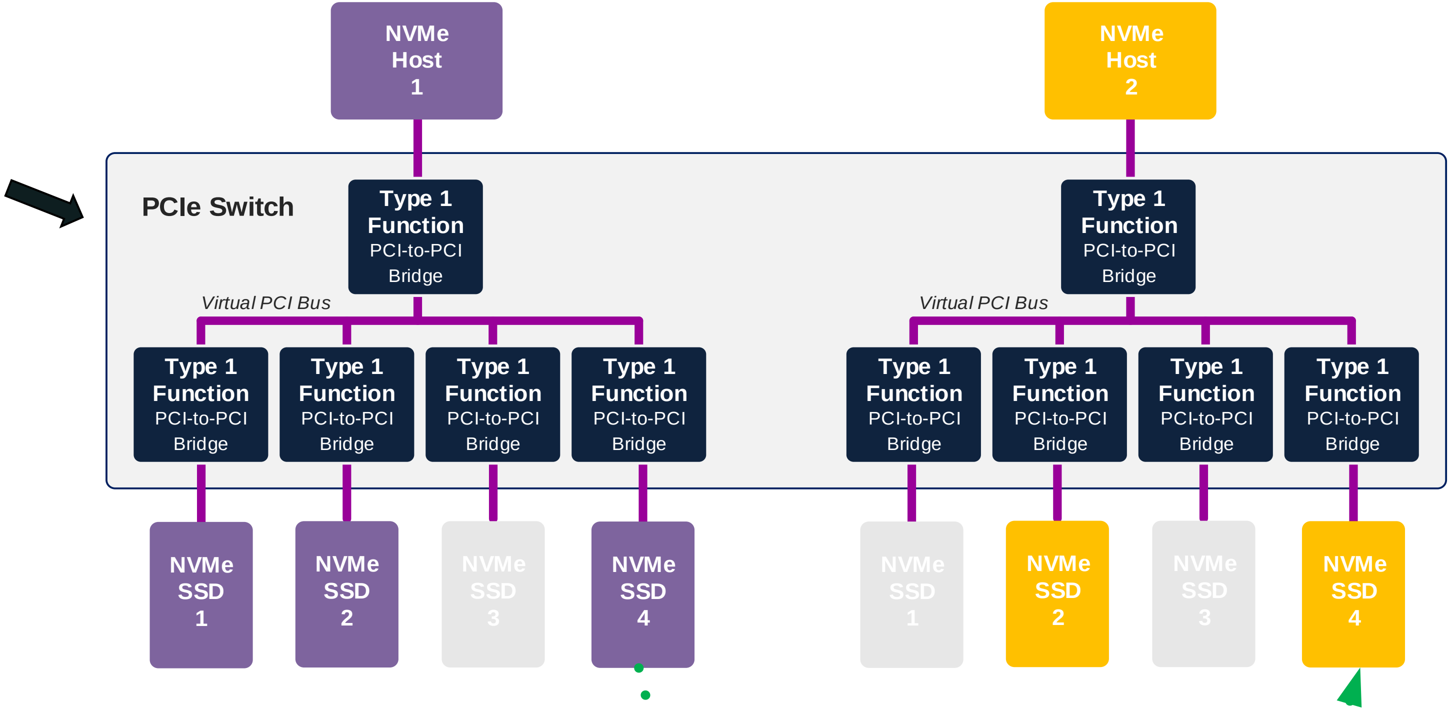
Non-Transparent Bridging (NTB)


NTB
NVMe Host 1
NVMe Host 2
PCIe Switch
Type 1 Function PCI-to-PCI Bridge
Type 0 Function NTB
Proprietary Logic
Type 0 Function NTB
Type 1 Function PCI-to-PCI Bridge
Virtual PCI Bus
Virtual PCI Bus
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
Type 1 Function PCI-to-PCI Bridge
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD
NVMe SSD

Dynamic Partitioning

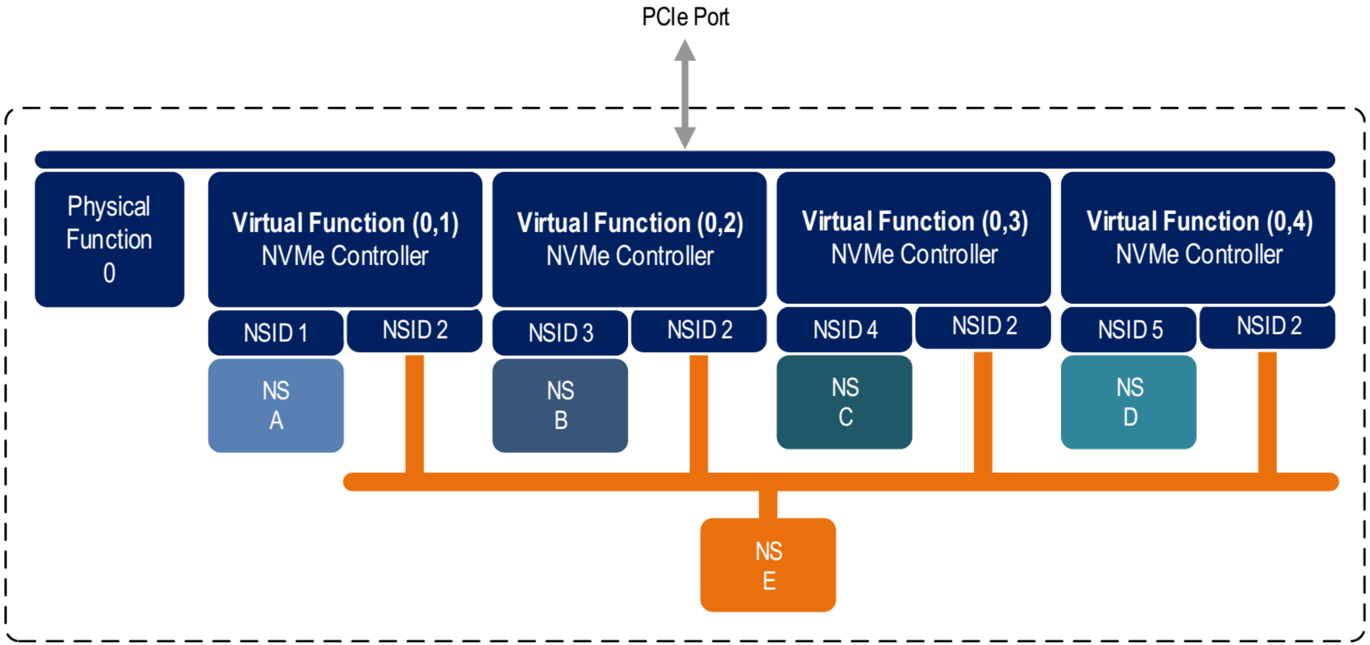


Physical View

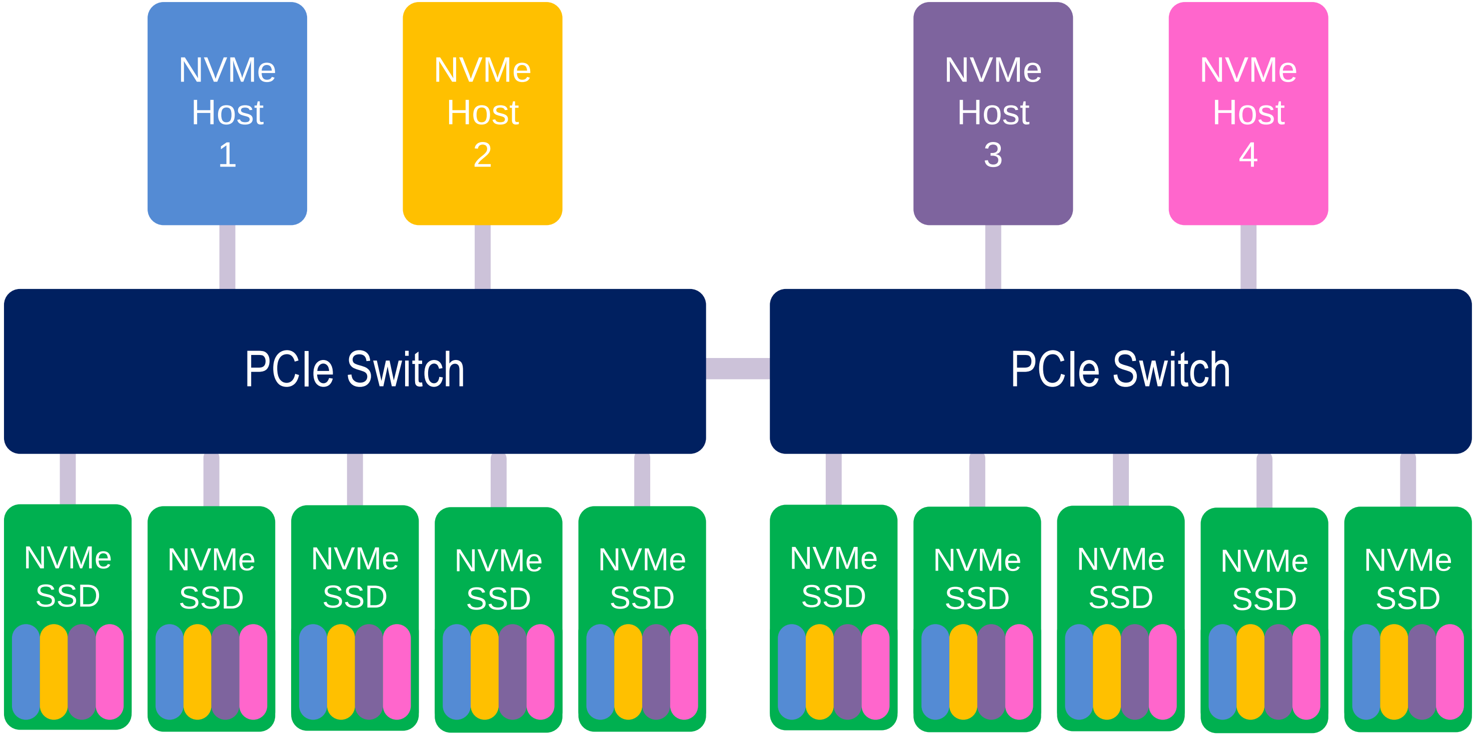


Functional View

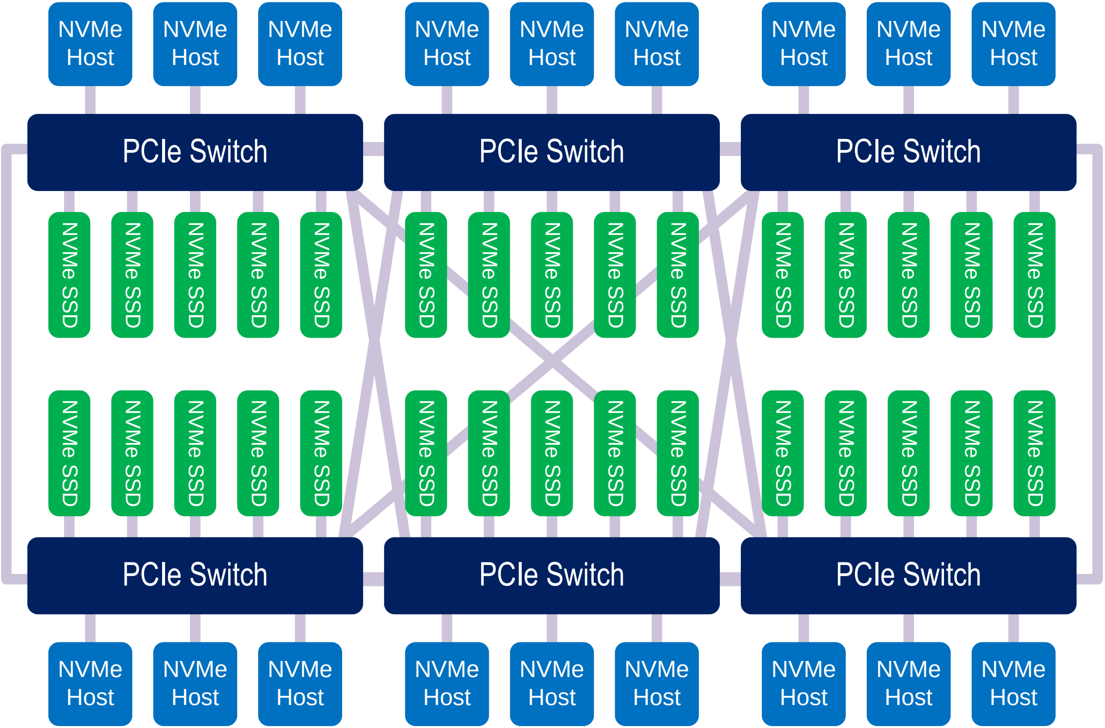
NVMe SR-IOV


PCIe Port
Physical Function 0
Virtual Function (0,1) NVMe Controller
Virtual Function (0,2) NVMe Controller
Virtual Function (0,3) NVMe Controller
Virtual Function (0,4) NVMe Controller
NSID 1
NSID 2
NSID 3
NSID 2
NSID 4
NSID 2
NSID 5
NSID 2
NS A
NS B
NS C
NS D
NS E

Multi-Host I/O Sharing



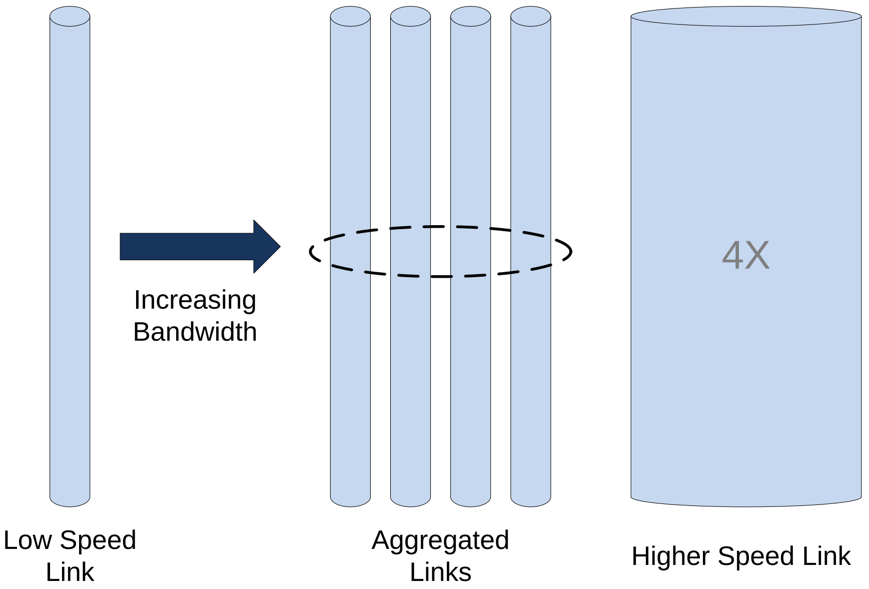
PCIe Fabric



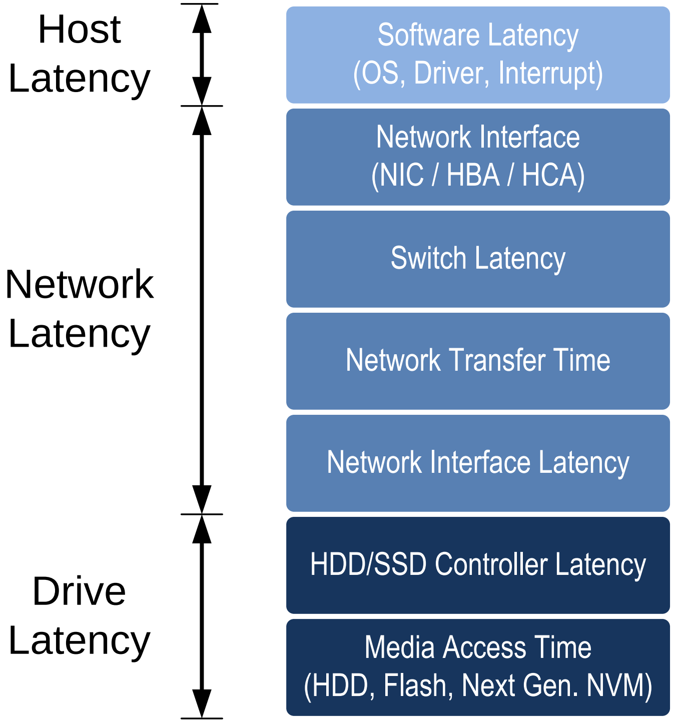
- Storage Functions
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
  - Ability to share other storage (SAS/SATA)
- Host-to-Host Communications
  - RDMA
  - Ethernet emulation
- Manageability
  - NVMe controller-to-host mapping
  - PCIe path selection
  - NVMe management
- Fabric Resilience
  - Supports link failover
  - Supports fabric manager failover

Fabric Performance

- A high performance fabric means:
  - High bandwidth
  - Low latency
- Increasing bandwidth is easy
  - Aggregate parallel links
  - Increase link speed (fatter pipe)
- Reducing latency is hard
  - Transfer latency is typically a small component of overall latency
  - Other sources of latency:
    - Software (drivers)
    - Complex protocols
    - Protocol translation
    - Fabric switches/hops



Latency



Host Latency

Software Latency (OS, Driver, Interrupt)

Network Latency

Network Interface (NIC / HBA / HCA)

Switch Latency

Network Transfer Time

Network Interface Latency

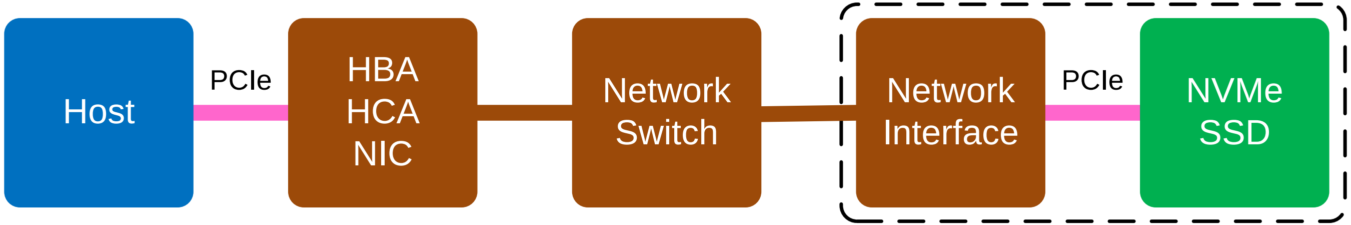
Drive Latency

HDD/SSD Controller Latency

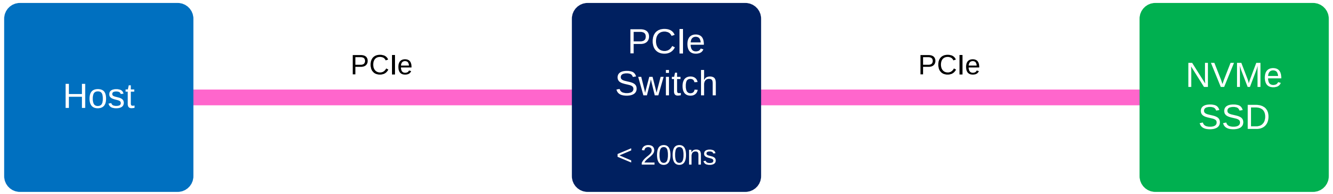
Media Access Time (HDD, Flash, Next Gen. NVM)

- Media Access Time
  - Hard drive – Milliseconds
  - NAND flash –Microseconds
  - Next-gen. NVM – Nanoseconds

The PCIe Advantage

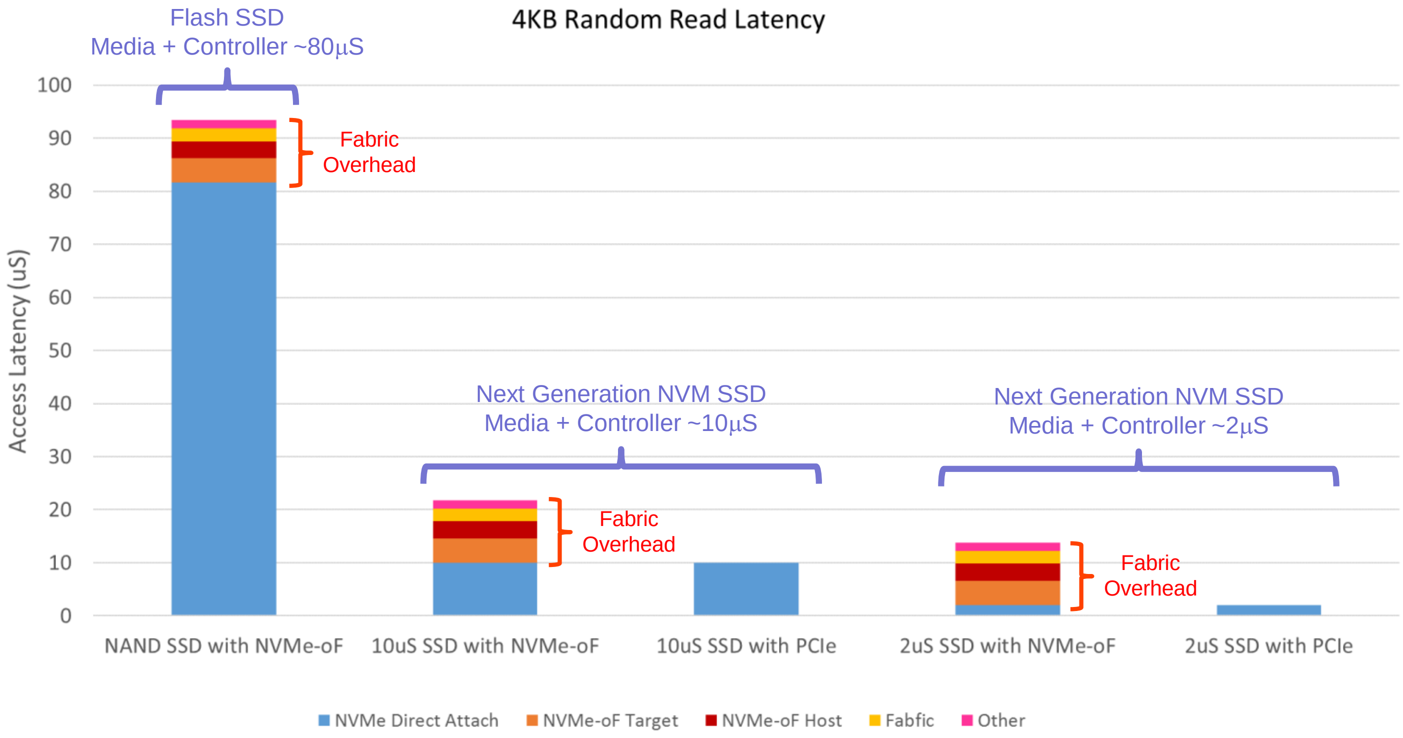


Other Flash Storage Networks



PCIe Fabric

The PCIe Latency Advantage

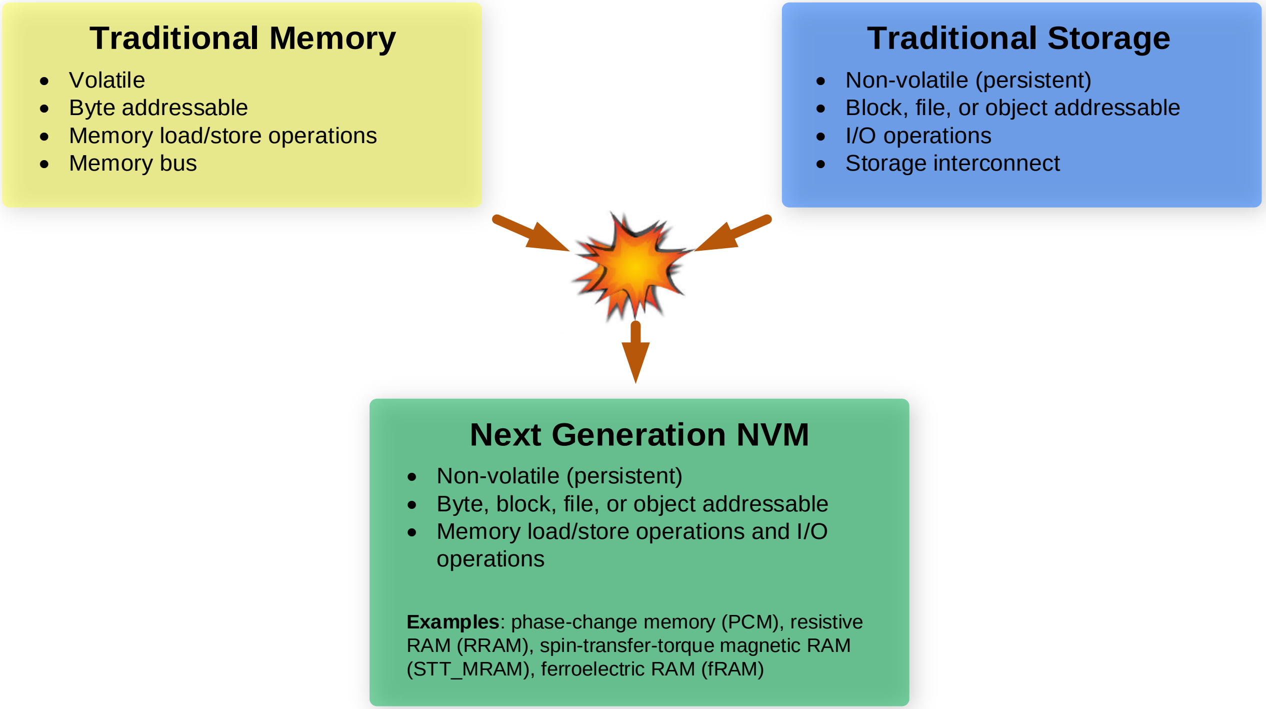


Latency data from Z. Guz et al., “NVMe-over-Fabrics Performance Characterization and the Path to Low-Overhead Flash Disaggregation” in SYSTOR ‘17

