FNET-301A-1:Networking Flash with Ethernet and Fibre Channel

Curt Beckmann, Principal Architect

Brocade Storage Networking, Broadcom

and

J Metz, R&D Engineer, Advanced Storage

Cisco Systems

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

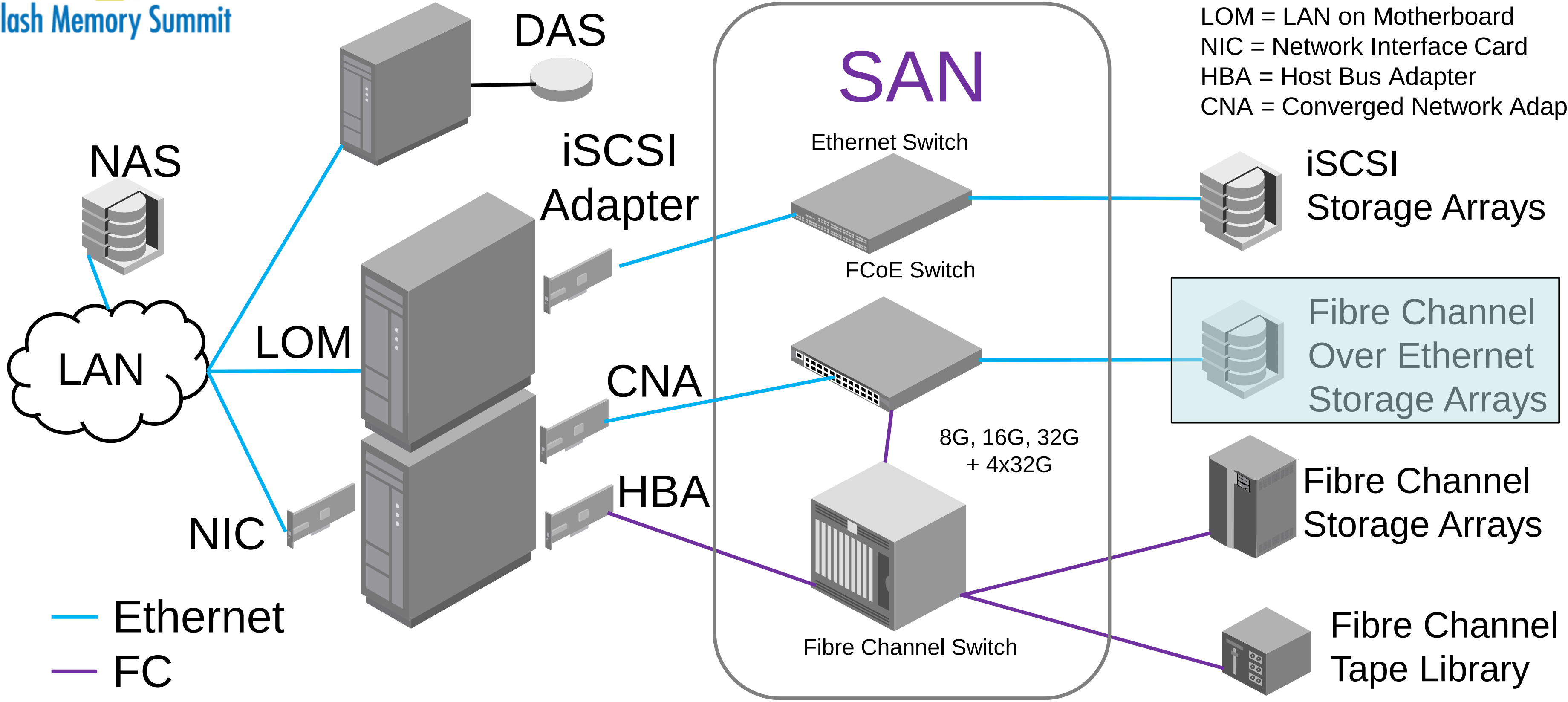
“Traditional” (20th C) shared storage concepts