PCIe Fabric Characteristics

Property	Ideal Characteristic	PCIe Fabric	Notes
Cost	Free	Low	• PCIe built into virtually all hosts and NVMe drives
Complexity	Low	Medium	• Builds on existing NVMe ecosystem with no changes • PCIe fabrics are an emerging technology • Requires PCIe SR-IOV drives for low-latency shared storage
Performance	High	High	• High bandwidth • The absolute lowest latency
Power consumption	None	Low	• No protocol translation
Standards-based	Yes	Yes	• Works with standard hosts and standard NVMe SSDs
Scalability	Infinite	Limited	• PCIe hierarchy domain limited to 256 bus numbers • PCIe has limited reach (cables) • PCIe fabrics have limited scalability (less than 256 SSDs and 128 hosts)

Persistent Memory & Next Gen. NVM



Traditional Memory

- Volatile
- Byte addressable
- Memory load/store operations
- Memory bus

Traditional Storage

- Non-volatile (persistent)
- Block, file, or object addressable
- I/O operations
- Storage interconnect

Next Generation NVM

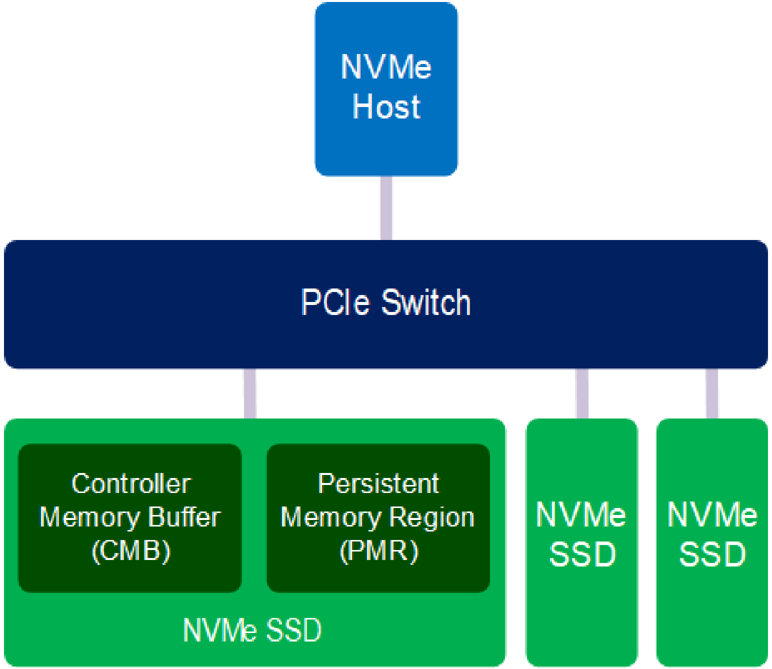
- Non-volatile (persistent)
- Byte, block, file, or object addressable
- Memory load/store operations and I/O operations

Examples: phase-change memory (PCM), resistive RAM (RRAM), spin-transfer-torque magnetic RAM (STT_MRAM), ferroelectric RAM (fRAM)

NVMe and Memory Operations

- Controller Memory Buffer (CMB)
  - PCI memory space exposed to host (byte addressable)
  - May be used to store commands & data
  - Contents **do not** persist across power cycles and resets

- Persistent Memory Region (PMR)
  - PCI memory space exposed to host (byte addressable)
  - May be used to store data
  - Content persist across power cycles and resets


NVMe Host
PCIe Switch
Controller Memory Buffer (CMB)
Persistent Memory Region (PMR)
NVMe SSD
NVMe SSD
NVMe SSD

Storage is Not Just About CPU I/O Anymore

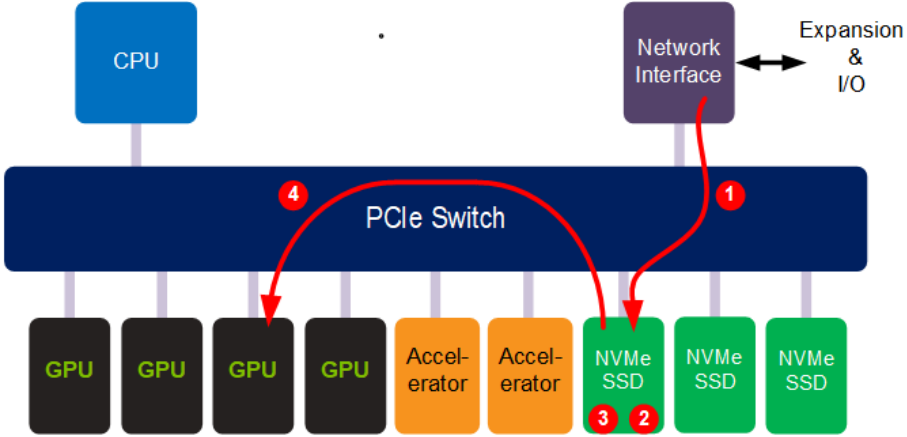
- NVMe together with a PCIe fabric allow direct network to storage and accelerator to storage communications

Example:

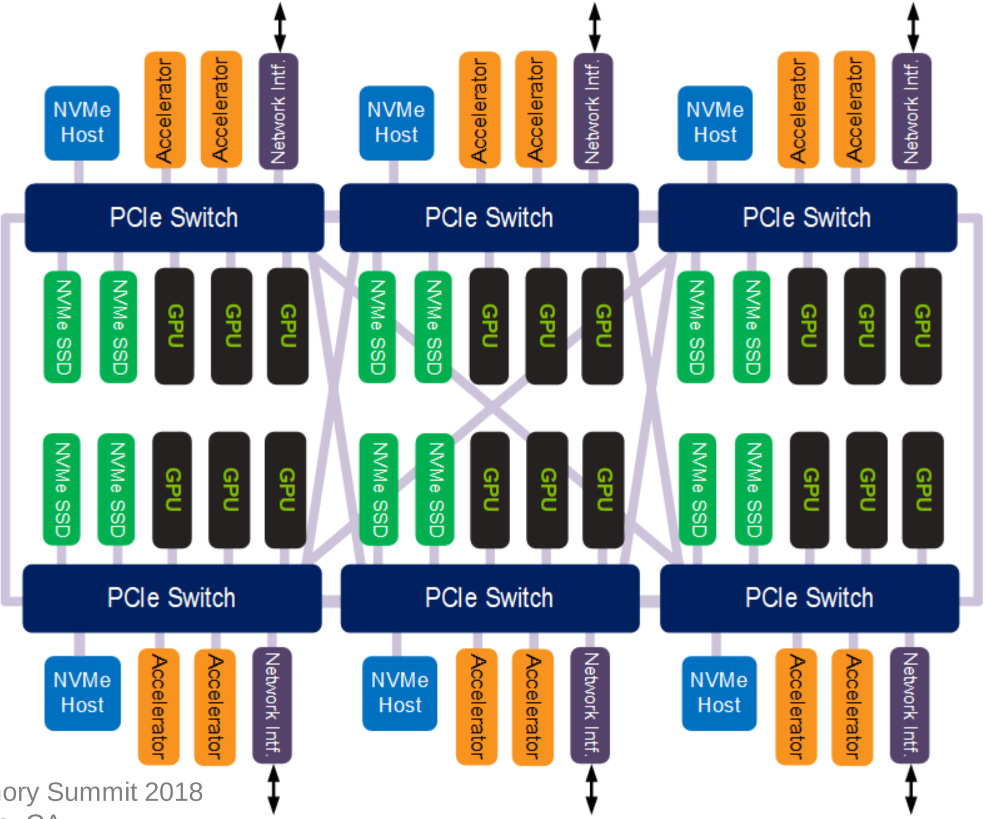
1. Data transferred from network to NVMe CMB
2. NVMe block write operation imitated from CMB to NVM

… sometime later …

3. NVMe block read operation initiated from NVM to CMB
4. GPU/Accelerator transfers data from NVMe CMB for processing



Putting it All Together



- NVMe Storage Functions
  - Dynamic partitioning (drive-to-host mapping)
  - NVMe shared I/O (shared storage)
- Direct accelerator-to-NVMe and network-to-NVMe transfers
- Byte addressable persistent memory

Summary

- PCIe fabrics build on the existing PCIe and NVMe ecosystem
  - Work with standard NVMe SSDs, OS drivers, and PCIe infrastructure
- PCIe fabrics support both byte addressable memory and traditional storage operations
- PCIe fabrics are well suited for applications that require low cost, the absolute lowest latency, and limited scalability
  - NVMe SSD sharing inside a rack and small clusters
- PCIe fabrics are not well suited for long reach applications or where a high degree of scalability is required
  - NVM Express over Fabrics (NVMe-oF^{TM}) is well suited for these applications

Motti Beck

Motti Beck is Sr. Director of Marketing, Enterprise Data Center market segment at Mellanox Technologies, Inc. Before joining Mellanox, Motti was a founder of several start-up companies including BindKey Technologies that was acquired by DuPont Photomask (today Toppan Printing Company LTD) and Butterfly Communications that was acquired by Texas Instrument. Prior to that he was a Business Unit Director at National Semiconductors. Motti hold B.Sc in computer engineering from the Technion - Israel Institute of Technology.

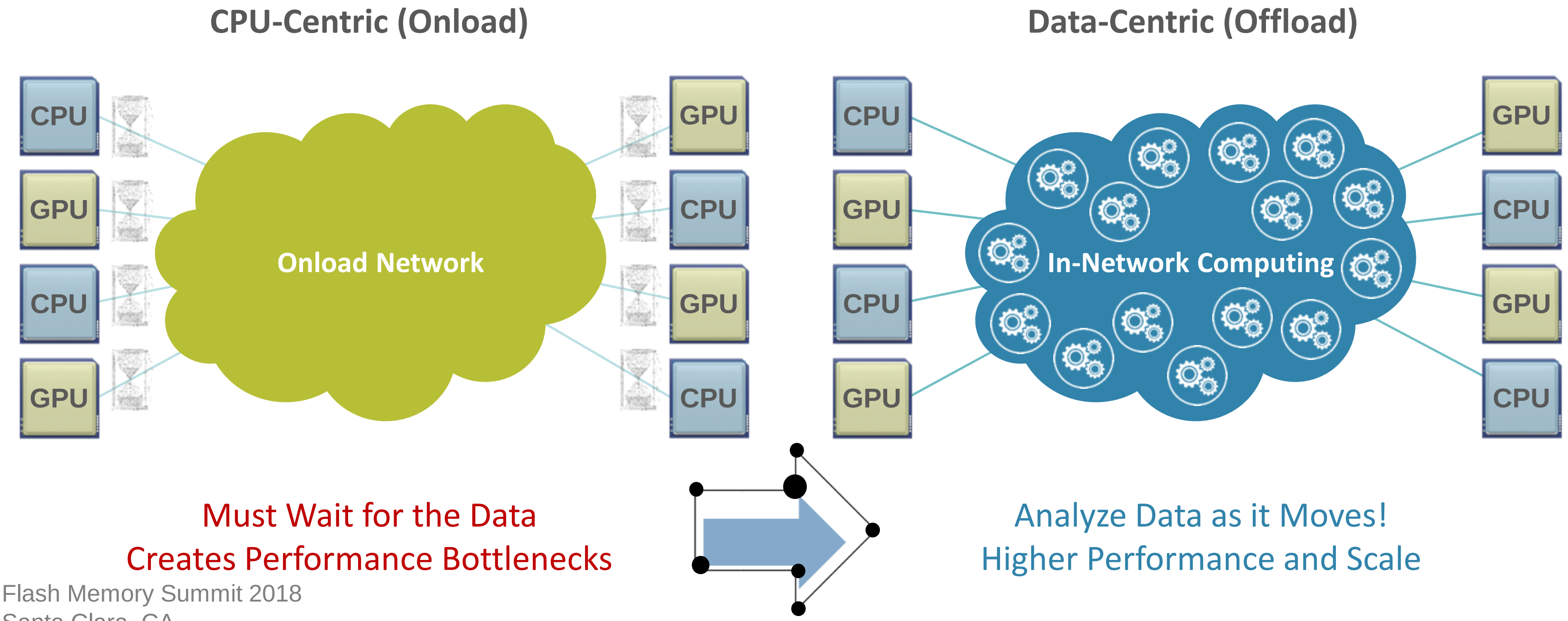
InfiniBand Networked Flash Storage

Superior Performance, Efficiency and Scalability

Motti Beck – Sr. Director Enterprise Market Development, Mellanox Technologies

The Need for Intelligent and Faster Interconnect

Faster Data Speeds and In-Network Computing Enable Higher Performance and Scale



Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale

In-Network Processing Enables Higher Efficiency

- Higher Scalability
- Lower latency
- Higher ROI

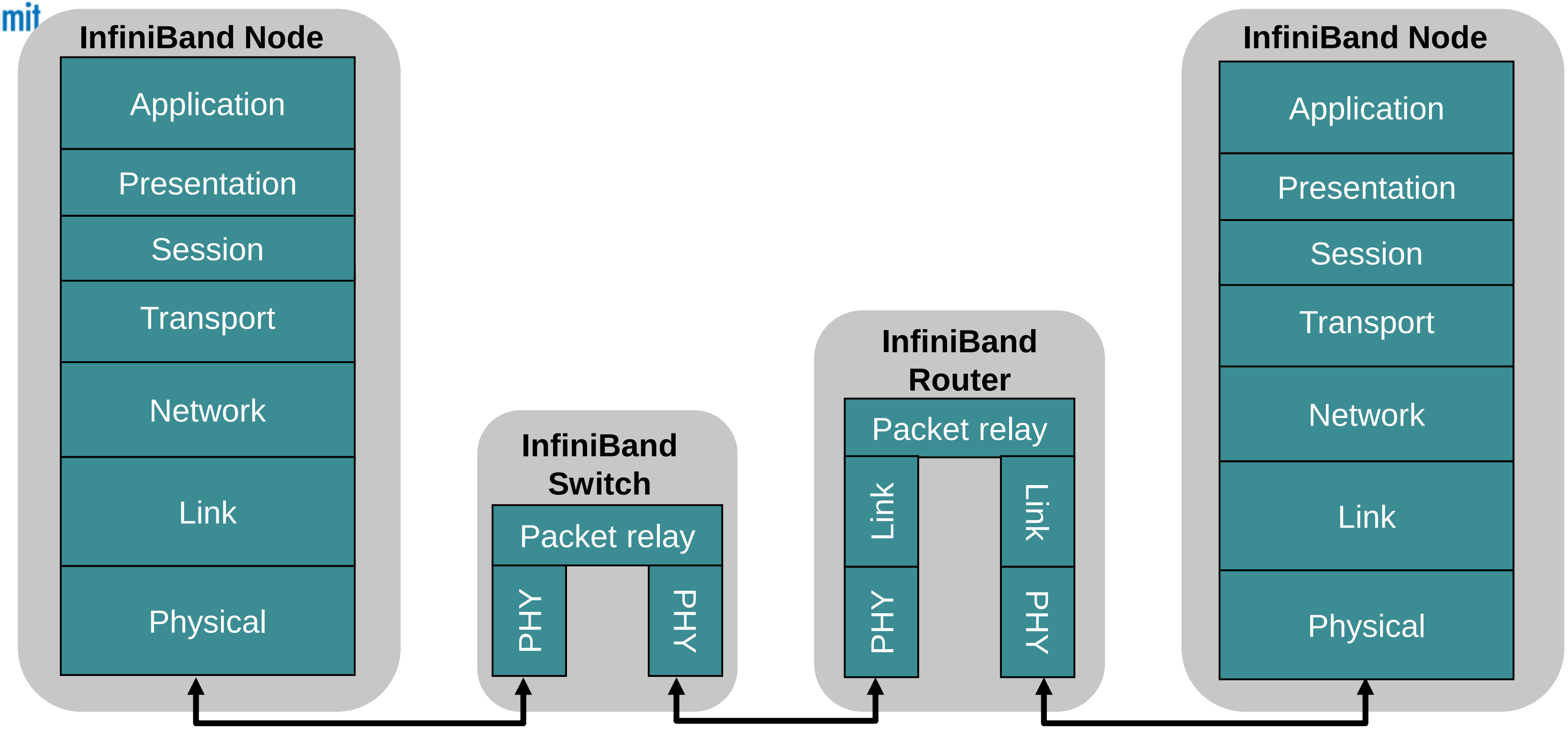
InfiniBand Technical Overview

- What is InfiniBand?
  - InfiniBand is an open standard, interconnect protocol developed by the InfiniBand® Trade Association: http://www.infinibandta.org/home
  - First InfiniBand specification was released in 2000

- What does the specification includes?
  - The specification is very comprehensive
  - From physical to applications

- InfiniBand SW is open and has been developed under OpenFabrics Alliance
  - http://www.openfabrics.org/index.html

InfiniBand Protocol Layers



InfiniBand Node

Application

Presentation

Session

Transport

Network

Link

Physical

InfiniBand Switch

Packet relay

PHY

PHY

InfiniBand Router

Packet relay

Link

Link

PHY

PHY

InfiniBand Node

Application

Presentation

Session

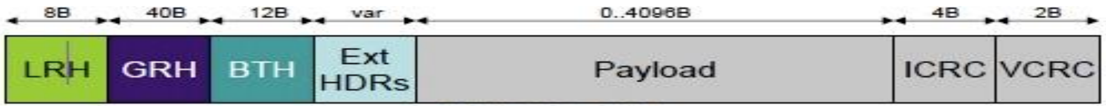
Transport

Network

Link

Physical

Packet Format

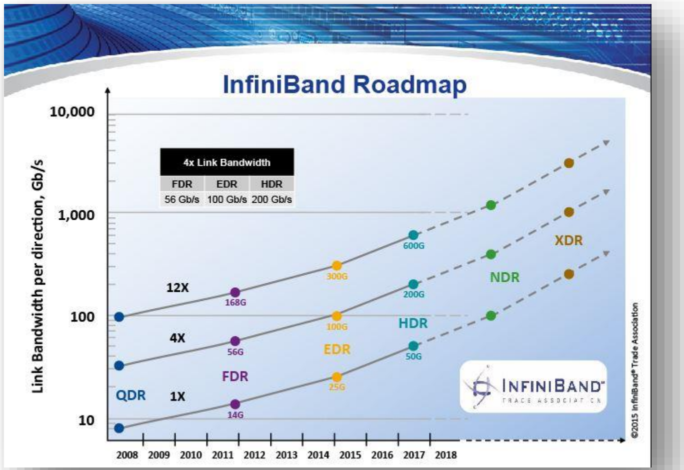


8B	40B	12B	var	0..4096B	4B	2B
LRH	GRH	BTH	Ext HDRs	Payload	ICRC	VCRC

InfiniBand Data Packet

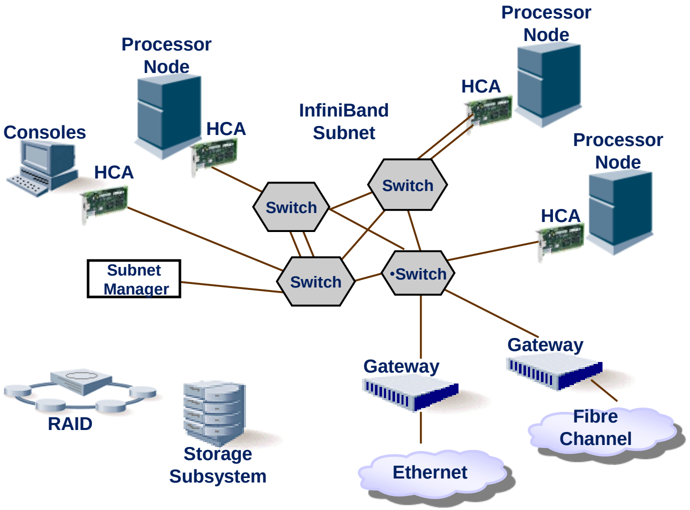
InfiniBand Architecture Highlights

- Reliable, lossless, self-managed fabric
- Hardware based transport protocol- Remote Direct Memory Access (RDMA)
- Centralized fabric management – Subnet Manger (SM)



InfiniBand Roadmap

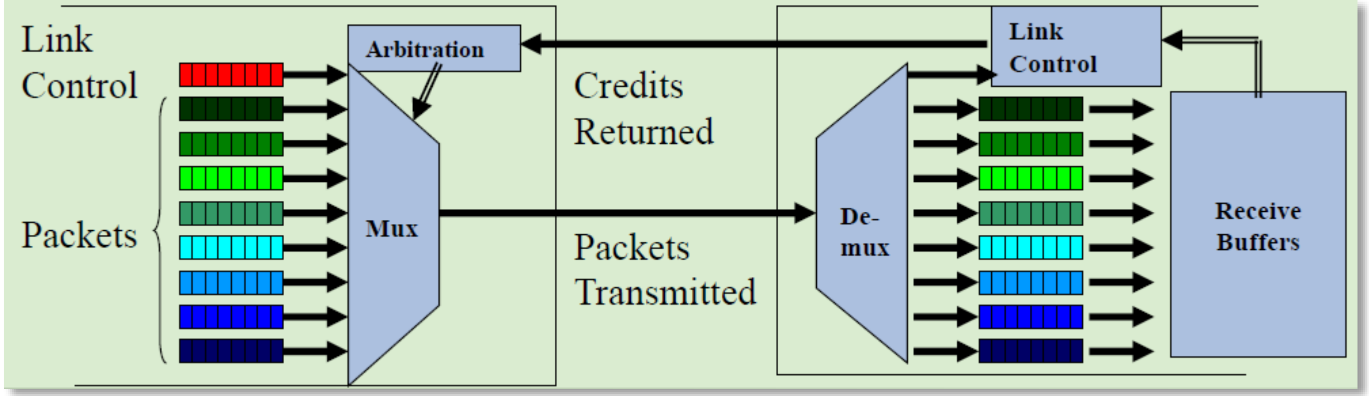
4x Link Bandwidth		
FDR	EDR	HDR
56 Gb/s	100 Gb/s	200 Gb/s



Reliable, Lossless, Self-Managed Fabric

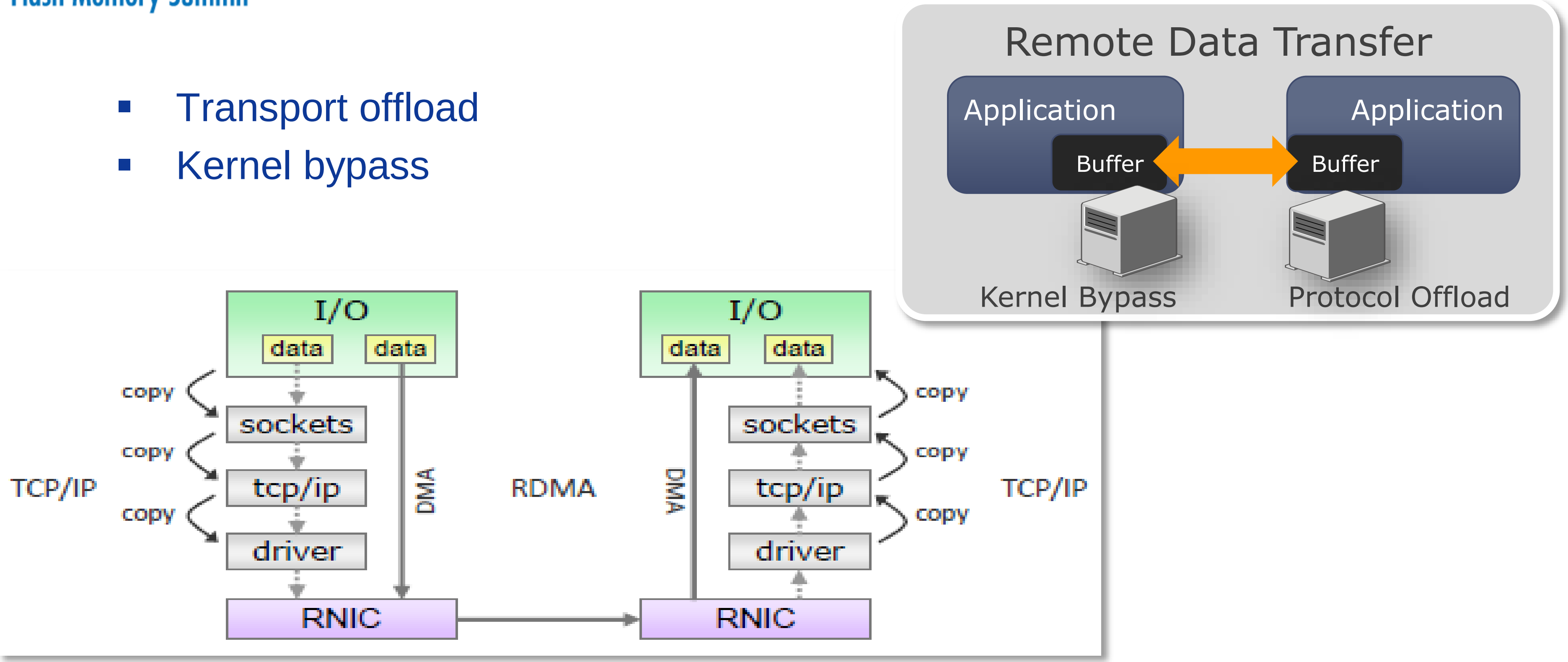
- Credit-based link-level flow control
  - Link Flow control assures **NO packet loss** within fabric even in the presence of congestion
  - Link Receivers grant packet receive buffer space credits per Virtual Lane
  - Flow control credits are issued in 64 byte units

- Separate flow control per Virtual Lanes provides:
  - Alleviation of head-of-line blocking
  - Virtual Fabrics – Congestion and latency on one VL does not impact traffic with guaranteed QOS on another VL even though they share the same physical link



Remote Direct Memory Access RDMA

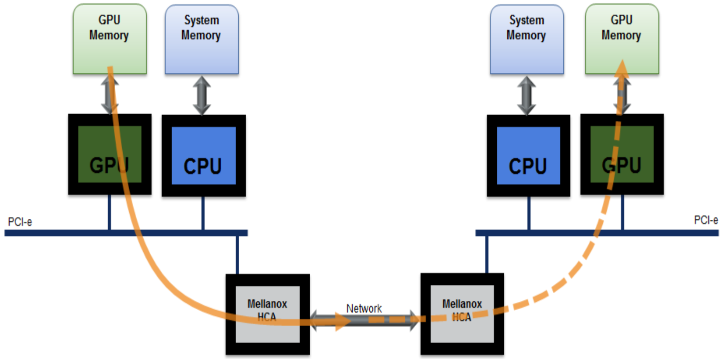
- Transport offload
- Kernel bypass

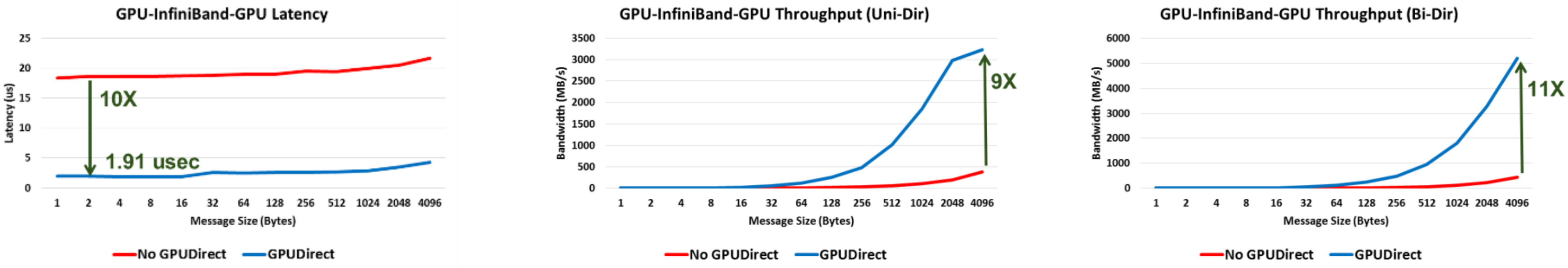


10X Better Performance with GPUDirect™ RDMA

- Purpose-built for Acceleration of Deep Learning
- Lowest communication latency for acceleration devices
- No unnecessary system memory copies and CPU overhead
- Enables GPUDirect™ RDMA and ASYNC, ROCm and others
- InfiniBand and RoCE

GPUDirect™ RDMA, GPUDirect™ ASYNC





Scaling HPC and ML with GPUDirect over InfiniBand on vSphere 6.7

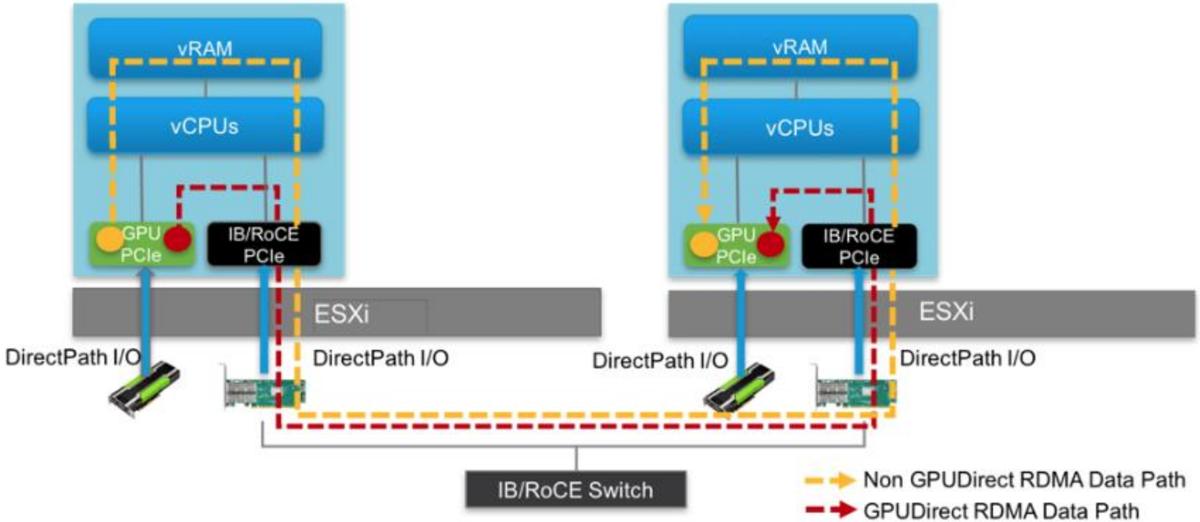
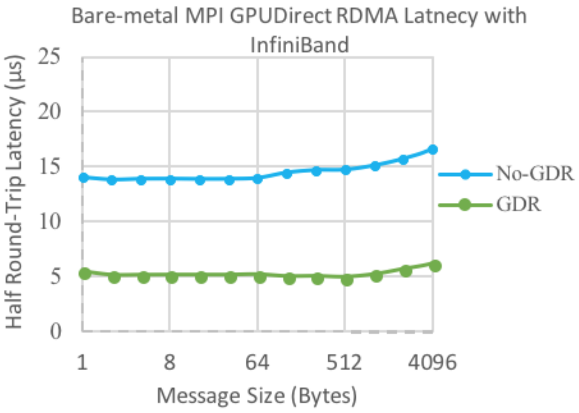
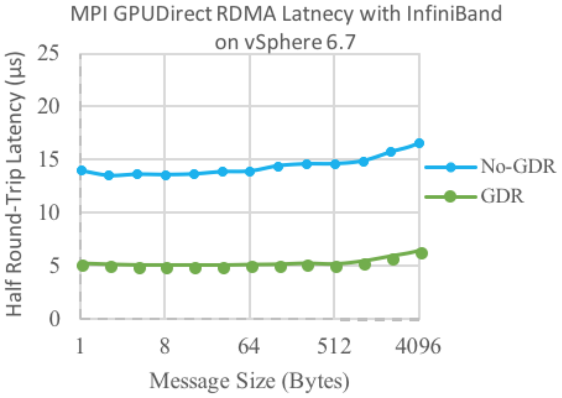


Figure 3: Testbed virtual cluster architecture showing the no-GPUDirect RDMA vs. GPUDirect RDMA data path with DirectPath I/O on vSphere 6.7





Source: Scaling HPC and ML with GPUDirect RDMA on vSphere 6.7

Subnet Management

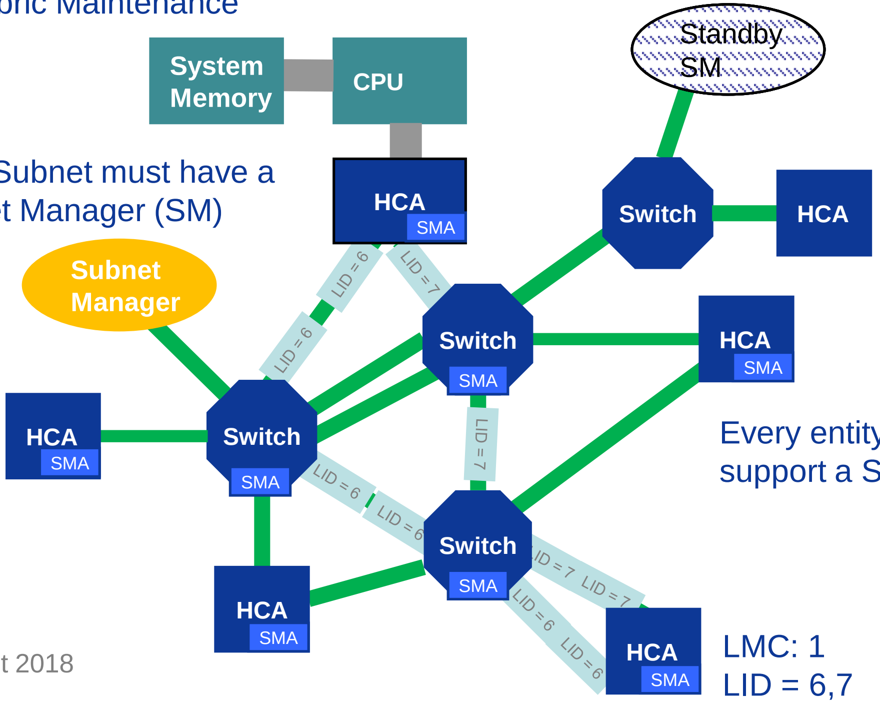
Topology Discovery
Fabric Initialization
Fabric Maintenance

Initialization uses
Directed Route MADs:

LID Route | Directed Route Vector | LID Route

Each Subnet must have a Subnet Manager (SM)

Management use unreliable datagrams (MAD)



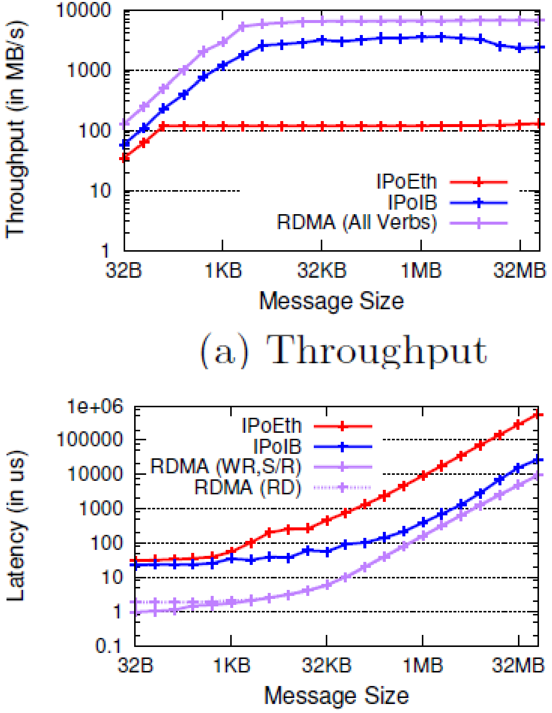
Every entity (HCA, Switch or Router) must support a Subnet Management Agent (SMA)

LMC: 1
LID = 6,7

Multipathing: LMC Supports Multiple LIDS

InfiniBand Superior Performance*

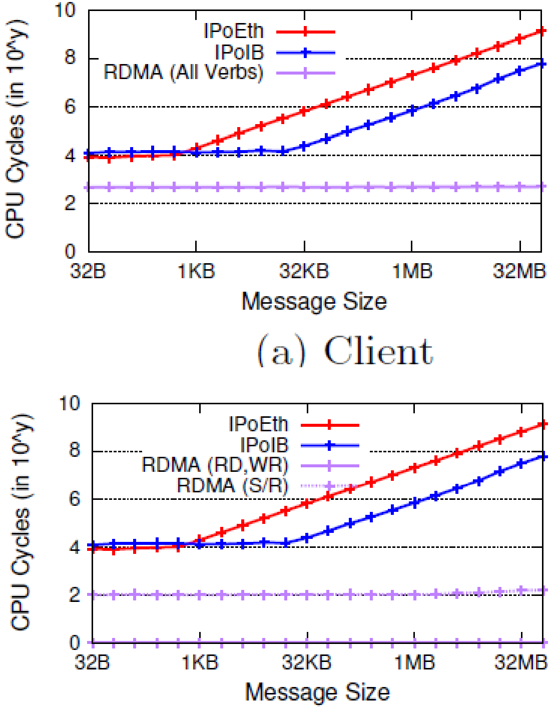
Network Throughput and Latency



(a) Throughput

(b) Latency

CPU Overhead for Network Operations

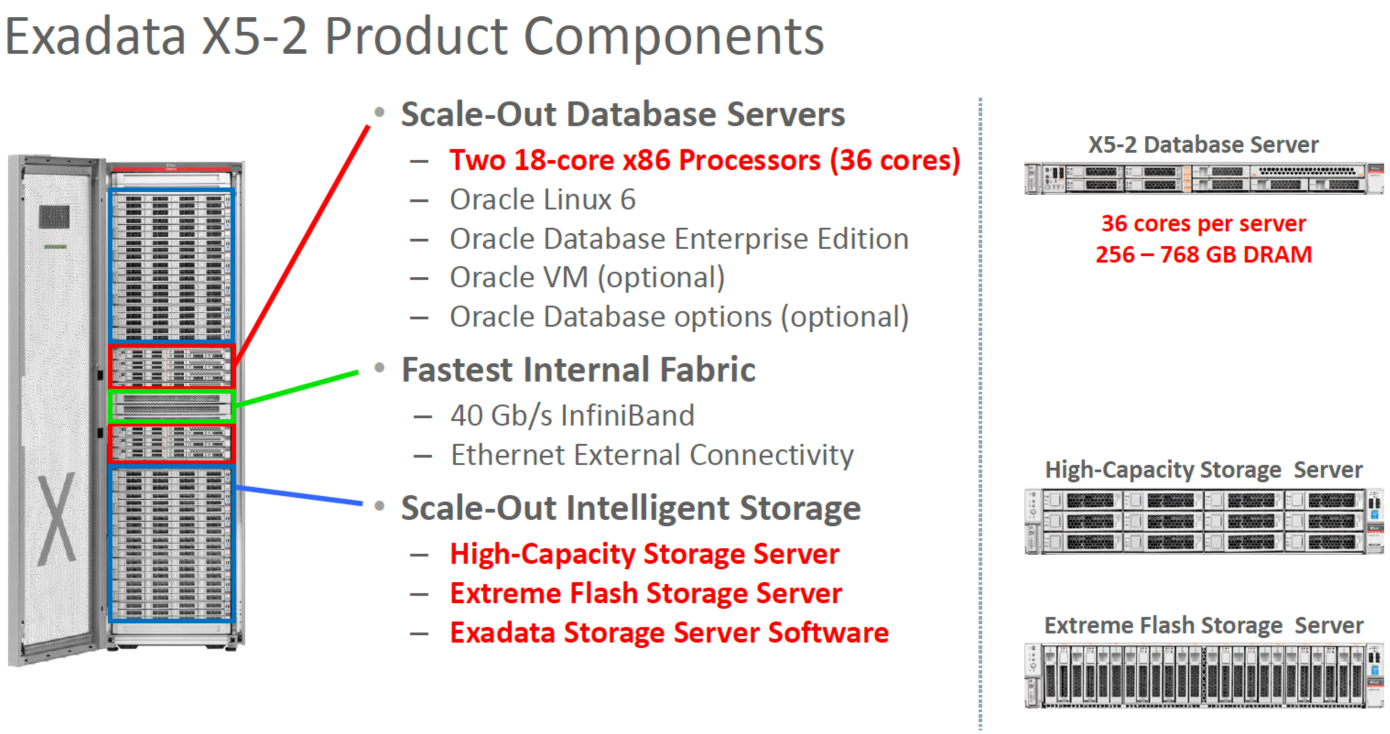


(a) Client

(b) Server

* Source: Brown University Research: "The End of Slow Networks: It's Time for a Redesign"

InfiniBand Enables Most Cost Effective Database Storage



Exadata X5-2 Product Components

- **Scale-Out Database Servers**
  - **Two 18-core x86 Processors (36 cores)**
  - Oracle Linux 6
  - Oracle Database Enterprise Edition
  - Oracle VM (optional)
  - Oracle Database options (optional)
- **Fastest Internal Fabric**
  - 40 Gb/s InfiniBand
  - Ethernet External Connectivity
- **Scale-Out Intelligent Storage**
  - **High-Capacity Storage Server**
  - **Extreme Flash Storage Server**
  - **Exadata Storage Server Software**

X5-2 Database Server

36 cores per server
256 – 768 GB DRAM

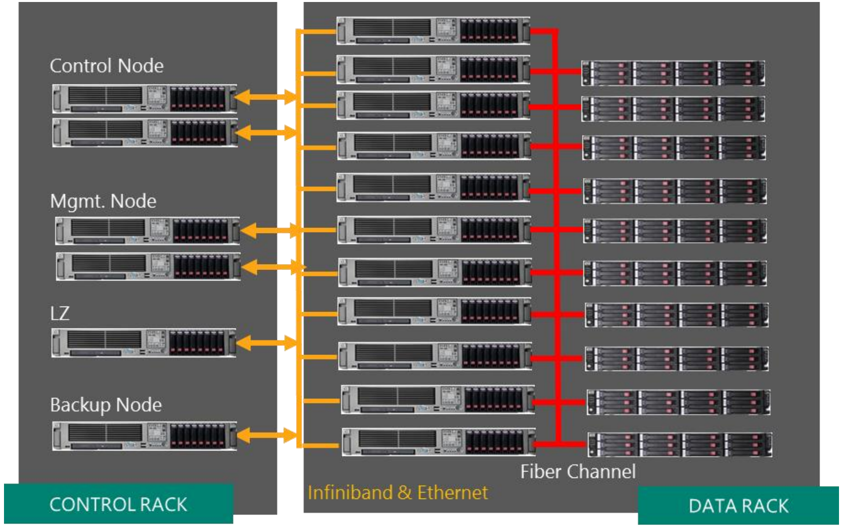
High-Capacity Storage Server

Extreme Flash Storage Server

Source: Oracle

InfiniBand Networking Storage enables Higher Efficiency

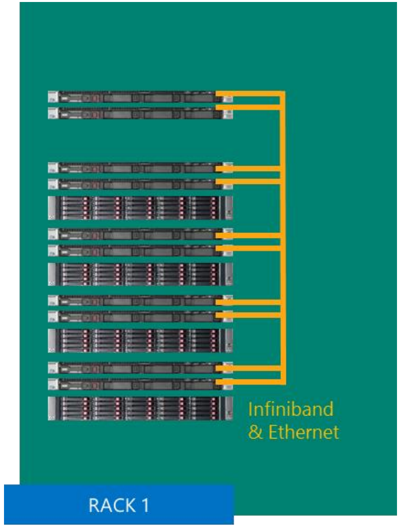
PDW* V1 Reference: The Basic Full Rack



Per RACK details

- 160 cores on 10 compute nodes
- 1.28 TB of RAM on compute
- Up to 30 TB of temp DB
- Up to 150 TB of user data

Parallel Data Warehouse
10X Faster & Lower Capital Cost



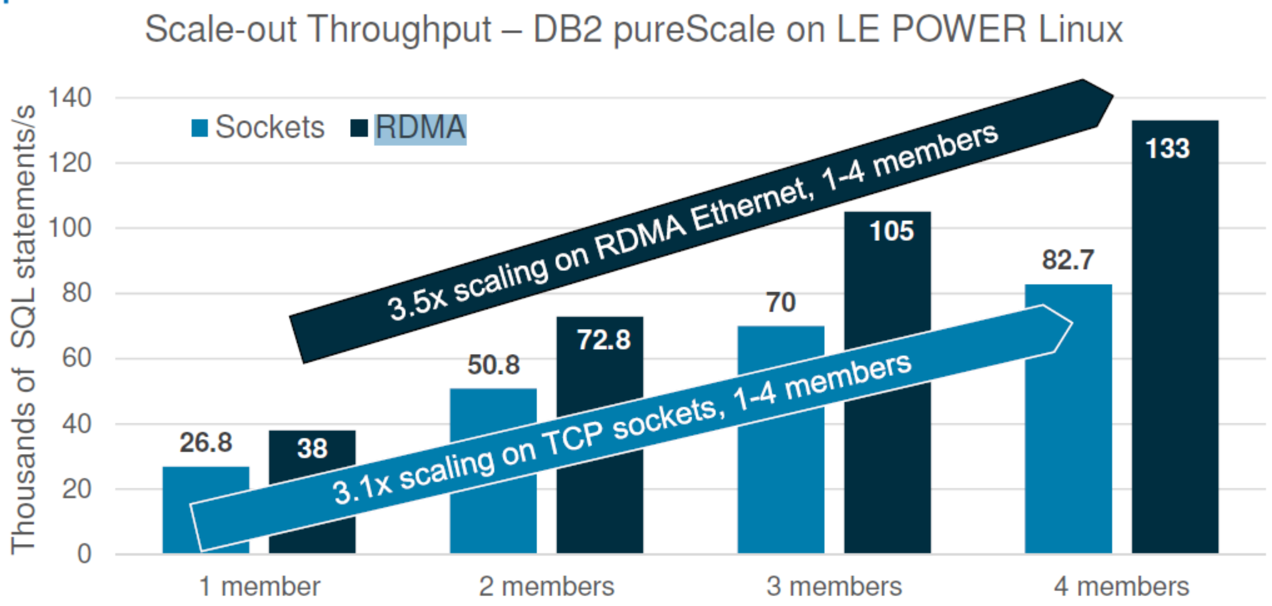
Per RACK Details

- 128 cores on 8 compute nodes
- 2TB of RAM on compute
- Up to 168 TB of temp DB
- Up to 1PB of user data

*Parallel Data Warehouse

Source: Big Data Integration with SQL Server PDW 2012

RDMA enables Higher Scalability with IBM DB2 pureScale

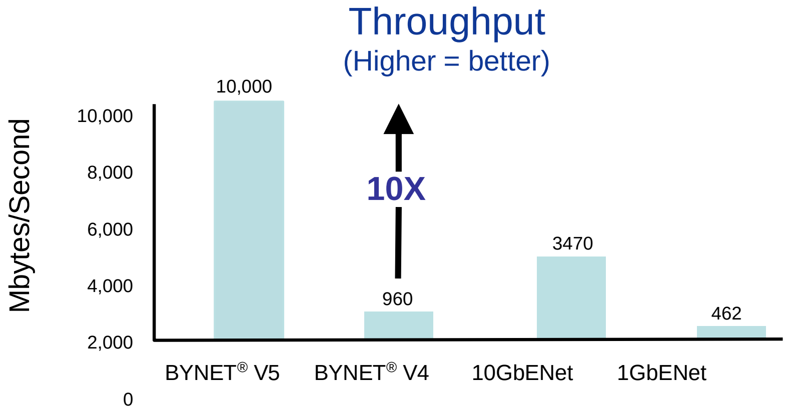


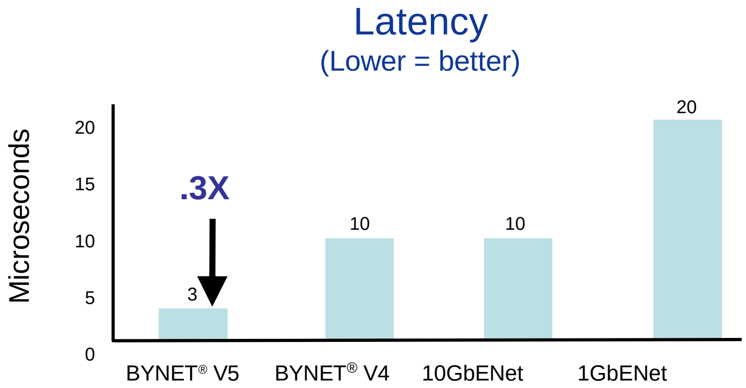
- 80% read / 20% write OLTP workload
- POWER8 4c/32t, 160 GB LBP
- 10 Gb RoCE RDMA Ethernet / 10 Gb TCP sockets

Source: IBM

Teradata BYNET® V5 Performance

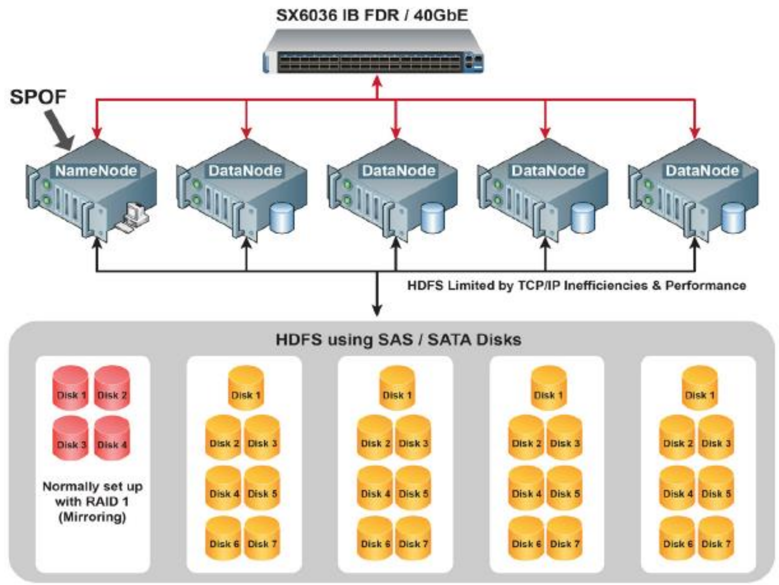
- BYNET’s basic link performance enhanced with InfiniBand
  - Dual InfiniBand links provide 10GB per second
  - 10X higher than previous BYNET®
- Message delays decreased
  - Latency in interconnect reduced by 2/3