- 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

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

"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

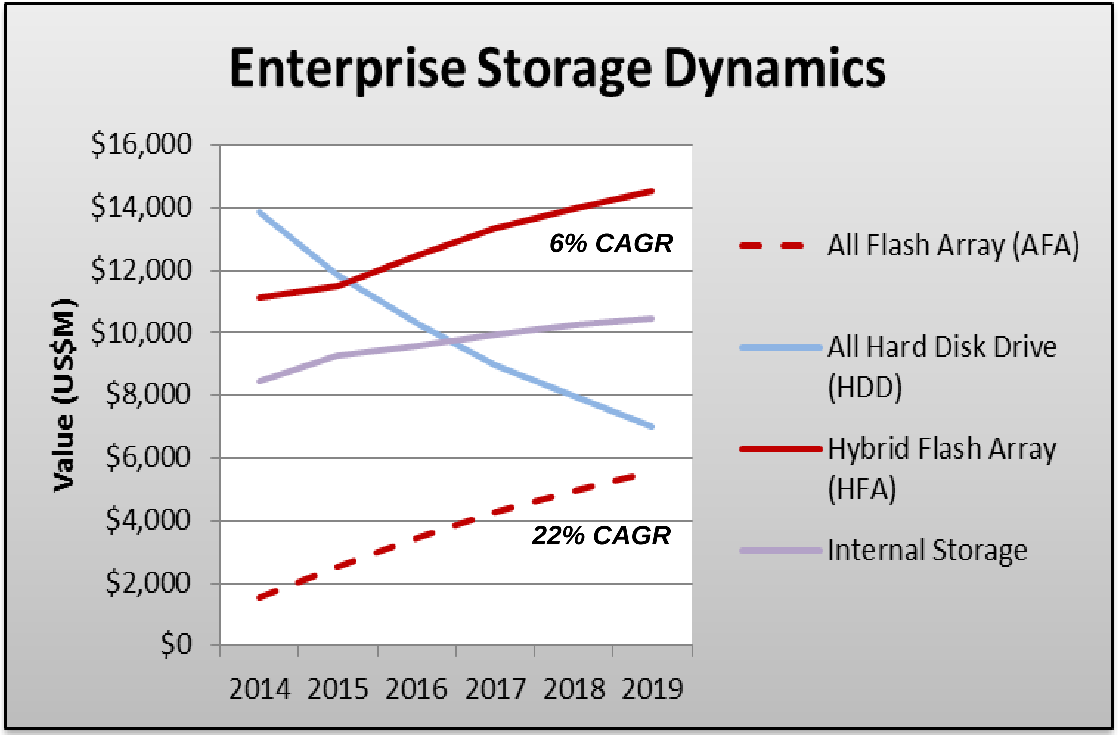
How Fibre Channel differs from Ethernet: Tech

- 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

FC-NVMe Spec Status

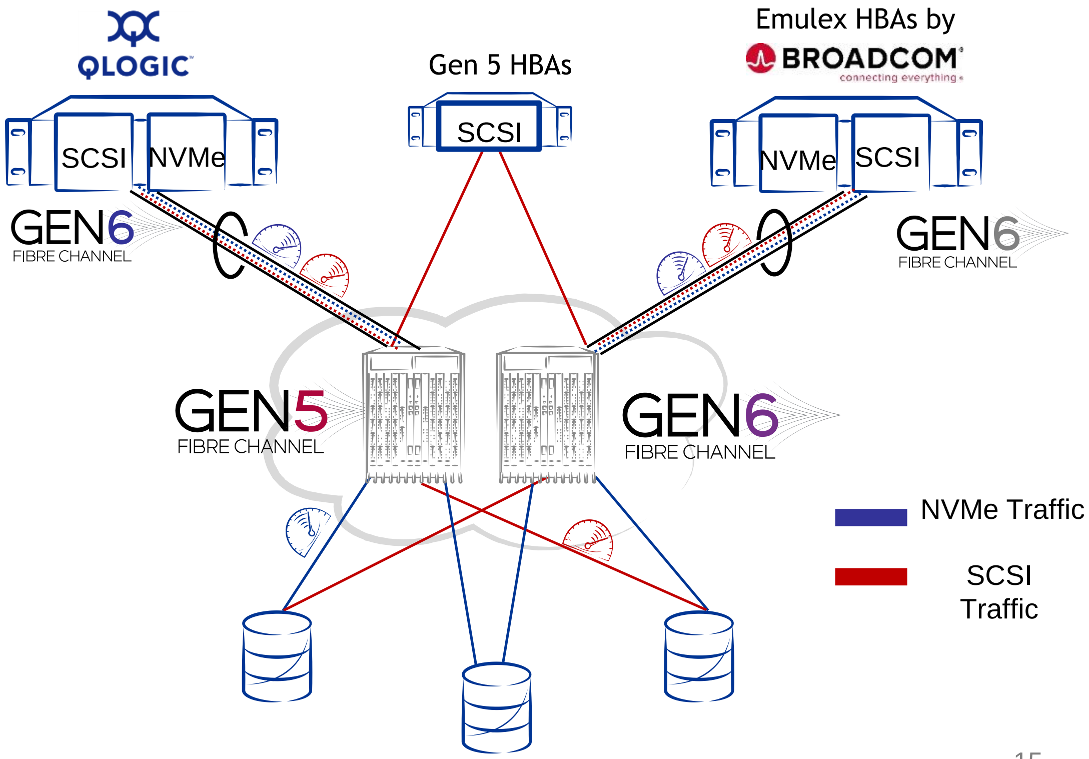
- Why move to NVMe/FC?
  - It’s like SCSI/FC tuned for SSDs and parallelism
  - Simpler, more efficient, and (as we’ll see) faster
- FC-NVMe standard effort is overseen by T11
  - T11 and INCITS finalized FC/NVMe early 2018
- Several vendors are shipping GA products
- FCIA plugfest last week: XX participants