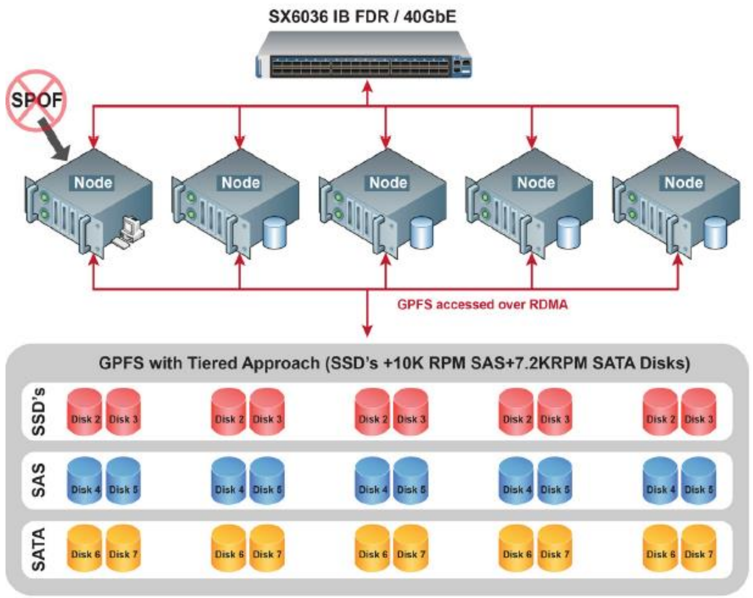


InfiniBand Unleashed the Power of Flash

Hadoop HDFS Architecture

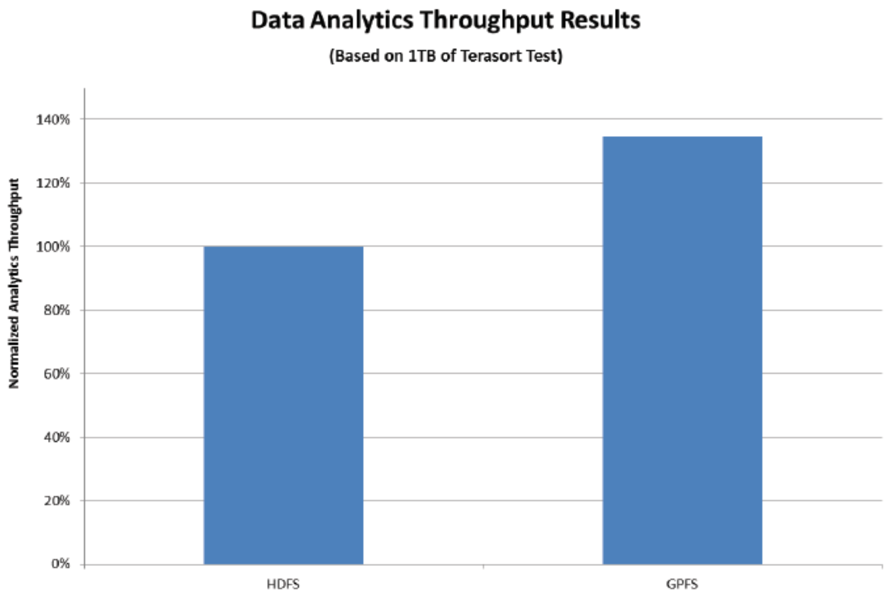


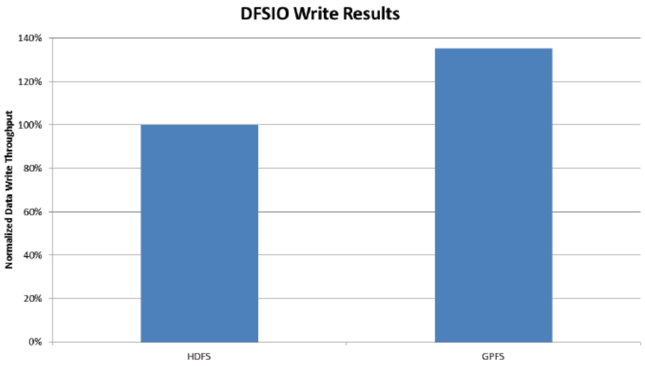
Hadoop GPFS Architecture

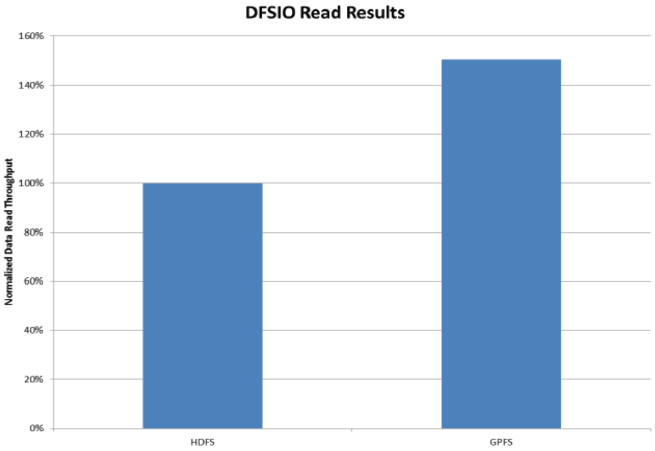


Source: Driving IBM BigInsights Performance Over GPFS Using InfiniBand+RDMA

InfiniBand Accelerate Big Data Analytics

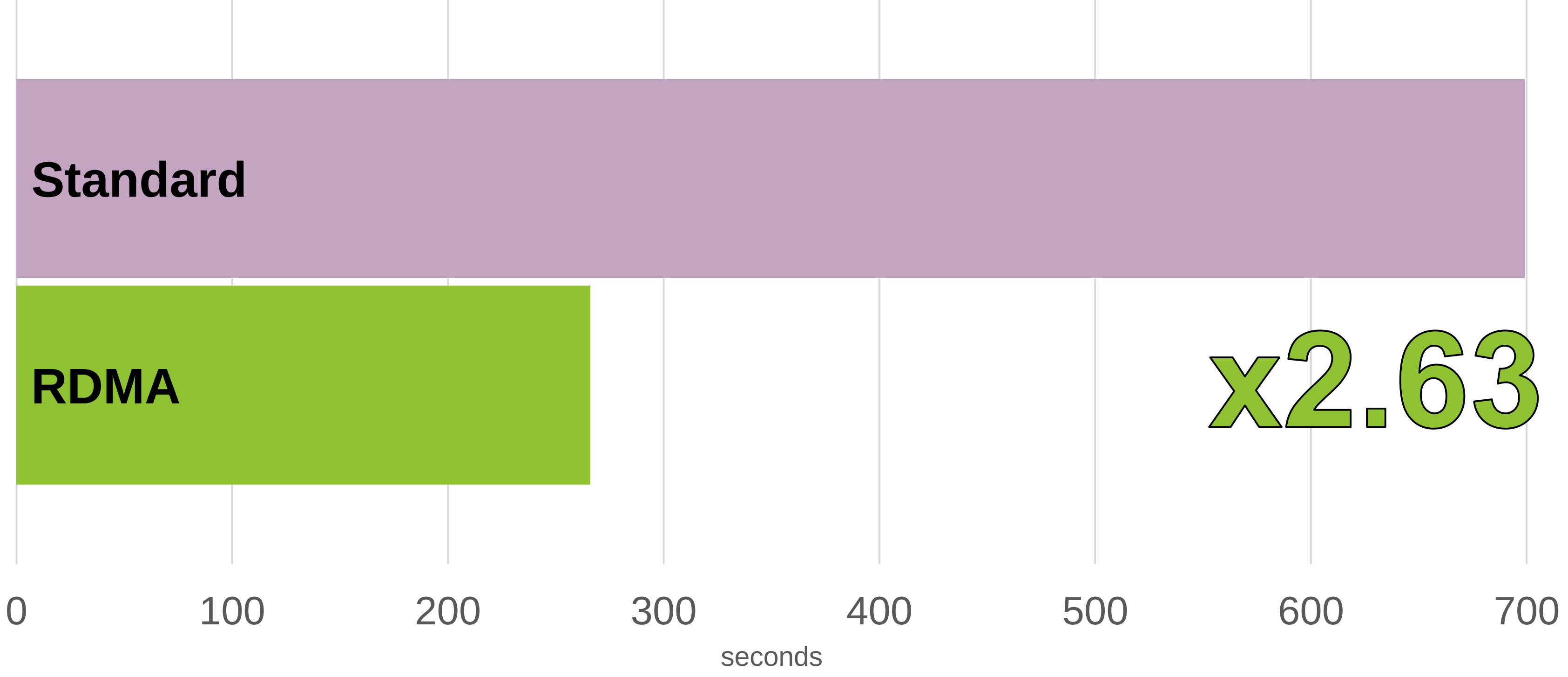






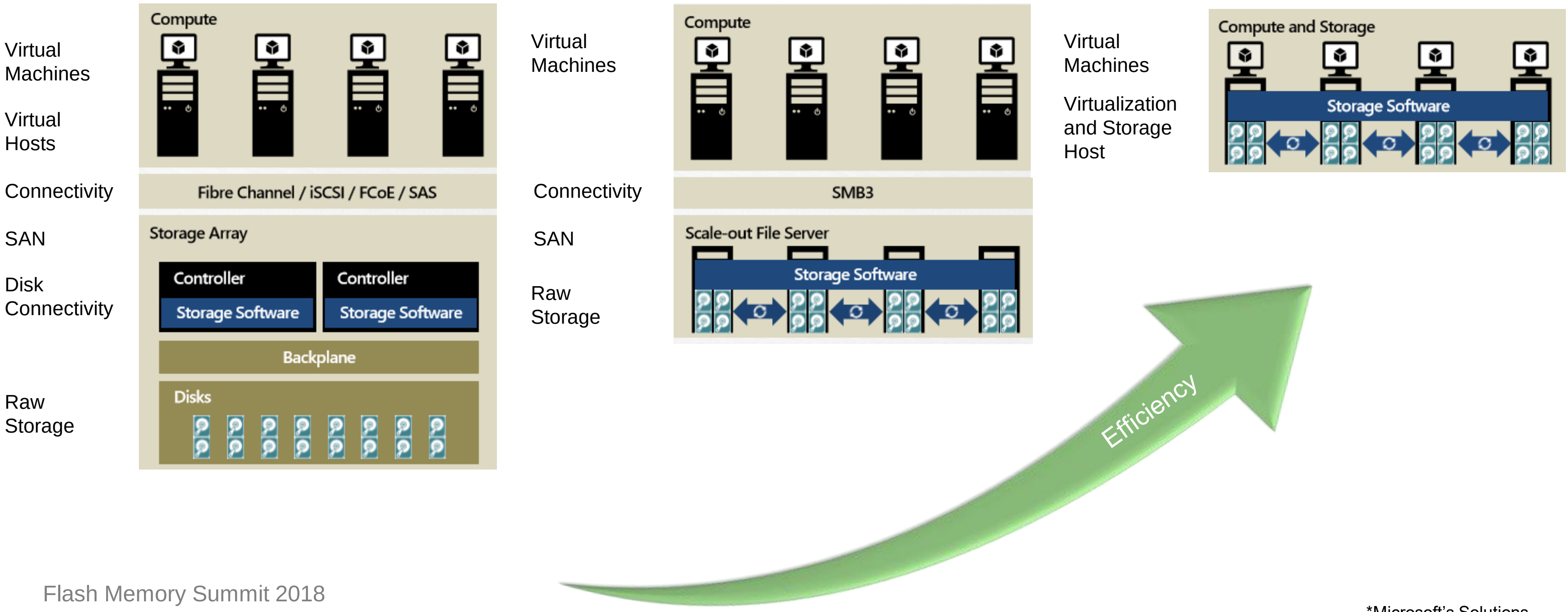
Source: Driving IBM BigInsights Performance Over GPFS Using InfiniBand+RDMA

TeraSort - Performance Results



Source: #HWCSAIS18

RDMA Enables Higher Performance SDS Solutions



Traditional Solution

Virtual Machines

Virtual Hosts

Connectivity

SAN

Disk Connectivity

Raw Storage

Compute

Fibre Channel / iSCSI / FCoE / SAS

Storage Array

Controller

Storage Software

Controller

Storage Software

Backplane

Disks

Converged Solution

Virtual Machines

Connectivity

SAN

Raw Storage

Compute

SMB3

Scale-out File Server

Storage Software

Hyperconverged Solution

Virtual Machines

Virtualization and Storage Host

Compute and Storage

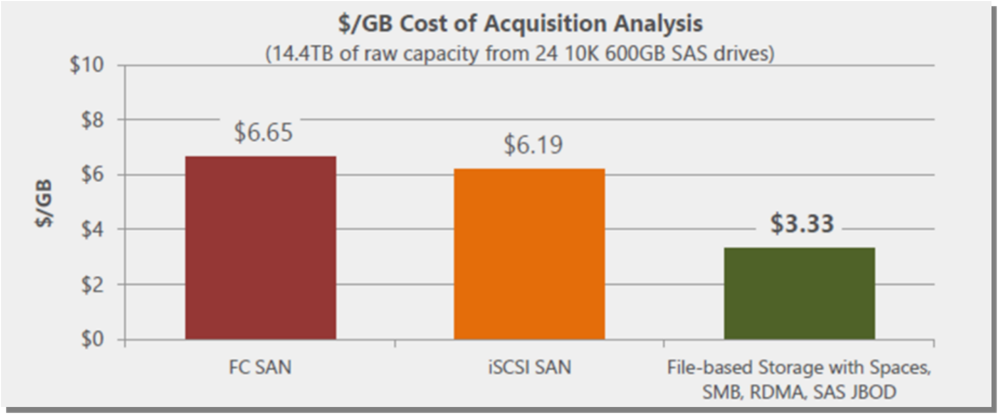
Storage Software

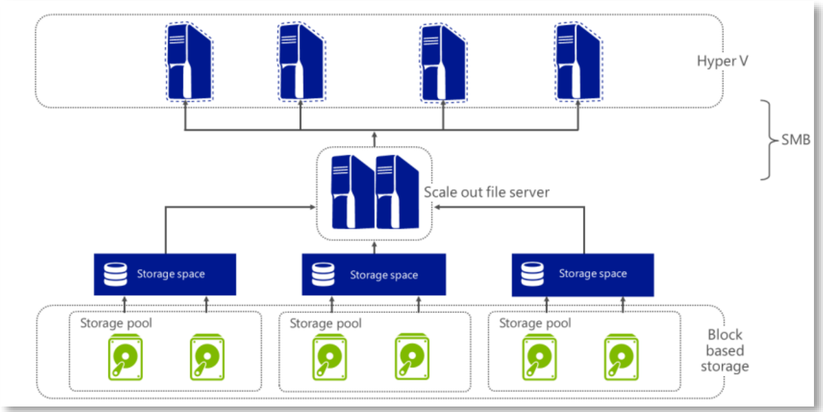
Efficiency

*Microsoft's Solutions

InfiniBand Cuts SAN Cost by 50%

- Delivers SAN-like functionality from the Windows Stack
  - Using SMB Direct (SMB 3.0 over RDMA)
- Utilize inexpensive, industry-standard, commodity hardware
  - Eliminate the cost of proprietary hardware and software from SAN solutions

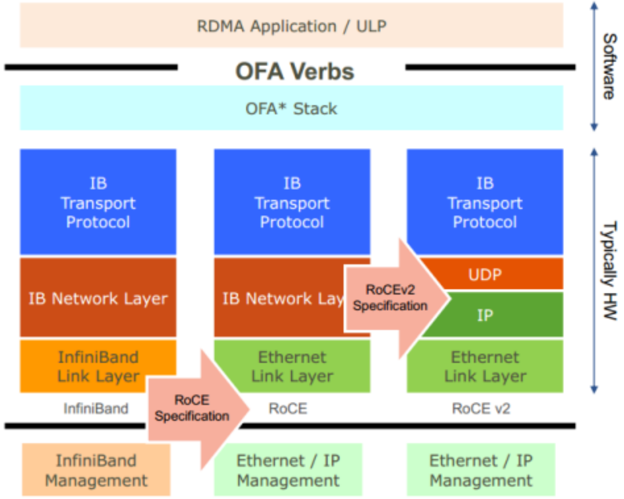




Source: Microsoft

RoCE – RDMA (InfiniBand) over Converged Ethernet

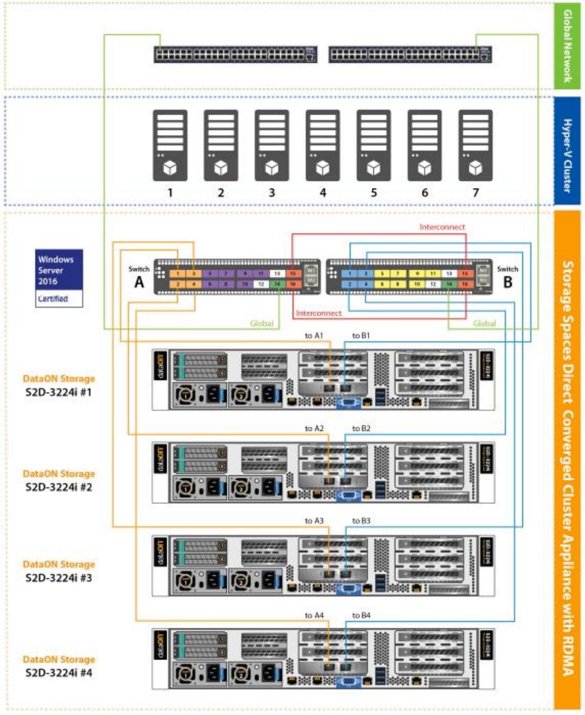
- InfiniBand transport over Ethernet
- API Compatible
- Efficient, light-weight transport, layered directly over
  - Ethernet – RoCE
  - UDP – RoCEv2
- Takes advantage of DCB Ethernet
  - PFC, ETS, and QCN



InfiniBand	LRH (L2 Hdr)	GRH (L3 Hdr)	BTH+ (L4 Hdr)	IB Payload	ICRC	VCRC		
RoCE	MAC	ET RoCE	GRH	BTH+	IB Payload	ICRC	FCS	
RoCEv2	MAC	ET IP	IP Proto UDP	UDP Port=RoCE	BTH+	IB Payload	ICRC	FCS

DataON WSSD* Hyper-Converged Infrastructure

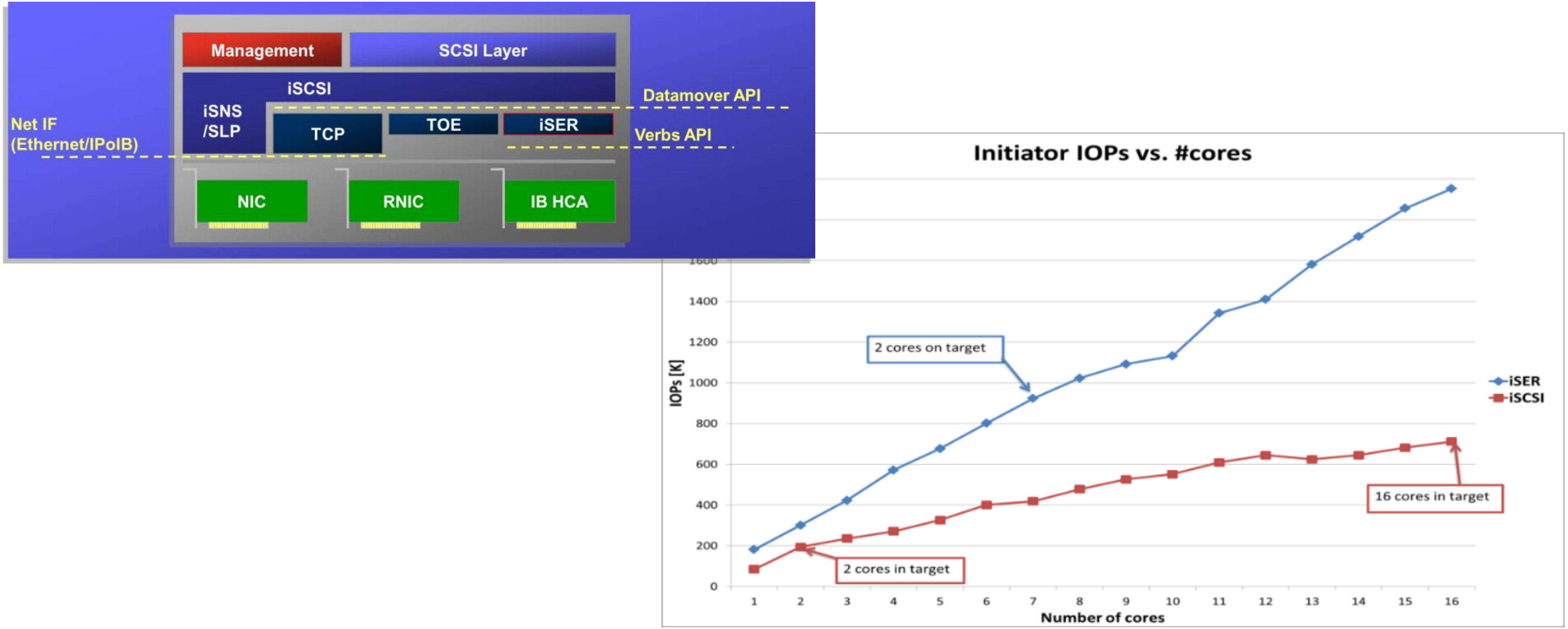
- Microsoft’s WSSD Certified
- RoCE networking
- Increased efficiency
  - 30X** vs. previous solution



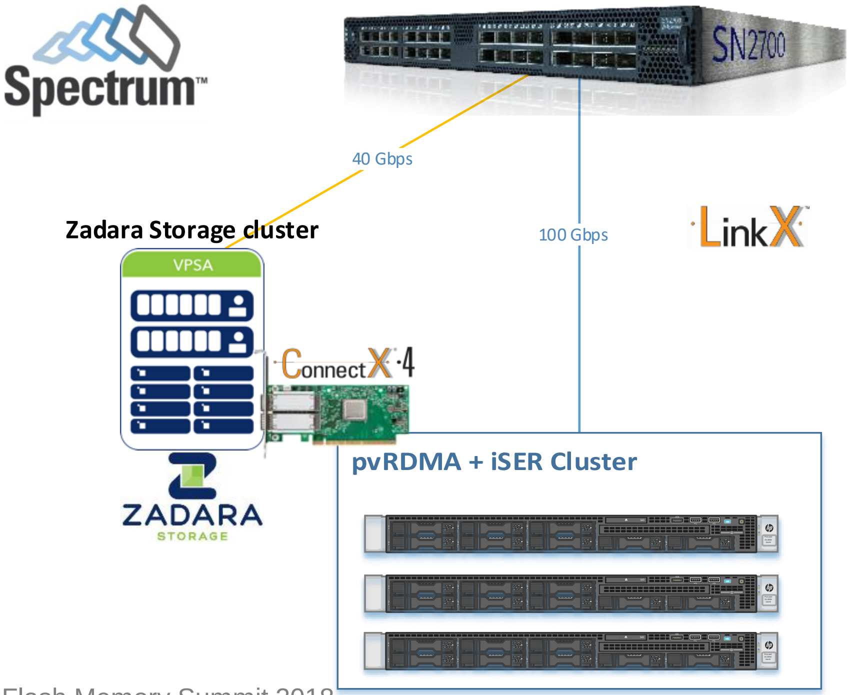
*Windows Server Software-Defined

** Source: DataON

iSER – iSCSI with RDMA Extensions

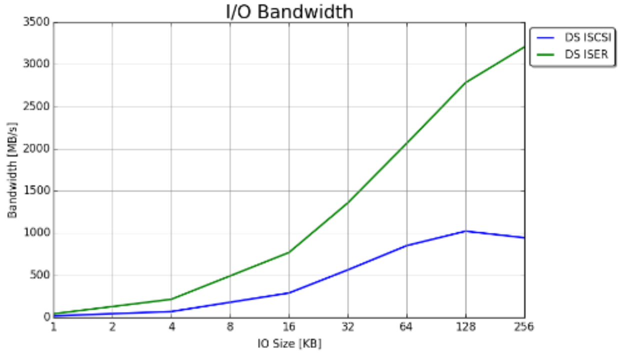


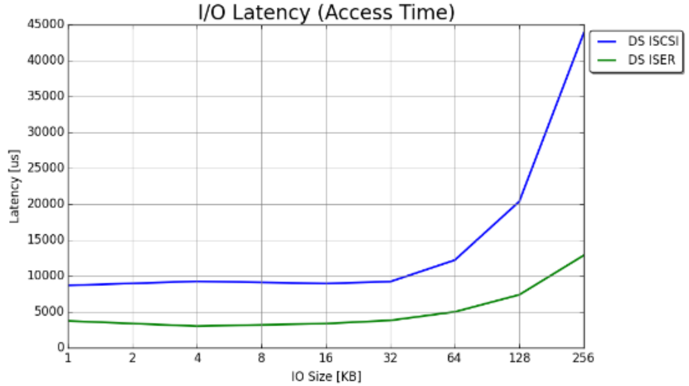
iSER Delivers 3X Higher Efficiency vs. iSCSI



Zadara Storage cluster

pvRDMA + iSER Cluster





RDMA enabled Networking Powers Modern Storage Platforms

DataDirect NETWORKS

dataon DataONSTORAGE

DELLEMC

FUJITSU

Hewlett Packard Enterprise

ORACLE

Microsoft

Micron

NetApp

NIMBUS DATA

IBM

xiv tms

SEAGATE

TOSHIBA

TERADATA

WD Western Digital

VIRIDENT

Higher Performance, Higher Efficiency and Higher Scalability

Curt Beckmann

Principal Architect at Brocade, is recently back to the Bay Area after 2 years in Paris where he held the role of CTO for Brocade Europe and last year wrote the 'NVMe over Fibre Channel for Dummies book'. Prior to that he led the architecture and development of storage virtualization ASICs for Rhapsody Networks, which was central to that firm’s successful acquisition by Brocade. He also led the ASIC/hardware design team for Nortel’s largest-unit-selling network switch. Beckmann’s combination of winning designs and customer-facing experience make him uniquely qualified to evaluate the design considerations of customer needs.

NVMe over Fibre Channel

Curt Beckmann

Principal Architect

Brocade Storage Networking, Broadcom

Today’s Presentation Topics

- Background: The why and how of sharing storage
- Enterprise and other storage categories
- The impact of Flash on Storage protocols
- The current state of NVMe/FC

Storage began as direct-attached. Why share it?

- Stored data as a durable *Information Asset*
  - Not like transient compute artifact (e.g. call stack)
  - Memory v Storage: Error handling? SLA?
- Desire to scale and leverage
  - Want to scale-out compute, re-use assets
- Stranded storage capacity
  - Spare capacity only usable by direct attached CPU

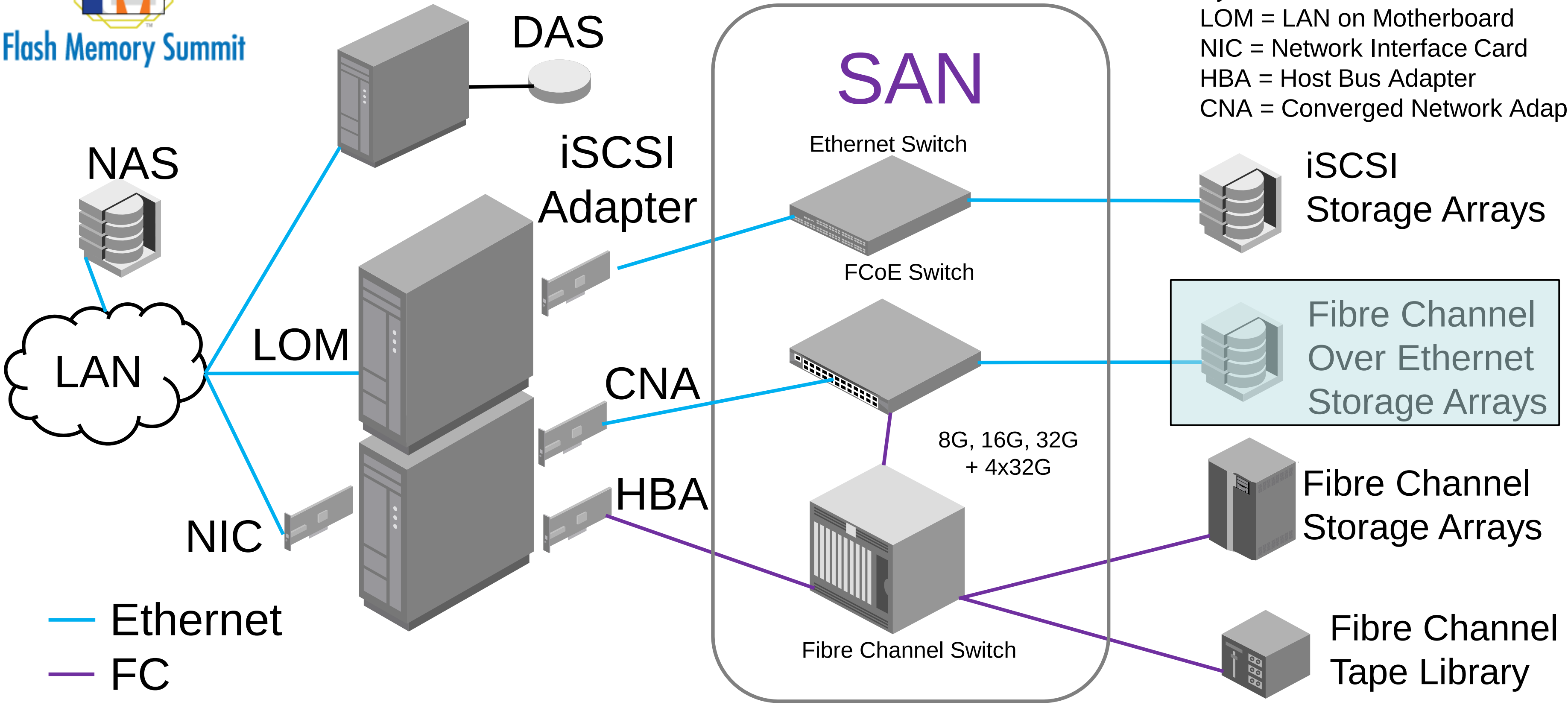
“Traditional” (20th C) shared storage concepts

- Files: “NAS”:
  - Enet/IP/L4: NFS, SMB/CIFS…
- Blocks (structured, strictly consistent, mission critical): “SAN”
  - Networked SCSI: SAS, FCP…
- Enduring wish: Consistency / Availability / Partition (CAP) Theorem
  - Span, cost, performance, availability/reliability, size

- Ethernet / IP / Layer 4: Rose to dominance in 1990’s
  - Best-effort/retry, Internet-wide, “converged”, commodity (span/cost)
- Fibre Channel: born in Ethernet/IP heyday
  - Lossless, DC-wide, storage-centric, “Enterprise” (performance/availability)

Storage Types

DAS = Direct Attached Storage
NAS = Network Attached Storage
iSCSI – Internet Small Computer Systems Interface
LOM = LAN on Motherboard
NIC = Network Interface Card
HBA = Host Bus Adapter
CNA = Converged Network Adapter



Source: http://www.ieee802.org/3/ad_hoc/bwa/public/sep11/kipp_01a_0911.pdf

“Recent” (21st C) shared storage concepts

- InfiniBand (and Omni-Path… etc?):
  - Lossless, DC-wide, compute-centric (HPC), popularized RDMA
- “3rd platform”: Mobile + Cloud, IoT
  - Virtualized, commoditized / converged, “shared nothing”, “cattle” v. “pets”
- New use cases, “evolved” choices for CAP theorem
  - Big Data / “SDS” / “Eventual Consistency” / AI-ML / DevOps (flexible) mindset
- Flash broke out of niche: scale, write endurance, $/GB
  - Flash’s disruptive speed has moved focus to various sluggish software
- NVMe stack slims away decades of SCSI baggage
  - “NVMe” is PCI-based, “NVMe-over-Fabrics” (coming slides) for shared use cases

Categorization (storage-oriented)

	CapEx*	Performance	Reliability	Maturity
Fibre Channel	1.00	High	High	High
NAS (NFS, etc, over IP)	0.68	Low-Medium	Medium	High
iSCSI	0.59	Medium-High	Medium	High
DAS	0.46	High	High	High
Mainframe (FICON)	1.63	High	High	High
InfiniBand	1.43	High	High	Low
SAS SAN	0.70	Medium	Medium	Low
FCoE	0.79	High	Medium	Medium
NVMe over Fabrics	n/a	High**	High**	Low

*Normalized to FC Cost/GB in 2016 prices (Ref: IDC)

**Projected

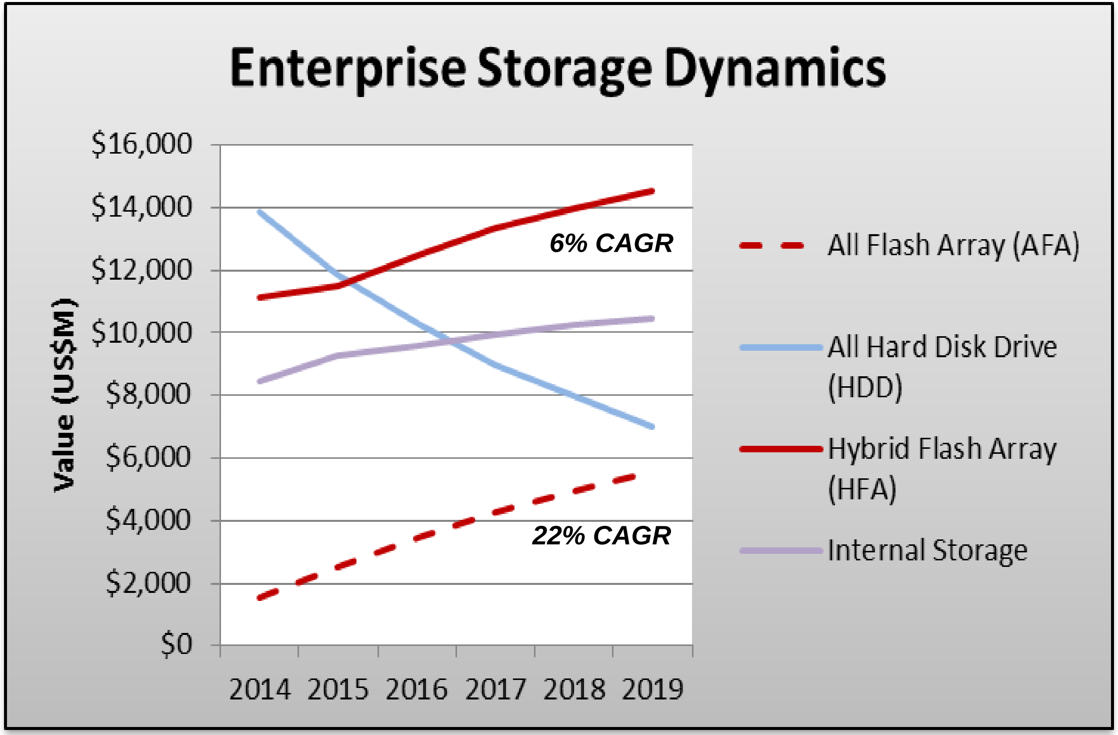
How Fibre Channel differs from Ethernet: Tech

- Technical:
  - Fewer, more coupled layers, limited application
  - Smaller address range, smaller header
  - Addresses assigned (not random or learned)
    - Scales bigger than typical subnet, but smaller than Internet
  - Not much multicast, no flooding
  - Always supported fabric topology (not just Spanning Tree)
  - Always built for reliable delivery (v. best effort)
  - Credit-based flow control is “always on”
  - Fabric provides fabric-resident services: Name server, etc

How Fibre Channel differs from Ethernet: Industry

- Industry:
  - Focus in critical “always on” use cases
    - Nearly always redundant fabrics dedicated to storage
  - Few switch / HBA firns mostly selling through storage vendors
  - Storage vendors certify products, mark them up, provide support
  - Interoperability driven by storage vendors
  - Vendor arrays loaded w enterprise features, virtualization
    - Rarely expose raw media
  - Upshot: most benchmarks are based on full featured arrays
    - With SSDs getting so fast, software features now a large fraction of the latency
    - When tested on raw media (Linux JBOFs), FC latency comparable to PCI-attached

Enterprise Flash Growing Well



Source: IDC September 2015 WW Quarterly Disk Storage Systems Forecast

NVMe over Fabrics Concepts

- NVMExpress.org defined specs
  - PCIe-based NVMe (1.0 in 2011, currently at 1.3)
  - NVMe-over-Fabrics (1.0 in 2016)
- Four early fabrics, one newcomer
  - (RDMA-based) InfiniBand, iWARP, RoCE(v2)
  - (no RDMA) Fibre Channel
  - (no RDMA, iSCSI-like newcomer) NVMe-over-TCP

FC-NVMe Spec Status

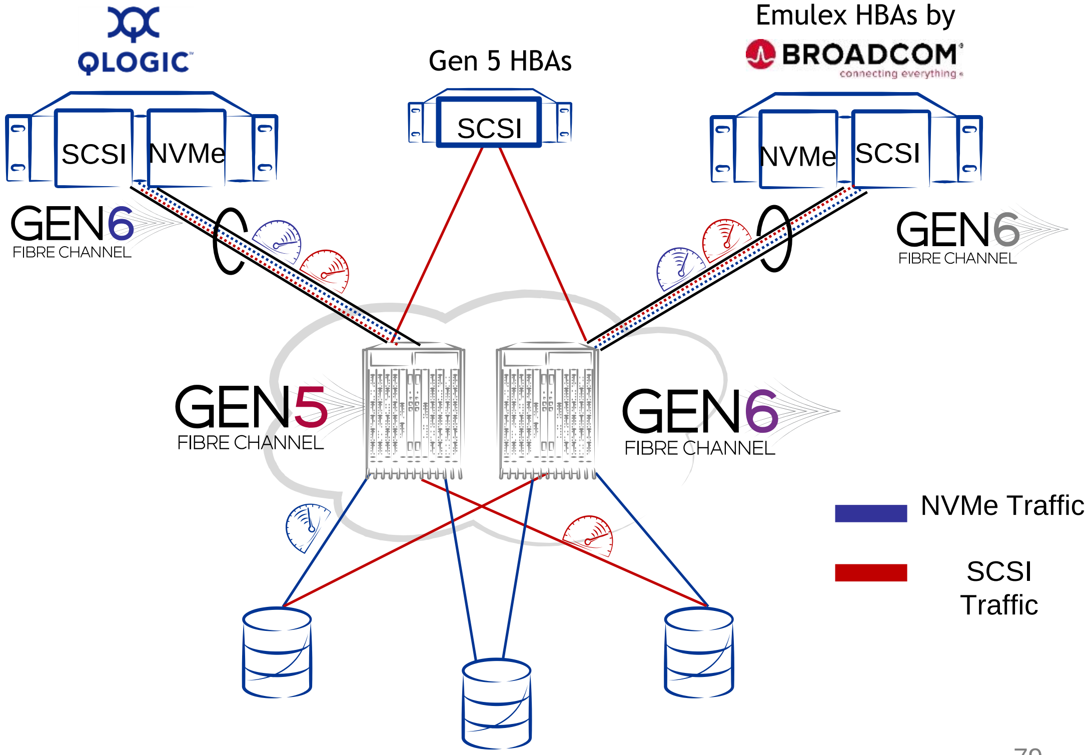
- Why move to NVMe/FC?
  - It’s like SCSI/FC tuned for SSDs and parallelism
  - Simpler, more efficient, and (as we’ll see) faster
- FC-NVMe standard effort is overseen by T11
  - T11 and INCITS finalized FC/NVMe early 2018
- Several vendors are shipping GA products
- FCIA plugfest last week: XX participants

Dual Protocol SANs lower risk, help NVMe adoption

- 80% of today's Flash arrays connect via FC
  - This is where most vital data assets (still!) live today
- High-value Assets require protection
  - Storage Teams avoid risk…part of job description
  - How can Storage Teams adopt NVMe with low risk?
    - Use familiar, trusted infrastructure, vendors and support
    - Dual protocol SAN offers that, and NVMe performance too…

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



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.

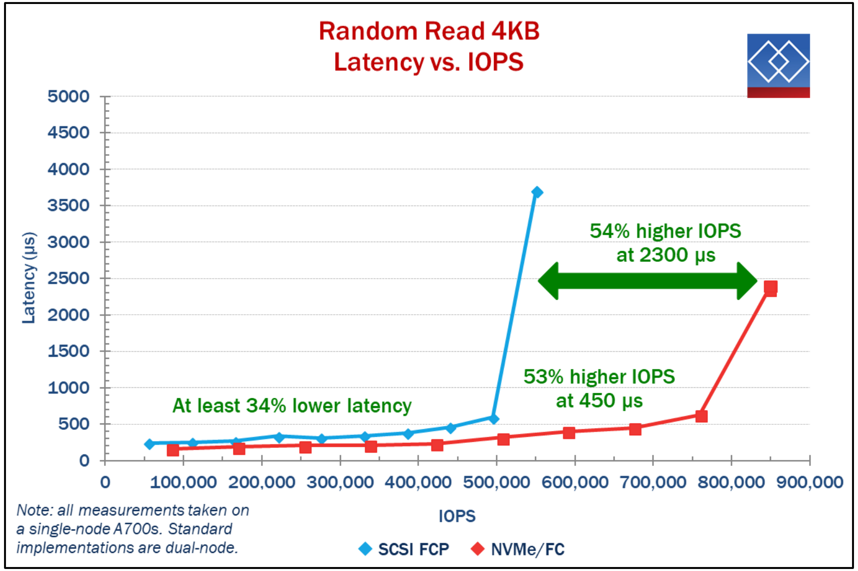
> ***NVMe/FC protects your investment:*** The benefits we observed were with existing hardware that supports 32GFC.

> ***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 overall array performance

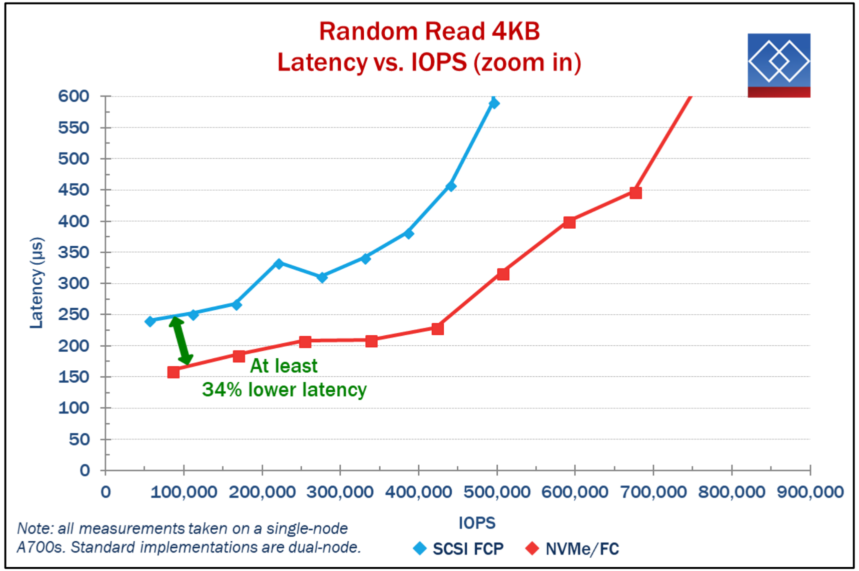
Results: 4KB Random Reads, full scale and zoomed in



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

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



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

Summary

- Shared storage
  - Data asset has value independent of any application
  - Need more protection!
    - Even if it adds some access time
- With slight inefficiency, SCSI has dominated
- SSDs are so fast, SCSI burden no longer slight
  - NVMe command set o

J Metz

J Metz is R&D Engineer, Office of the CTO, at Cisco, where he focuses on examining and deploying directions for storage strategy. He was previously Strategic Product Manager, Storage and Unified Fabric. With Cisco since 2010, he has previous experience with QLogic and Apple. He has also been President at Communiweb Communications and an Assistant Professor at the University of Central Florida. He holds a PhD from the University of Georgia, an MA from the University of South Dakota, and a BA from the University of Rhode Island

Flash Memory Summit

Ethernet-Networked Flash Storage

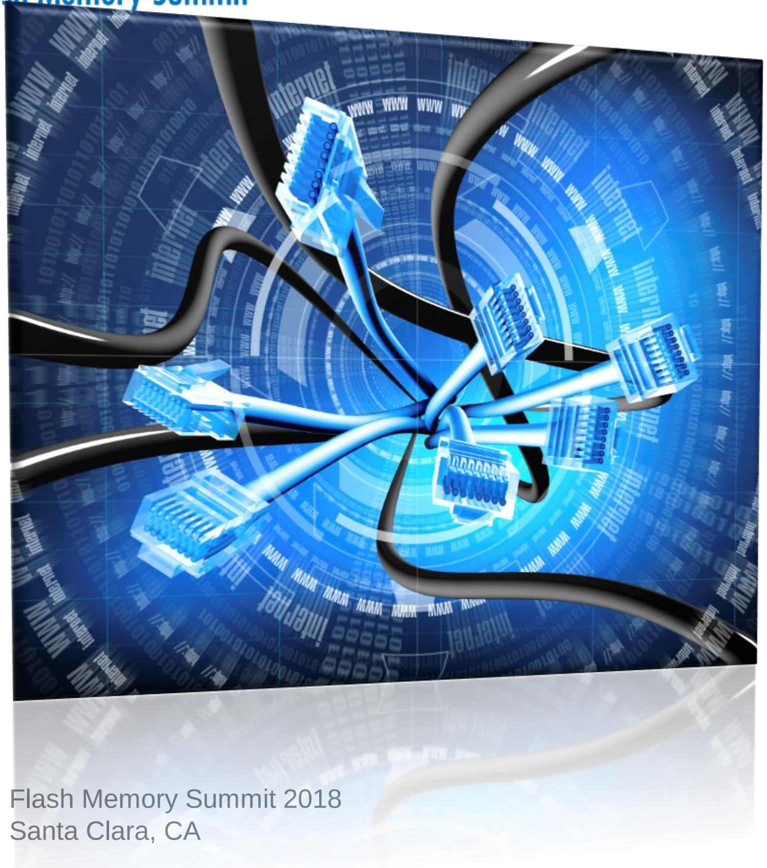
J Metz, Ph.D

R&D Engineer, Advanced Storage

Cisco Systems

@drjmetz

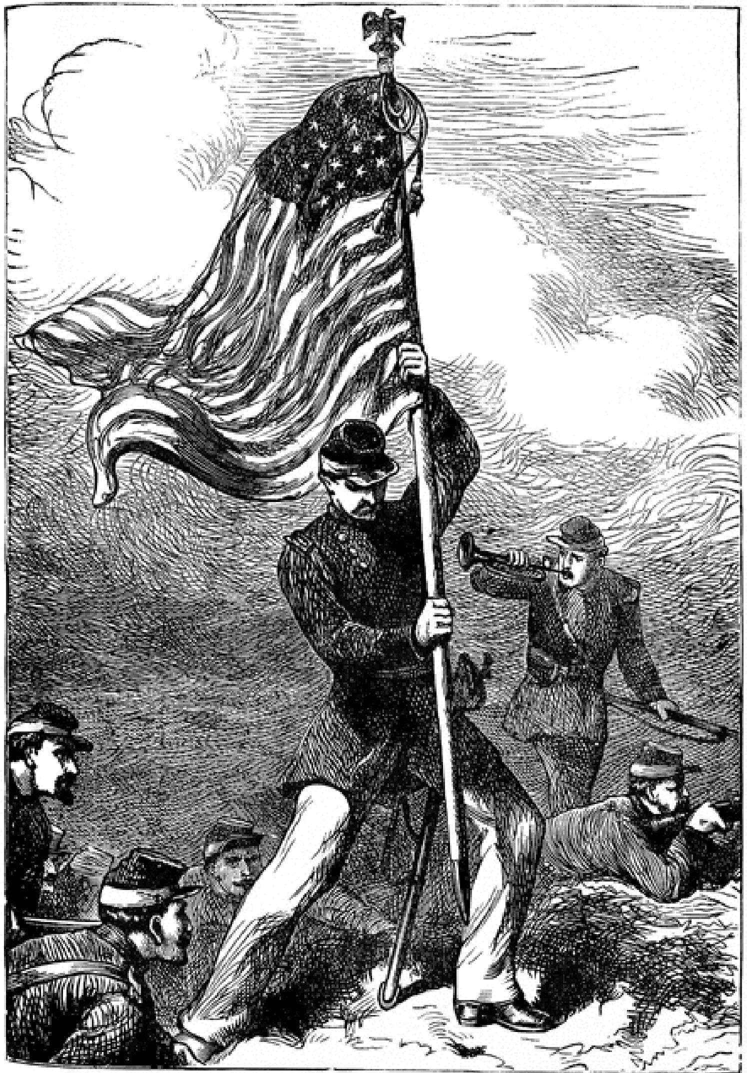
Agenda



- Ethernet Background and Roadmap
- Storage Use Cases
- Goodness of Fit

Planting a Flag

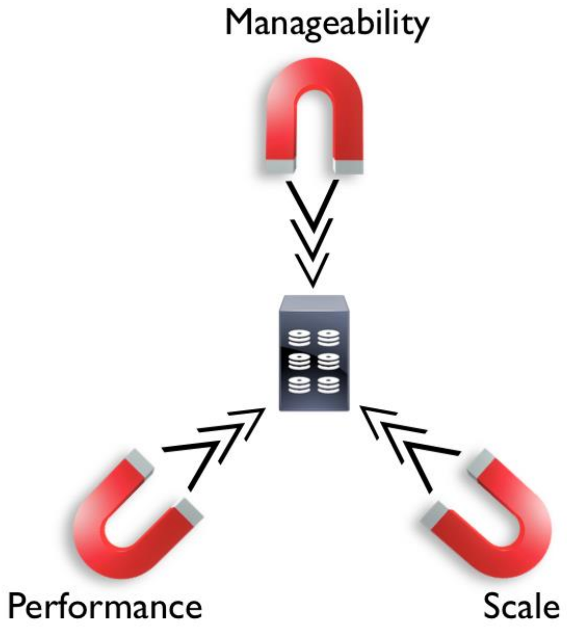
- Is there anyone who thinks Ethernet will *not* play a role in storage?