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 management 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.

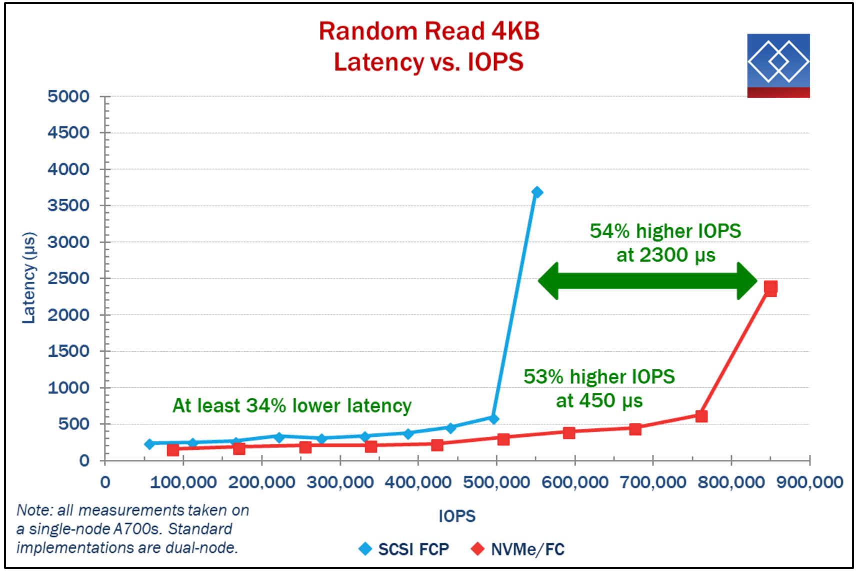
> ***NVMe/FC protects your investment:*** The benefits we observed were with existing hardware that supports 32GFC.

> ***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



Random Read 4KB
Latency vs. IOPS

54% higher IOPS at 2300 µs

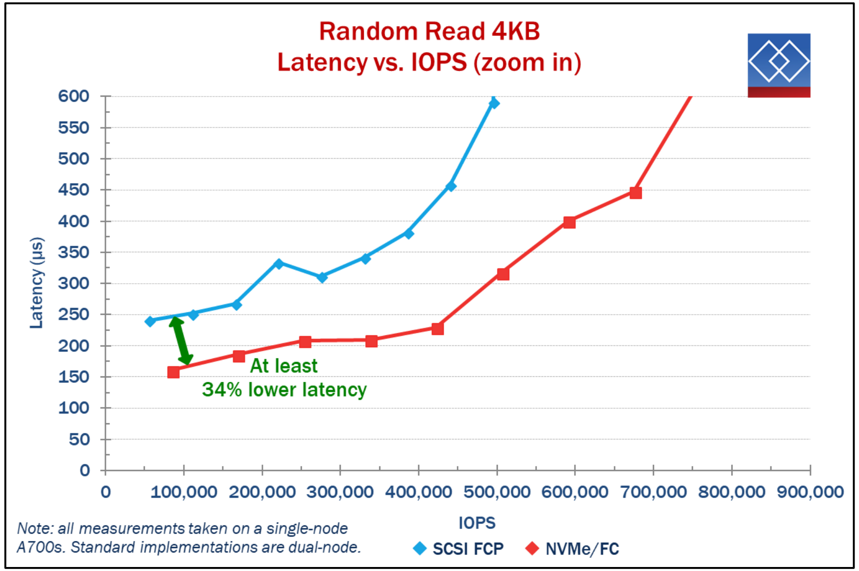
53% higher IOPS at 450 µs

At least 34% lower latency

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



Random Read 4KB
Latency vs. IOPS (zoom in)