Then the Question Is…

…how best to use Ethernet for Storage?

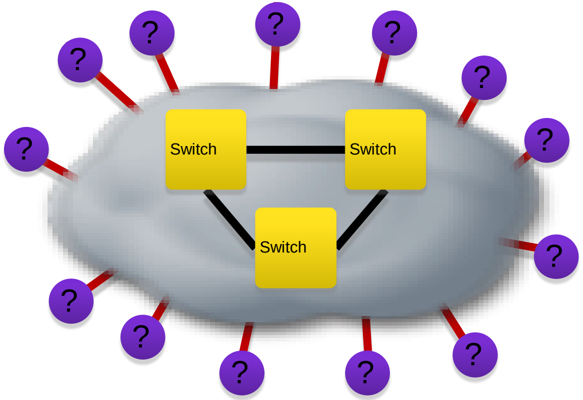
Storage Perspective



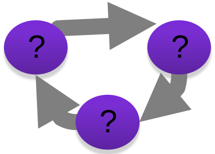
- There is a "sweet spot" for storage
  - Depends on the workload and application type
  - No "one-size fits all"
- What is the problem to be solved?
  - Deterministic or non-deterministic?
  - Highly scalable or highly performant?
  - Level of manageability?
- Understanding "where" the solution fits is critical to understanding "how" to put it together

Network Determinism

- Non-Deterministic
  - Provide any-to-any connectivity
  - Storage is unaware of packet loss – relies on ULPs for retransmission and windowing
  - Provide transport w/o worrying about services
  - East-West/North-South traffic ratios are undefined
- Examples
  - NFS/SMB
  - iSCSI
  - iSER
  - iWARP
  - (Some) NVMe-oF

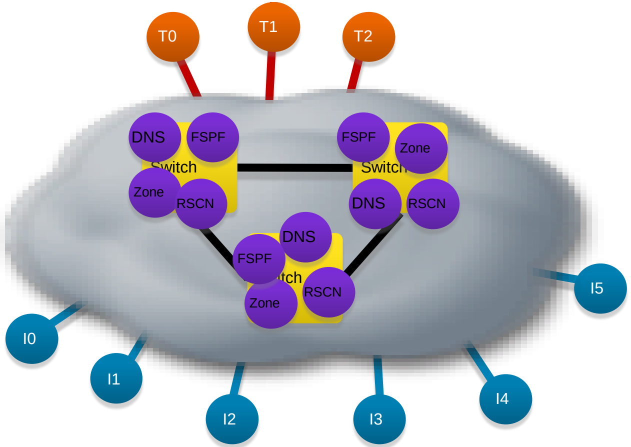


Fabric topology and traffic flows are highly flexible

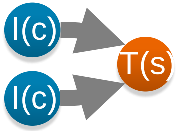


Client/Server Relationships are not pre-defined

Network Determinism (cont.)



Fabric topology, services and traffic flows are structured

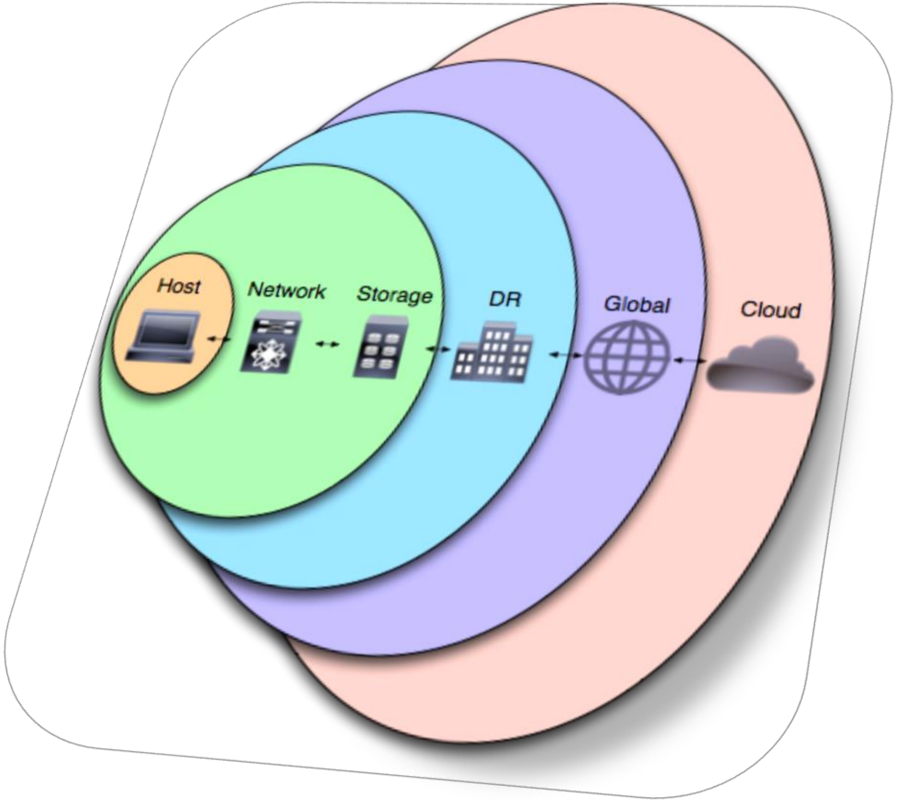


Client/Server Relationships are pre-defined

- Deterministic Storage
  - Goal: Provide 1:1 Connectivity
  - Designed for Scale and Availability
  - Well-defined end-device relationships (i.e., initiators/targets)
  - Only north-south traffic; east-west mostly irrelevant
- Examples
  - Fibre Channel
  - Fibre Channel over Ethernet
  - InfiniBand
  - RoCE
  - (Some) NVMe-oF

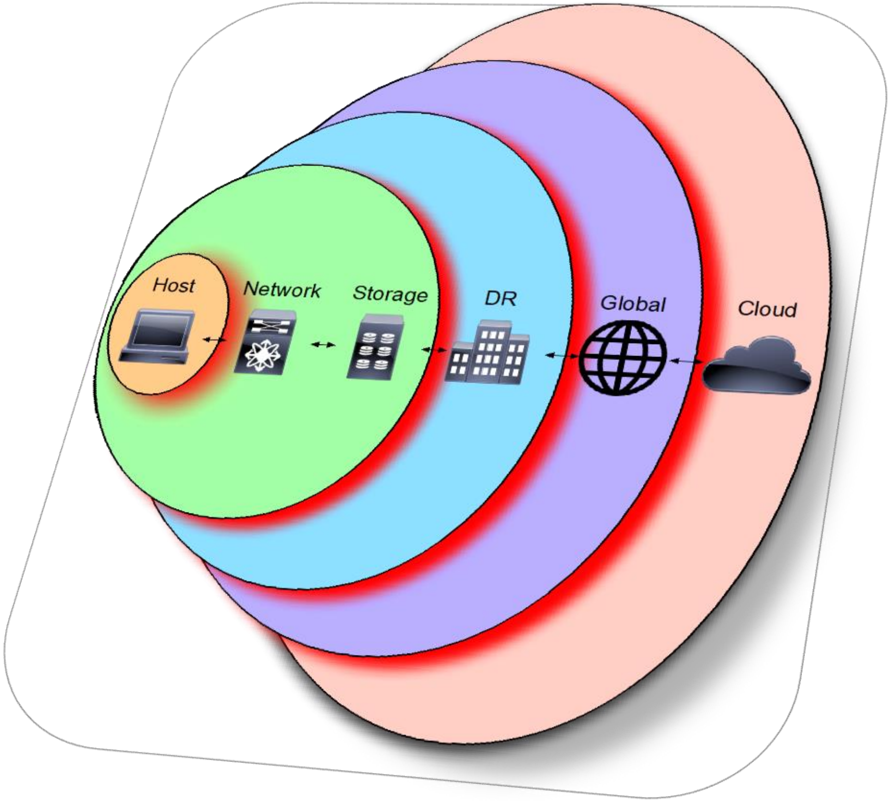
Big Picture

- Many ways to solve a problem
  - No “one-size-fits-all”
- Lots of overlap
  - Can easily get confused about which to choose
  - If two different approaches can do the same thing, how do you know what to do?

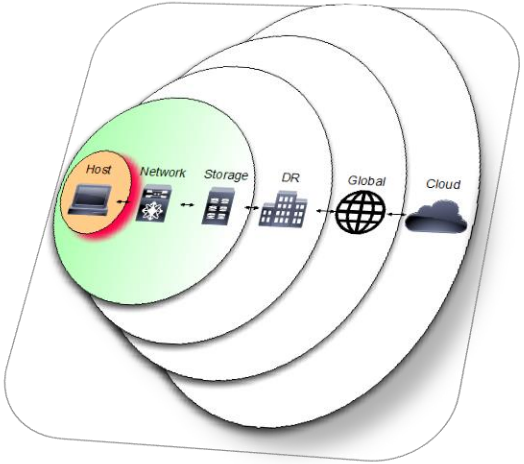


Big Picture

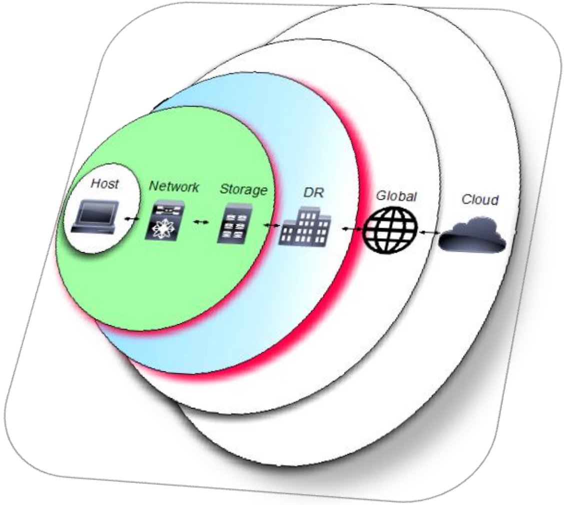
- When you miss the sweet spot, you risk major problems
  - Careful of the “Danger Zones”



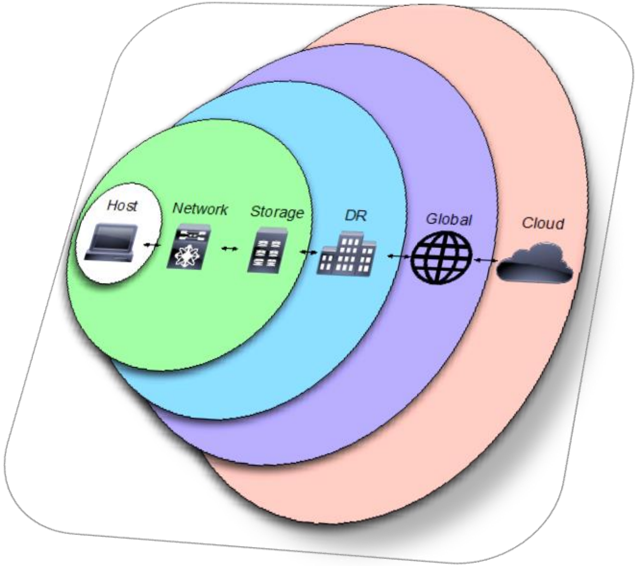
Scope Comparison



PCIe

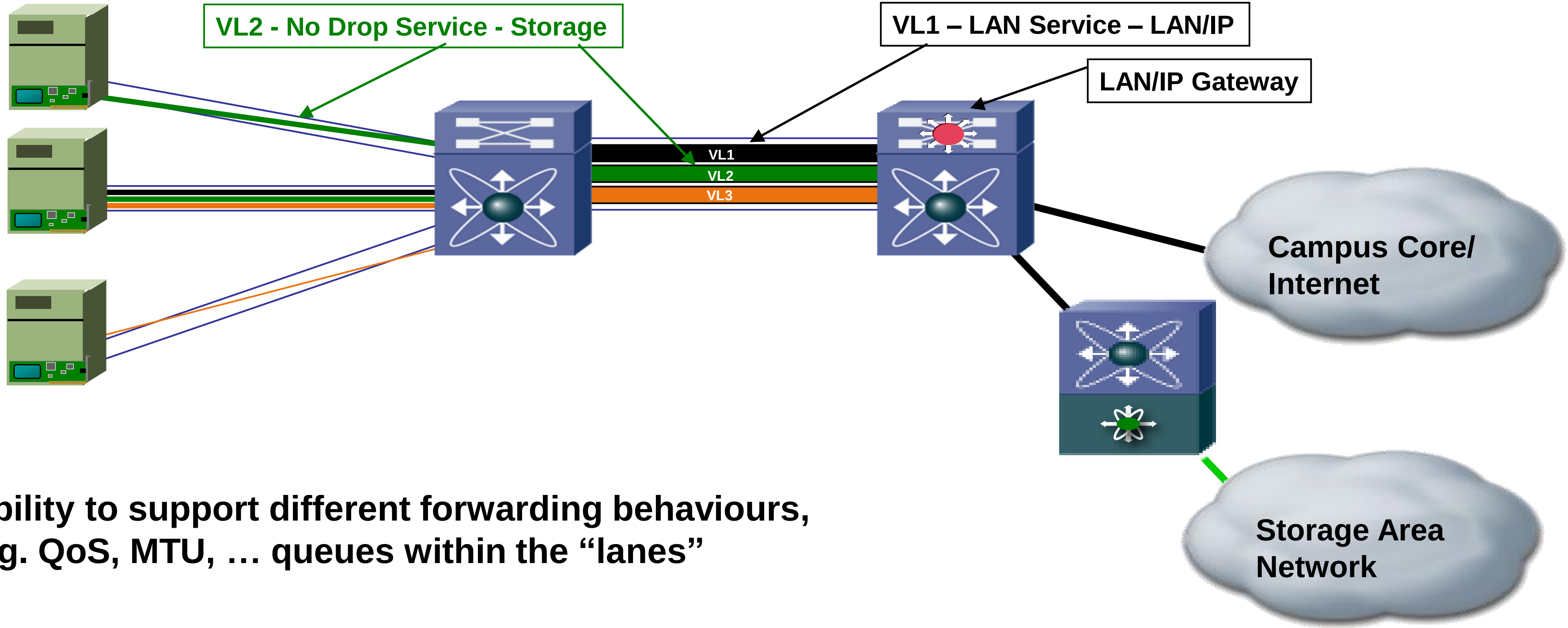


Fibre Channel
Ethernet (FCoE, iSCSI,iSER, NVMe-oF)
InfiniBand



Ethernet (NFS, SMB, Object)

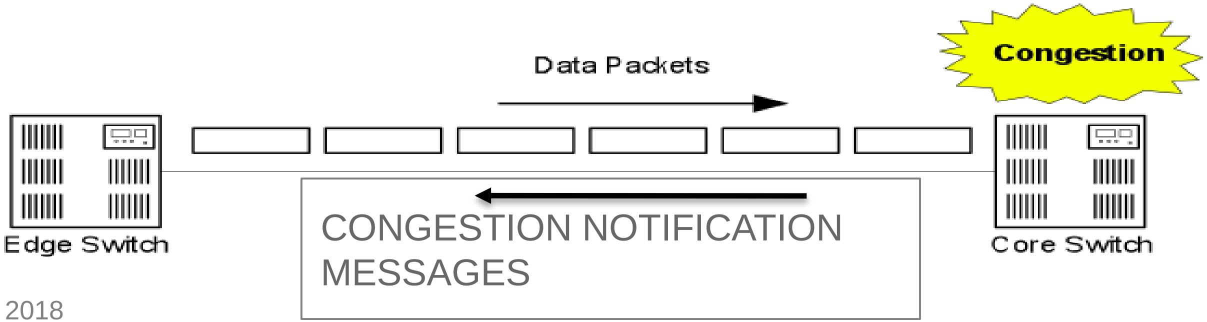
Ethernet Enhancements



Ability to support different forwarding behaviours, e.g. QoS, MTU, … queues within the “lanes”

Congestion Notification: BCN/QCN

- Principles
  - Push congestion from the core towards the edge of the network
  - Use rate-limiters at the edge to shape flows causing congestion
  - Tune rate-limiter parameters based on feedback coming from congestion points
- Inspired by TCP
- Self-Clocking Control loop
- Derived from FCC (Fbire Channel Congestion Control)



CONGESTION NOTIFICATION MESSAGES

DCTCP

Data Center TCP

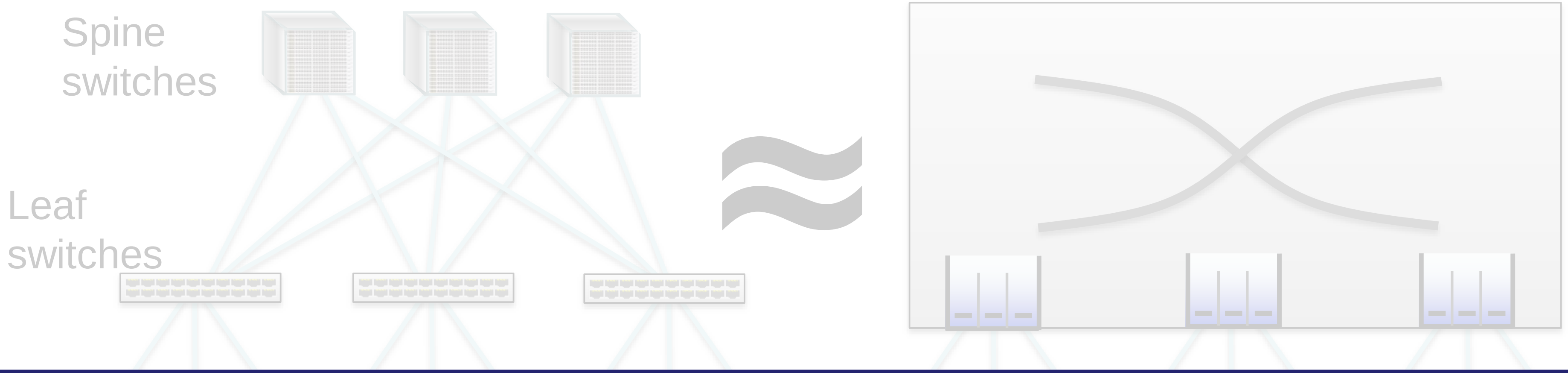
- Congestion indicated quantitatively (reduce load prior to packet loss)
- React in proportion to the extent of congestion, not its presence
  - Reduces variance in sending rates, lowering queuing requirements

ECN Marks	TCP	DCTCP
1 0 1 1 1 1 0 1 1 1	Cut window by 50%	Cut window by 40%
0 0 0 0 0 0 0 0 0 1	Cut window by 50%	Cut window by 5%

- Mark based on instantaneous queue length
  - Fast feedback to better deal with bursts

Leaf-Spine DC Fabric

Approximates ideal output-queued switch



- How close is Leaf-Spine to ideal OQ switch?
- What impacts its performance?
  - Link speeds, oversubscription, buffering

Comparison

	Ethernet	PCIe	Fibre Channel	InfiniBand
Intra-Host	No	*Yes*	No	No
Direct Attached (DAS)	***Yes***	*Yes*	*Yes*	*Yes*
Network Attached (NAS)	***Yes***	No	No	No
Storage-Area Network (SAN)	***Yes***	No	*Yes*	*Yes*
Deterministic Capability	***Yes***	*Yes*	*Yes*	*Yes*
Non-Deterministic Capability	***Yes***	No	No	No
Block Storage	***Yes***	*Yes*	*Yes*	*Yes*
File Storage	***Yes***	No	No	No
Object Storage	***Yes***	No	No	No
Global Distance	***Yes***	Hell no	No	No

Summary

- Ethernet
  - General Purpose network designed to solve many, many problems and do it well
  - Flexible for all but the most extreme conditions
  - Largest ecosystem of developers, vendors, and users
  - From the smallest system to the largest, there is no other networking technology more suited, or best understood

Ilker Cebeli

Ilker is a Senior Director of Product Planning at Samsung. He is responsible for leading the emerging memory, SSD, and all-flash-array related storage solutions and technologies. He has spent 25 years in enterprise computing, storage and networking working in various roles. Prior to joining to Samsung, Ilker worked at Micron, and was leading and directing emerging memory projects in memory division. Ilker also spent 15 years at Intel and he was responsible for Intel's Xeon™ product planning and server platform architecture definition.

NVMe over Fabrics

High Performance SSDs networked over Ethernet

Ilker Cebeli
Senior Director of Product Planning, Samsung

August 8th , 2017

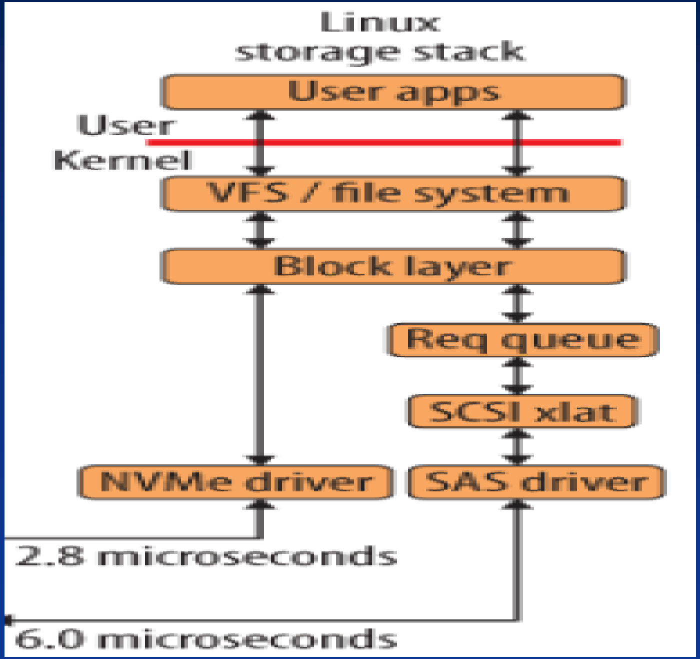
Disclaimer

This presentation and/or accompanying oral statements by Samsung representatives collectively, the “Presentation”) is intended to provide information concerning the SSD and memory industry and Samsung Electronics Co., Ltd. and certain affiliates (collectively, “Samsung”). While Samsung strives to provide information that is accurate and up-to-date, this Presentation may nonetheless contain inaccuracies or omissions. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of the information provided in this Presentation.

This Presentation may include forward-looking statements, including, but not limited to, statements about any matter that is not a historical fact; statements regarding Samsung’s intentions, beliefs or current expectations concerning, among other things, market prospects, technological developments, growth, strategies, and the industry in which Samsung operates; and statements regarding products or features that are still in development. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements in this Presentation. In addition, even if such forward-looking statements are shown to be accurate, those developments may not be indicative of developments in future periods.