At least 34% lower 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

Ethernet-Networked Flash Storage

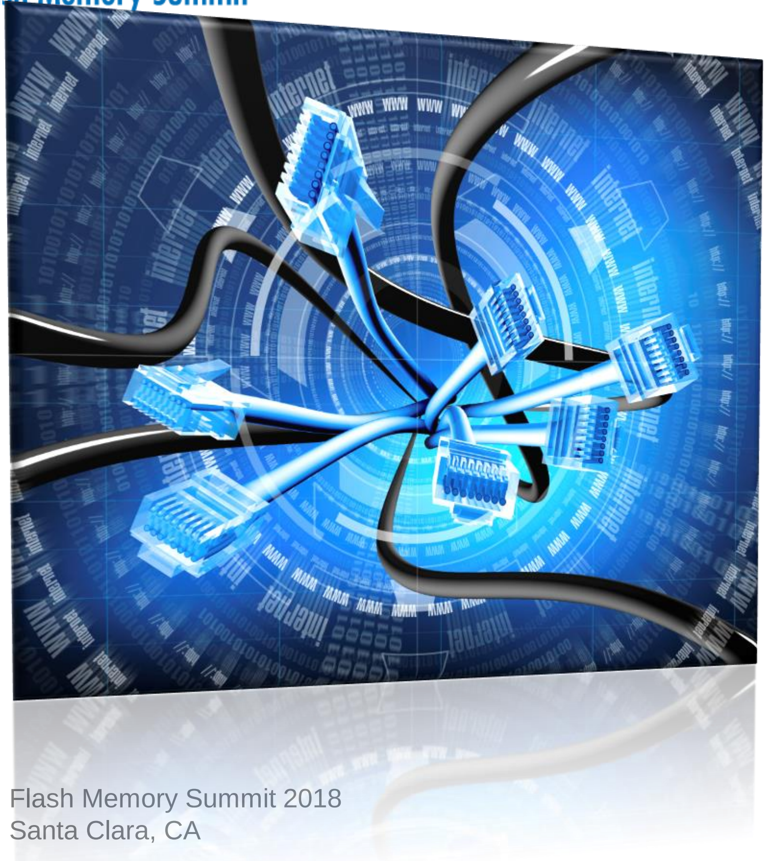
J Metz, Ph.D

R&D Engineer, Advanced Storage

Cisco Systems

@drjmetz

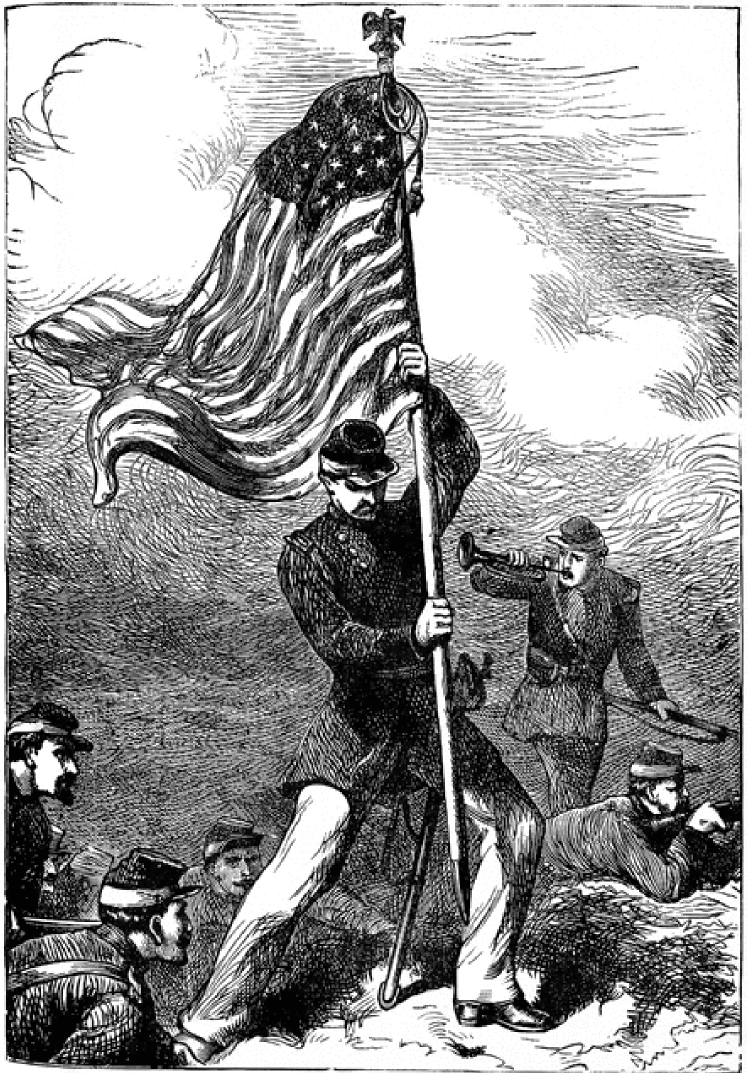
Agenda



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

Planting a Flag

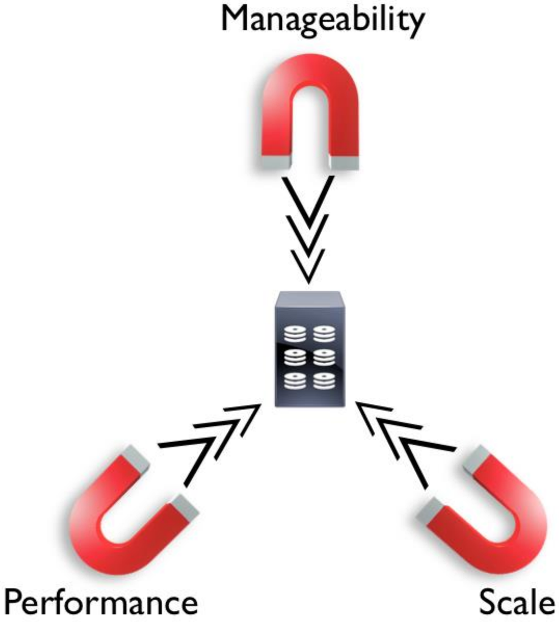
- Is there anyone who thinks Ethernet will *not* play a role in storage?



Then the Question Is…

…how best to use Ethernet for Storage?

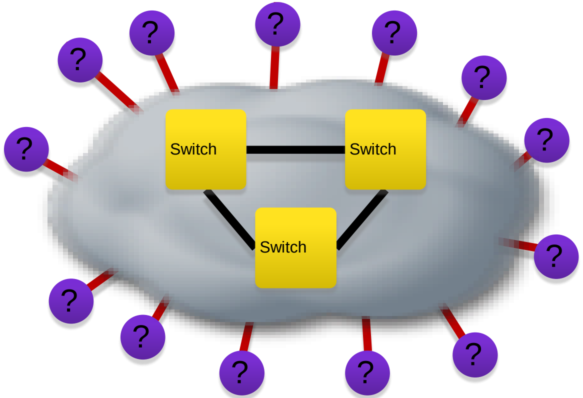
Storage Perspective



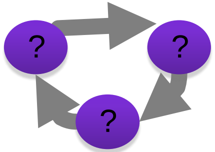