NVMe Technology – Background

- Optimized for flash
  - Traditional SCSI designed for disk
  - NVMe bypasses unneeded layers
  - Dramatically reducing latency

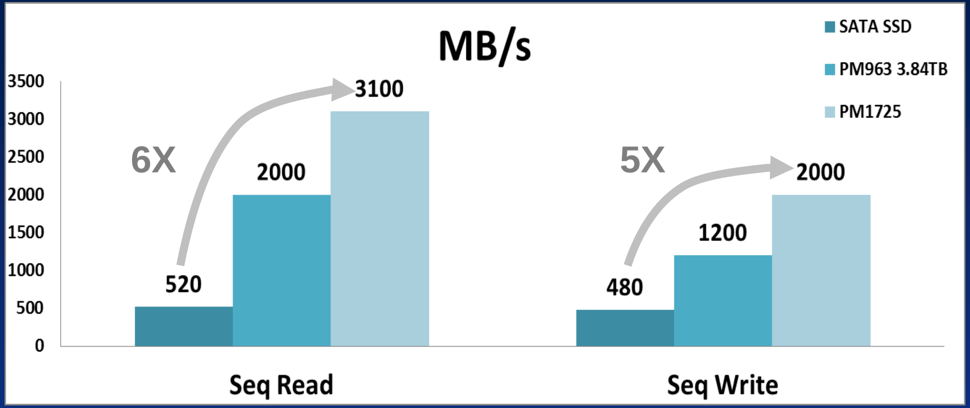


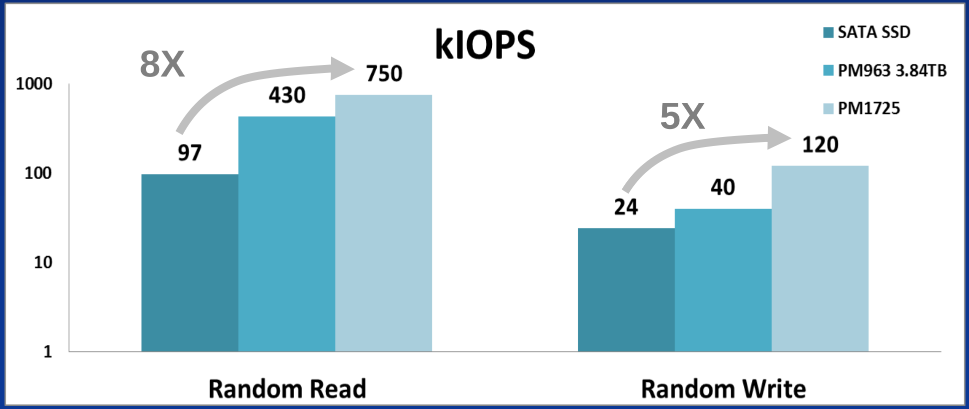
NVMe Design Advantages

- **Lower latency**
  - **Direct connection to CPU’s PCIe lanes**
- **Higher bandwidth**
  - **Scales with number of PCIe lanes**
- **Best in class latency consistency**
  - **Lower cycles/IO, fewer cmds, better queueing**
- **Lower system power**
- **No HBA required**

NVMe Technology – Background

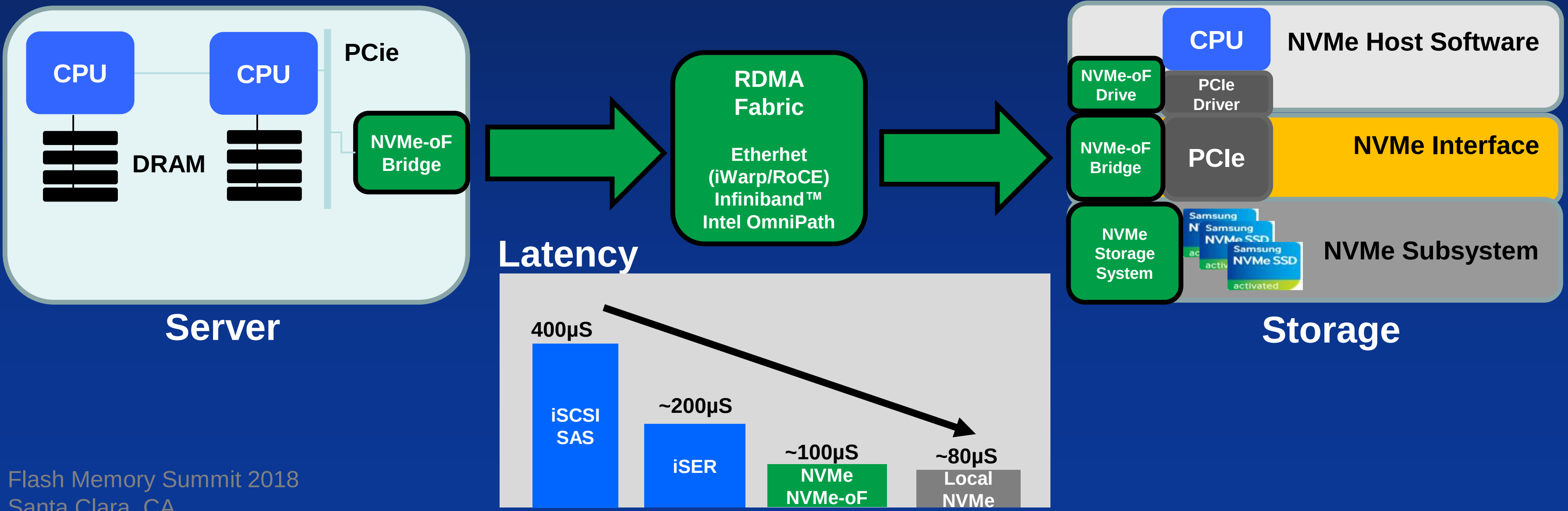
- NVMe outperforms SATA SSDs
  - 5X-6X more bandwidth,
  - 40-50% lower latency
  - Up to 8x more IOPS



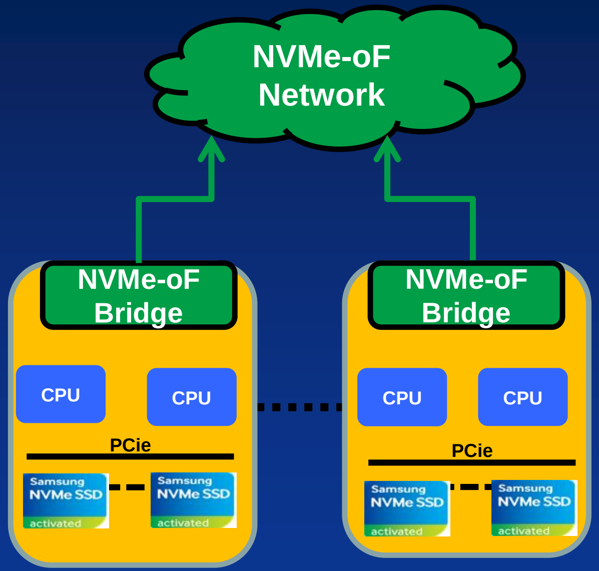


What is NVM Express Over Fabrics?

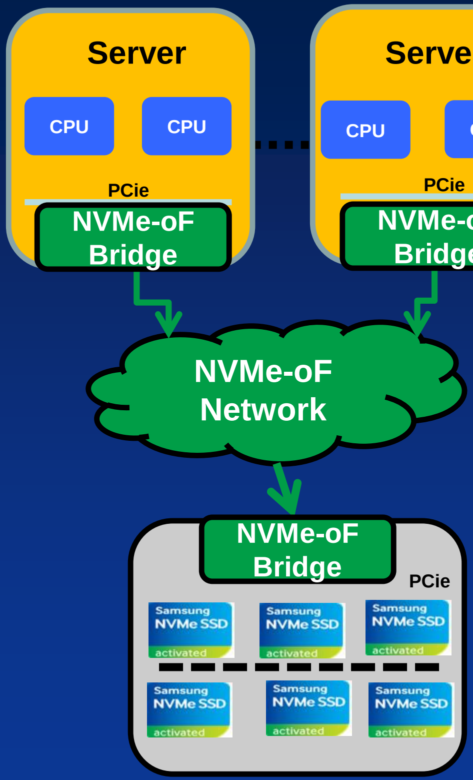
- A protocol interface to NVMe that enable operation over other interconnects (e.g., Ethernet, InfiniBand™, Fibre Channel).
- Shares the same base architecture and NVMe Host Software as PCIe
- Enables NVMe Scale-Out and low latency (<10µS latency) operations on Data Center Fabrics
- Avoids protocol translation (avoid SCSI)



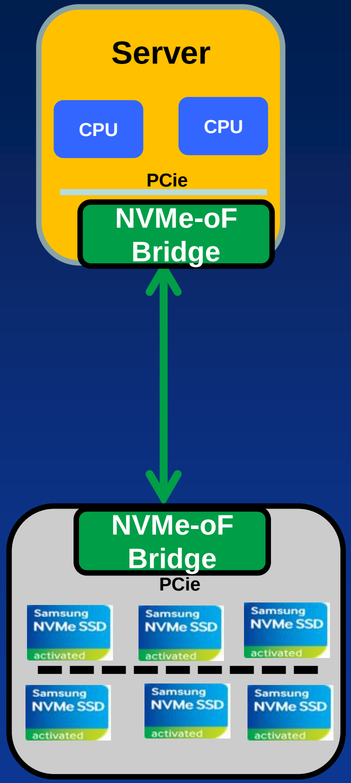
Some of the use cases for NVMe Over Fabrics


NVMe-oF
Network
NVMe-oF
Bridge
CPU
CPU
PCIe
Samsung NVMe SSD activated
Samsung NVMe SSD activated
NVMe-oF
Bridge
CPU
CPU
PCIe
Samsung NVMe SSD activated
Samsung NVMe SSD activated

1. Software-Defined Storage (SDS)
2. Hyper-Converged


Server
CPU
CPU
PCIe
NVMe-oF
Bridge
Server
CPU
CPU
PCIe
NVMe-oF
Bridge
NVMe-oF
Network
NVMe-oF
Bridge
PCIe
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated

Disaggregated JBOF Storage

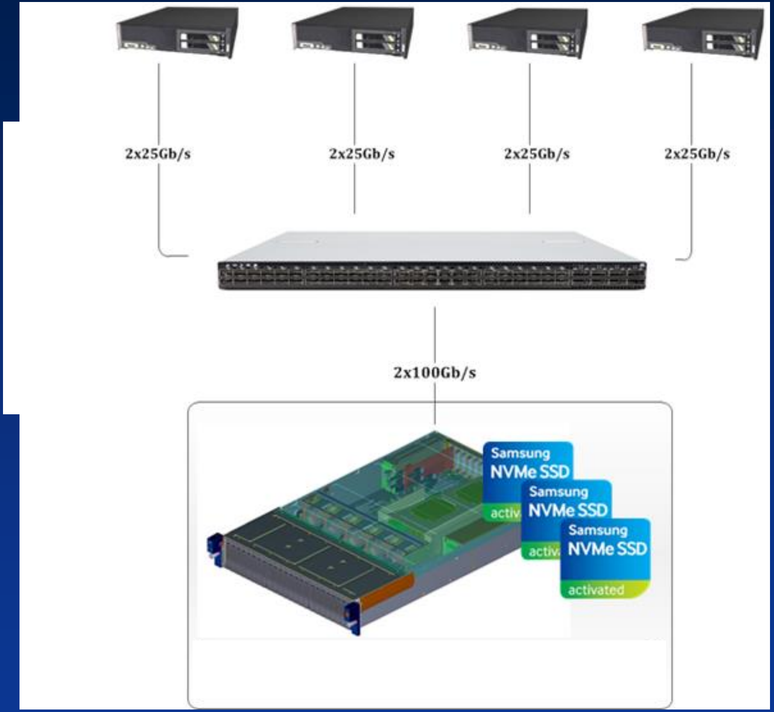

Server
CPU
CPU
PCIe
NVMe-oF
Bridge
NVMe-oF
Bridge
PCIe
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated
Samsung NVMe SSD activated

Direct Attached JBOF

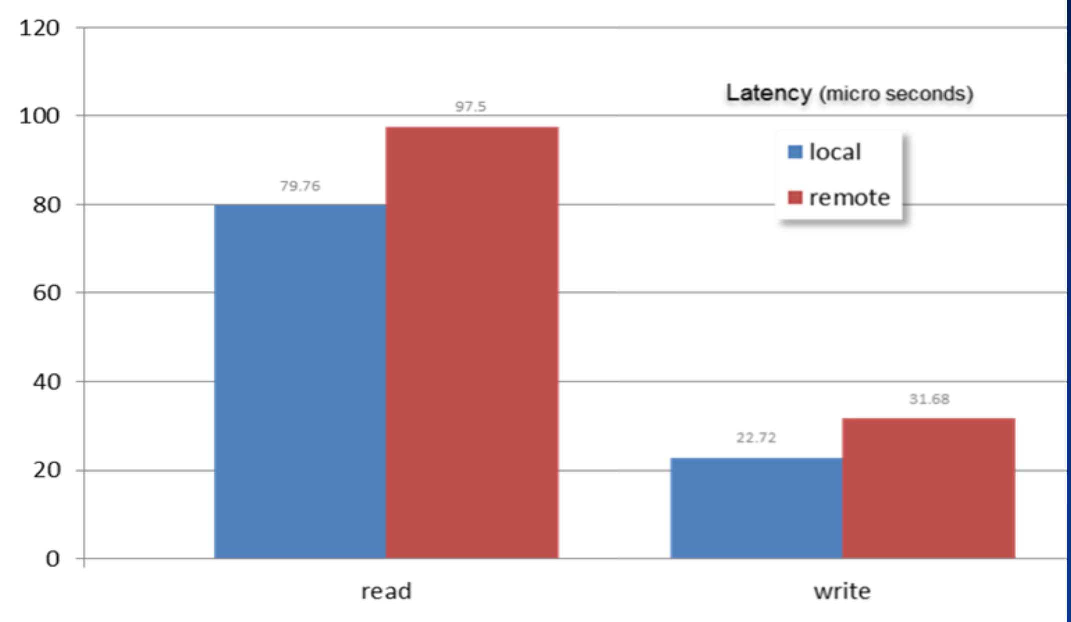
SAS DAS Replacement

Performance Test Configuration – 2016

- 1x NVMe-oF target
  - 24x NVMe 2.5” SSDs
  - 2x 100GbE NICs
  - Dual x86 CPUs
- 4x initiator hosts
  - 2x25GbE NICs each
- Open Source NVMe-oF kernel drivers



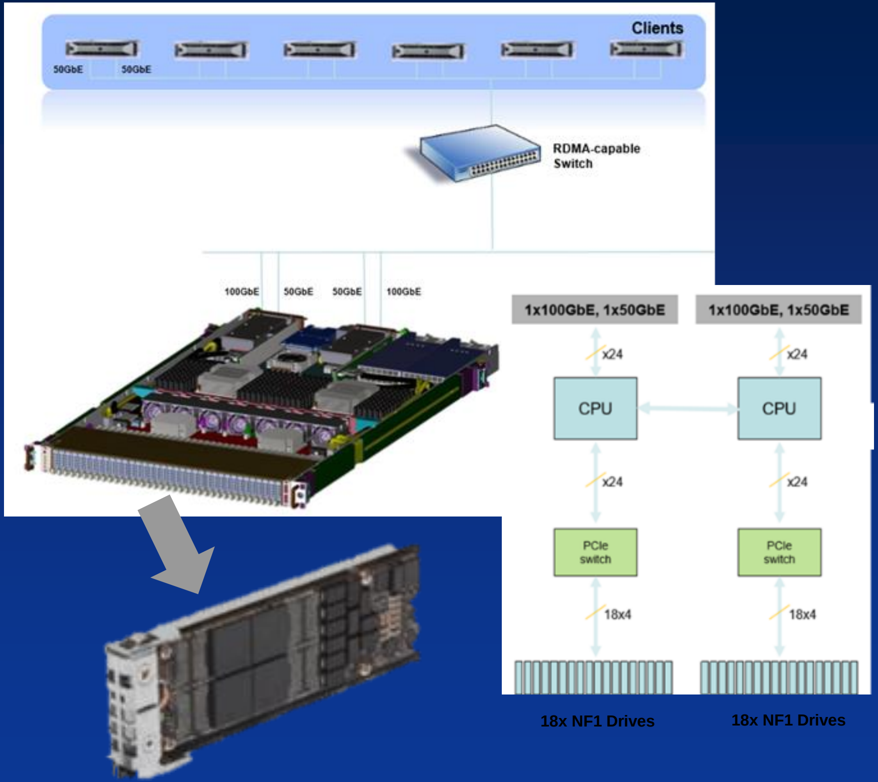
Local vs. Remote Latency Comparison – 2016



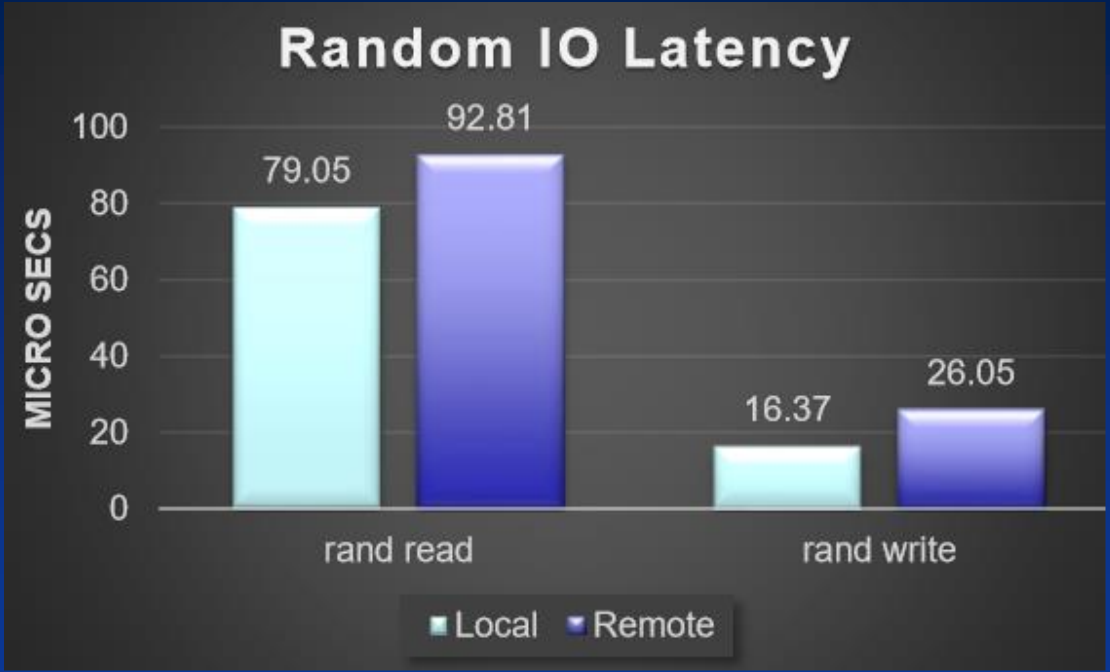
Read Gap	Write Gap
~17 us	~9 us

Performance Test Configuration – 2017

- 1x NVMeoF target
  - 36x NF1 SSDs
  - 2x 100GbE NICs, 2x 50GbE NICs
  - Dual x86 CPUs
- 6x initiator clients
  - 2x25Gb/s each
- Open Source NVMe-oF kernel drivers
- Ubuntu Linux 16.04/4.9 on Target



Local vs. Remote Latency Comparison - 2017



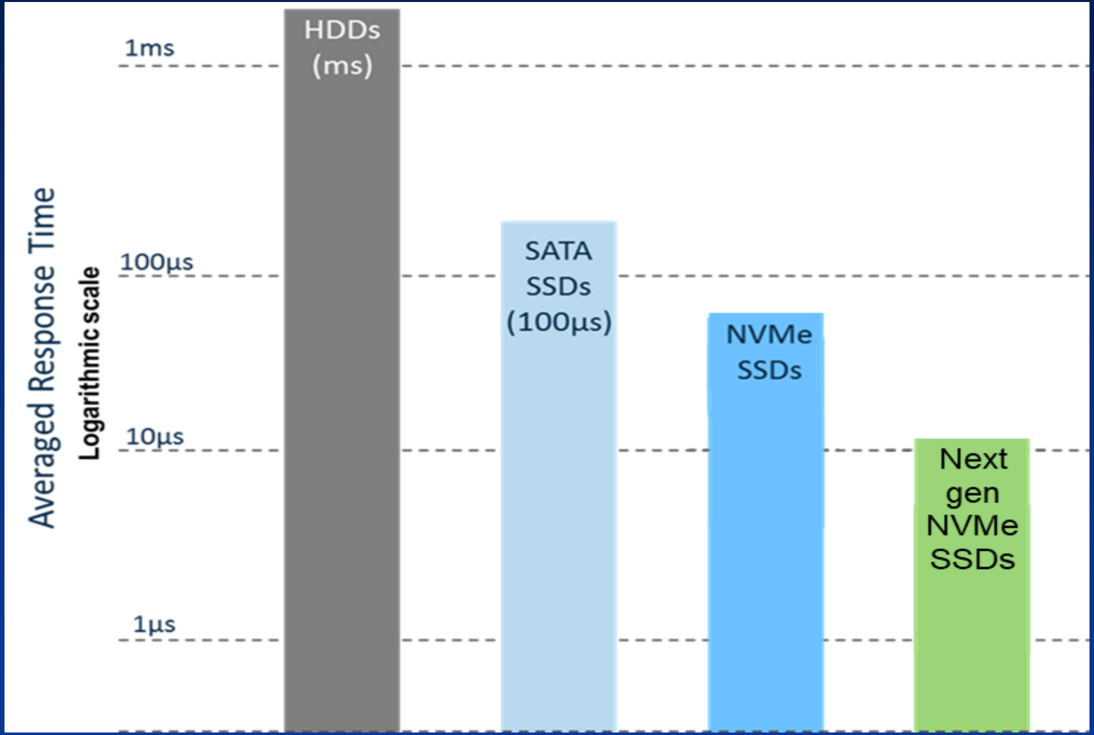
2017 Tests

Read Gap	Write Gap
~14 us	~10 us

2016 Tests

Read Gap	Write Gap
~17 us	~9 us

SSDs Will Continue to get Faster



2017 Tests

Read Gap	Write Gap
~14 us	~10 us

2016 Tests

Read Gap	Write Gap
~17 us	~9 us

Flash Memory Summit

THANK YOU

Sessions to Follow:

Forum W-32: NVMe over Fabrics (NVMe-oF) (NVMe over Fabrics (NVMe-oF) Track)

Session 204-C: Flash in Big Data Applications (Data Management Track)

Forum B-11: Flash-Memory Based Architectures: A Technical Discussion, Part 1 (Architectures Track)

Alan Weckel

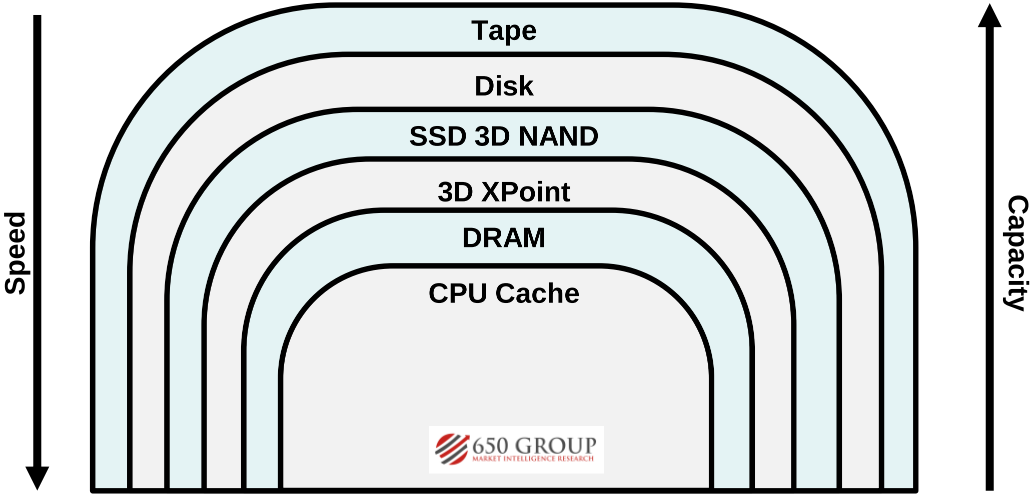
Alan Weckel is Technology Analyst/Co-Founder at 650 Group, where he is in charge of Ethernet switch, Cloud and data center research. He has written many articles for the trade and technical press, and is frequently quoted in such leading publications as Bloomberg, Businessweek, Forbes, Network World, and the Wall Street Journal. Before co-founding 650 Group, he was VP/analyst at Dell’Oro Group and had engineering and software development experience at Raytheon, General Electric Power Systems, and Cisco. He holds a BSEE and an MS in Management from Rensselaer Polytechnic Institute.

Flash Storage Networking, How the market is evolving

Alan Weckel (alan@650group.com)

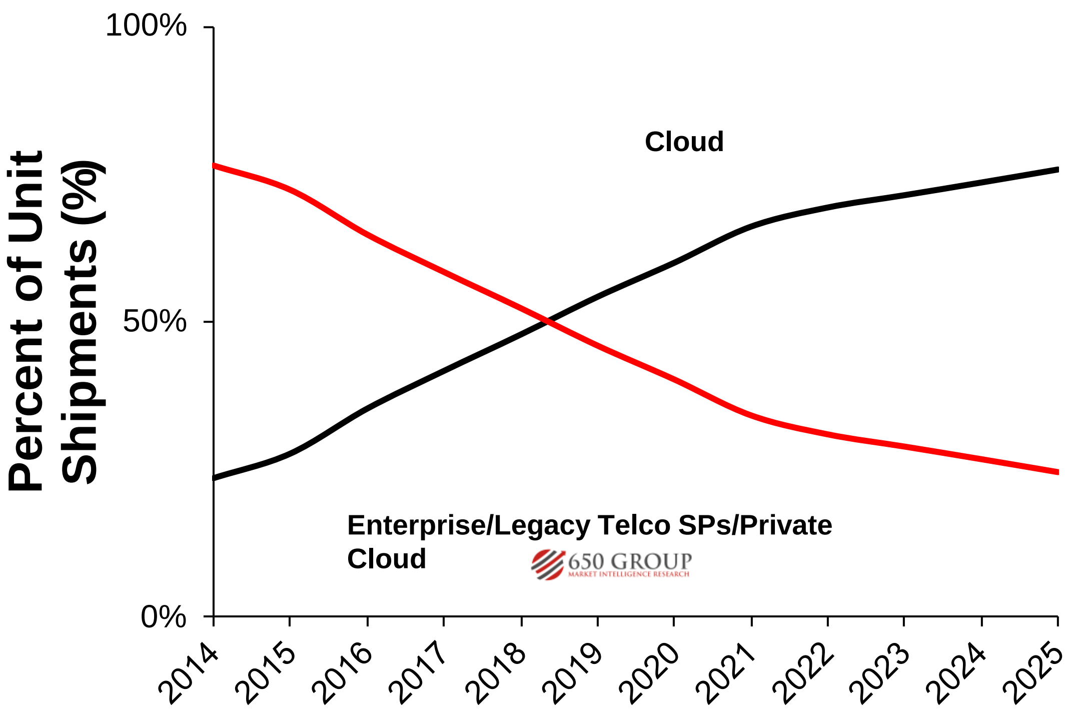
Trends changing how compute and storage are consumed

Storage: How and Where We Store Data is Changing



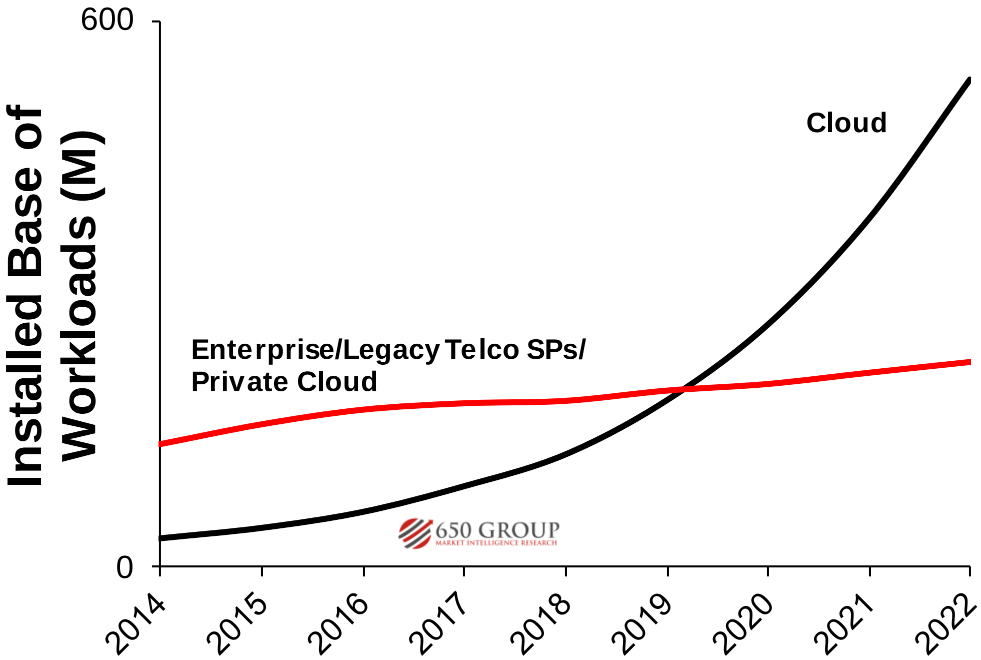
- Enterprise Storage Systems Market is Shrinking
  - Enterprises continue to buy systems
  - Enterprise market for converged and hyperconverged is growing
- Cloud Market is Growing
  - Hyperscalers buy components
  - Hyperscalers build their own software
- Areas of growth in Storage Systems Market
  - Cloud
  - All Flash Arrays
  - Hyperconverged

Server Shipments: Shipments into the Cloud



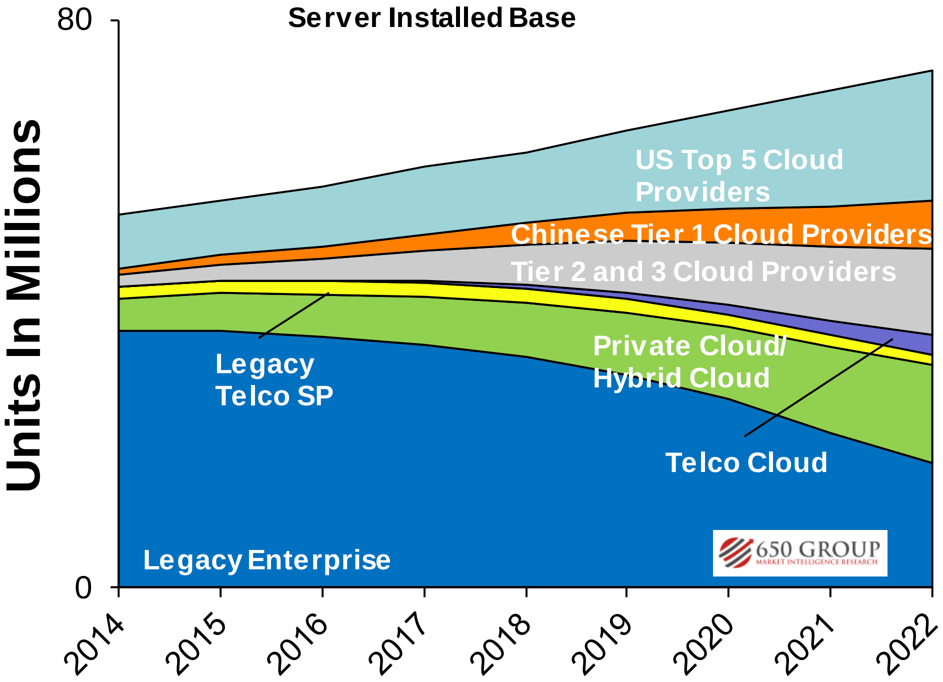
- Cloud servers will dominate compute
  - Higher-end processor
  - Smart NIC
  - Better software
  - Different type of storage
- Enterprise servers are being deployed in colocation facilities
- East/West traffic is no longer limited to one data center
  - Ethernet Based Architectures
  - Large amounts of data being moved across the world

Workloads: Installed Base by Deployment



- Enterprise workloads continue to grow
  - More workloads per server
  - Type of application is changing
  - Colocation becoming common
- Cloud workload grow exploding
  - All types of applications are growing
  - IoT will be a major driving of workload growth

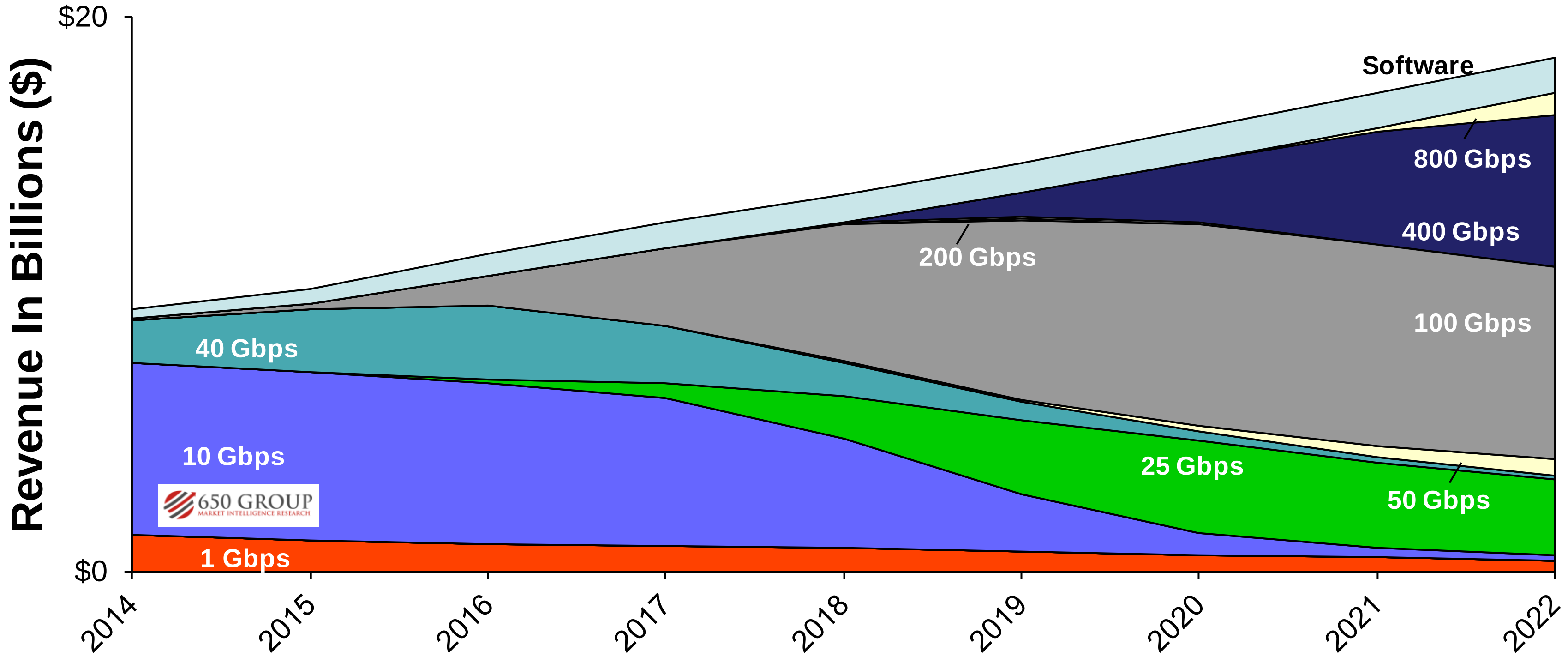
Server and Smart NICs: Server Installed Base



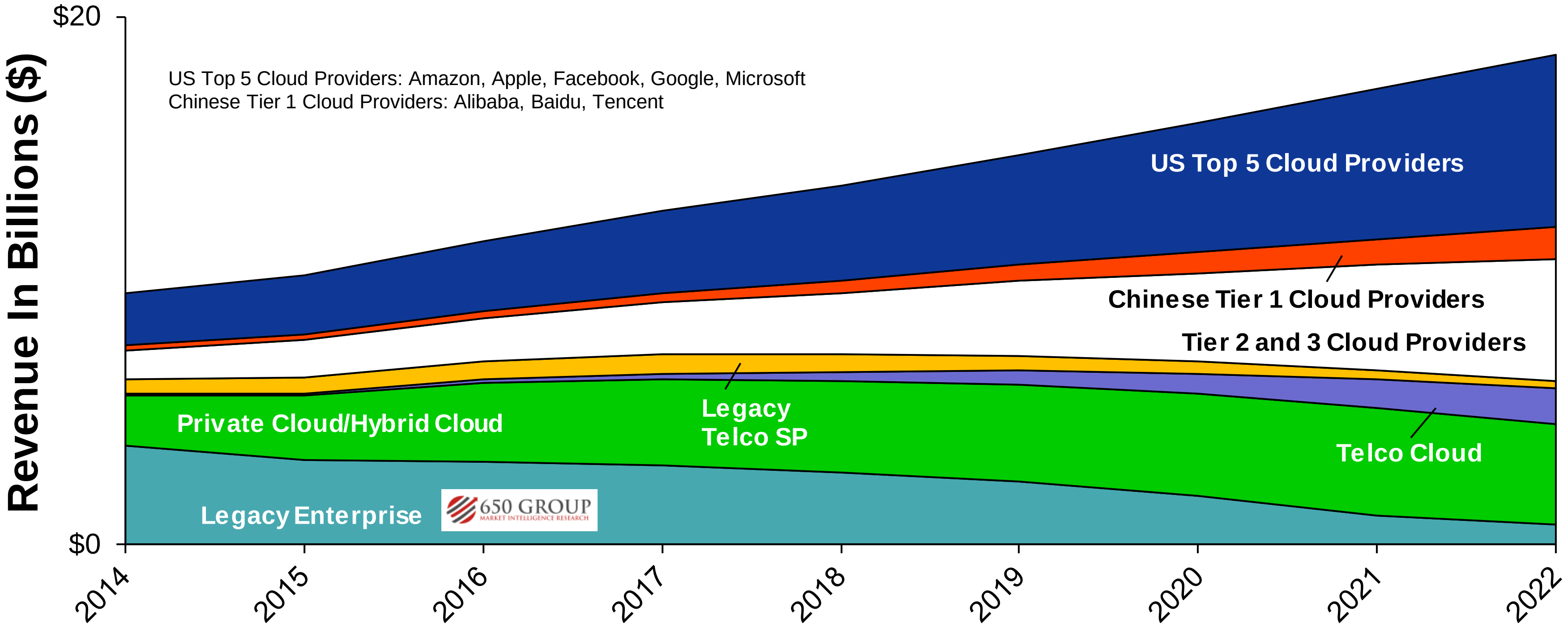
- Cloud is the new leader in technology transitions
  - Entire Telco market is smaller than Amazon
  - Cloud is moving from 2-3 to 3-4 technology generations ahead of the enterprise
- Tier 2 and 3 Clouds are increasingly riding on top of Tier 1 Cloud Infrastructure
- Clouds uses different architecture and buys different equipment then the enterprise

Ethernet Switch – Data Center

Ethernet Switch – Data Center: Total Market Revenue

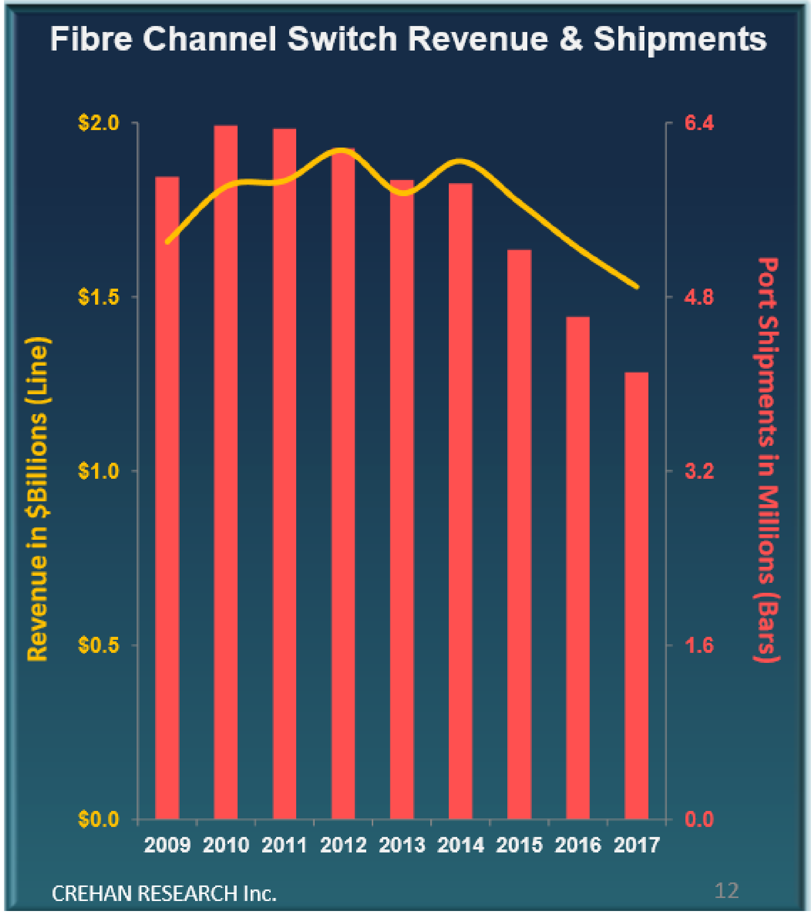


Ethernet Switch – Data Center: Total Market Revenue



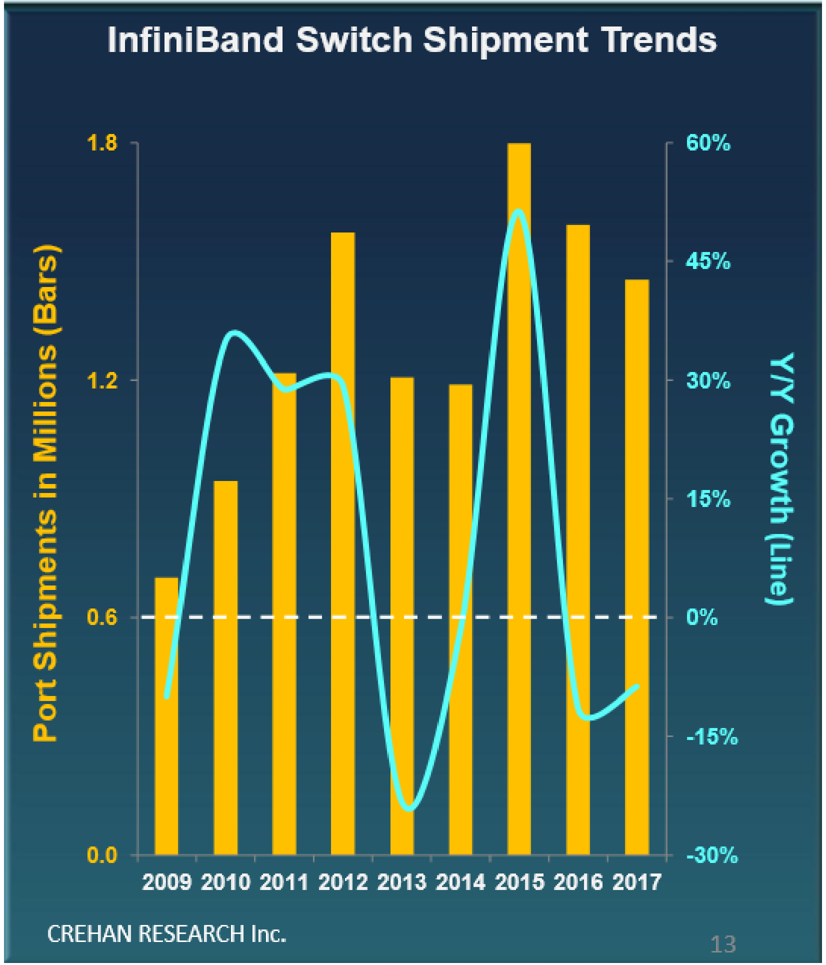
Crehan FC Data

- 2017 saw 3rd consecutive year of almost identical Y/Y change
  - Revenue down 7% to ~$1.5B
  - Shipments down 11% to ~4.1M ports
- 2H17 did improve 2% probably due to 32Gb product ramping

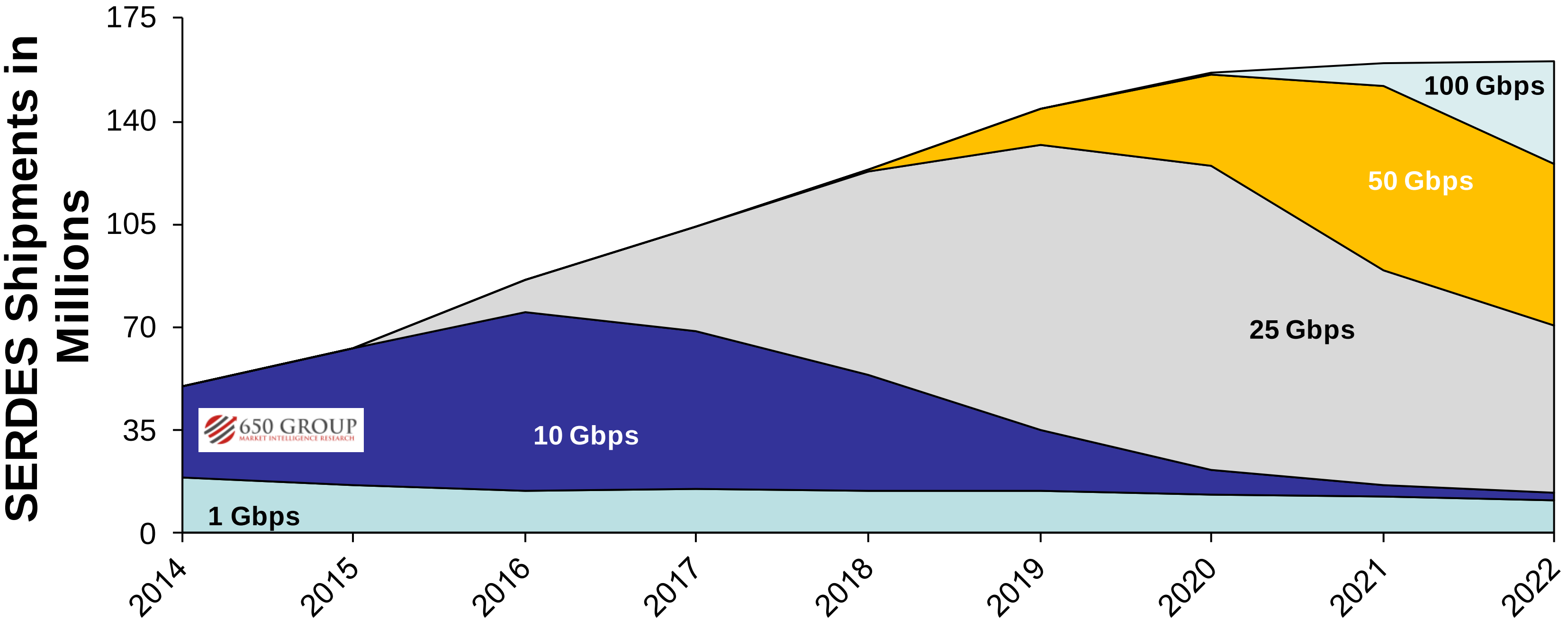


Crehan IB Data

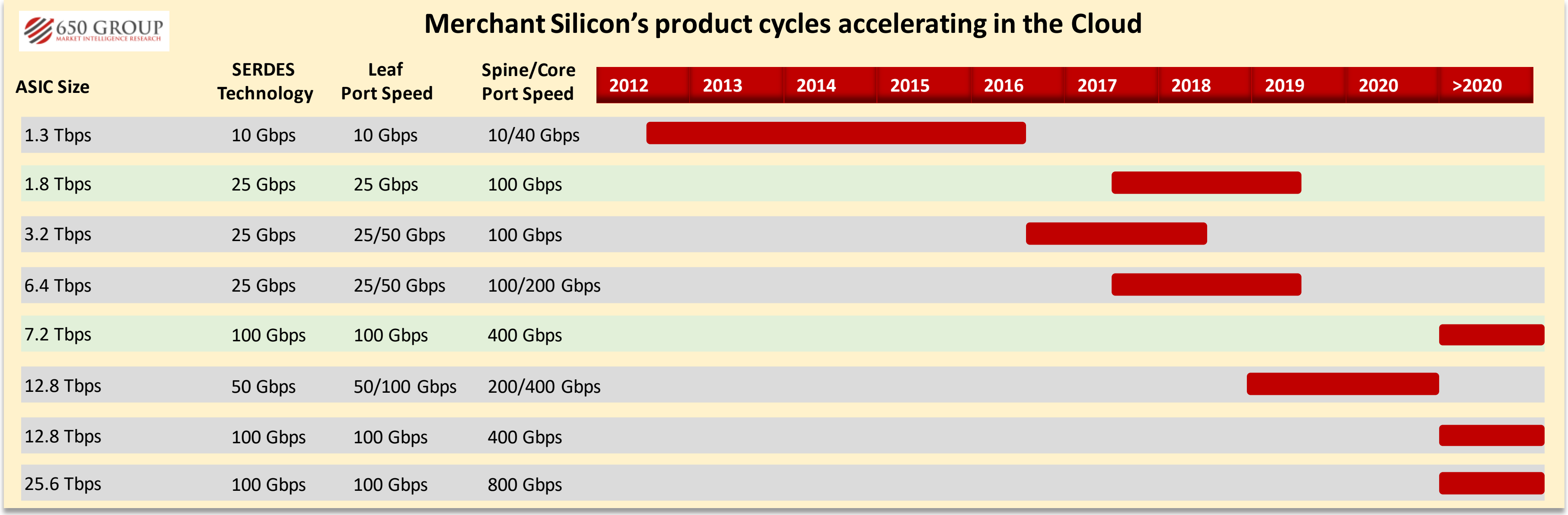
- 2017 saw revenue up and post shipments down
  - Revenue up 7% to ~$460M
  - Shipments down 9% to ~1.5M ports
- HDR/200Gb products announced but not yet shipping



Merchant Silicon – Data Center Switching: Total SERDES Shipments



Merchant Silicon – Data Center Switching: ASIC Usage in the Tier 1 Cloud



650 GROUP
MARKET INTELLIGENCE RESEARCH

Merchant Silicon's product cycles accelerating in the Cloud

ASIC Size	SERDES Technology	Leaf Port Speed	Spine/Core Port Speed	2012	2013	2014	2015	2016	2017	2018	2019	2020	>2020
1.3 Tbps	10 Gbps	10 Gbps	10/40 Gbps	🟥	🟥	🟥	🟥	🟥					
1.8 Tbps	25 Gbps	25 Gbps	100 Gbps						🟥	🟥	🟥		
3.2 Tbps	25 Gbps	25/50 Gbps	100 Gbps					🟥	🟥	🟥			
6.4 Tbps	25 Gbps	25/50 Gbps	100/200 Gbps						🟥	🟥	🟥		
7.2 Tbps	100 Gbps	100 Gbps	400 Gbps										🟥
12.8 Tbps	50 Gbps	50/100 Gbps	200/400 Gbps								🟥	🟥	
12.8 Tbps	100 Gbps	100 Gbps	400 Gbps										🟥
25.6 Tbps	100 Gbps	100 Gbps	800 Gbps										🟥

- Two waves of 400 Gbps
  - 8 X 50 Gbps
  - 4 X 100 Gbps

- Pace of Innovation Increasing
  - Four major silicon cycles in five years
  - Some technologies will get orphaned

Conclusion

- Speed of technology advancement is more rapid
- Ethernet is expanding into the Storage connectivity and Data Center transport markets at a rapid pace
- Cloud customers have different architectures and use different equipment then the enterprise
- 2019 will usher in Smart NICs and 200/400 Gbps which will expand the market for Ethernet

Thank You

Panel Q/A

Rob Davis, Ilker Cebeli, J Metz, Motti Beck, Curt Beckmann, Peter Onufryk, and Allen Weckel