- There is a “sweet spot” for storage
  - Depends on the workload and application type
  - No “one-size fits all”
- What is the problem to be solved?
  - Deterministic or non-deterministic?
  - Highly scalable or highly performant?
  - Level of manageability?
- Understanding “where” the solution fits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

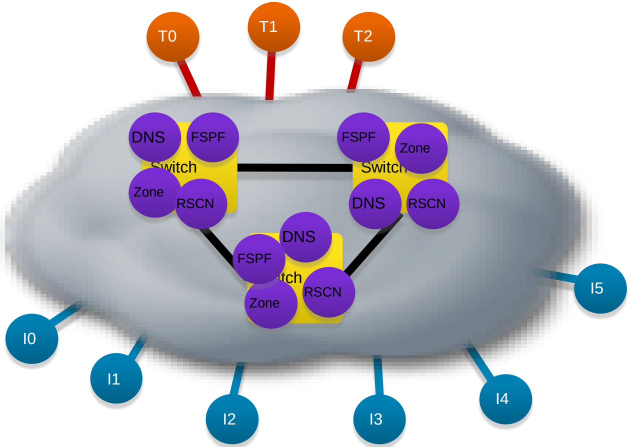


Fabric topology and traffic flows are highly flexible

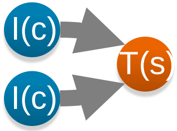


Client/Server Relationships are not pre-defined

Network Determinism (cont.)



Fabric topology, services and traffic flows are structured

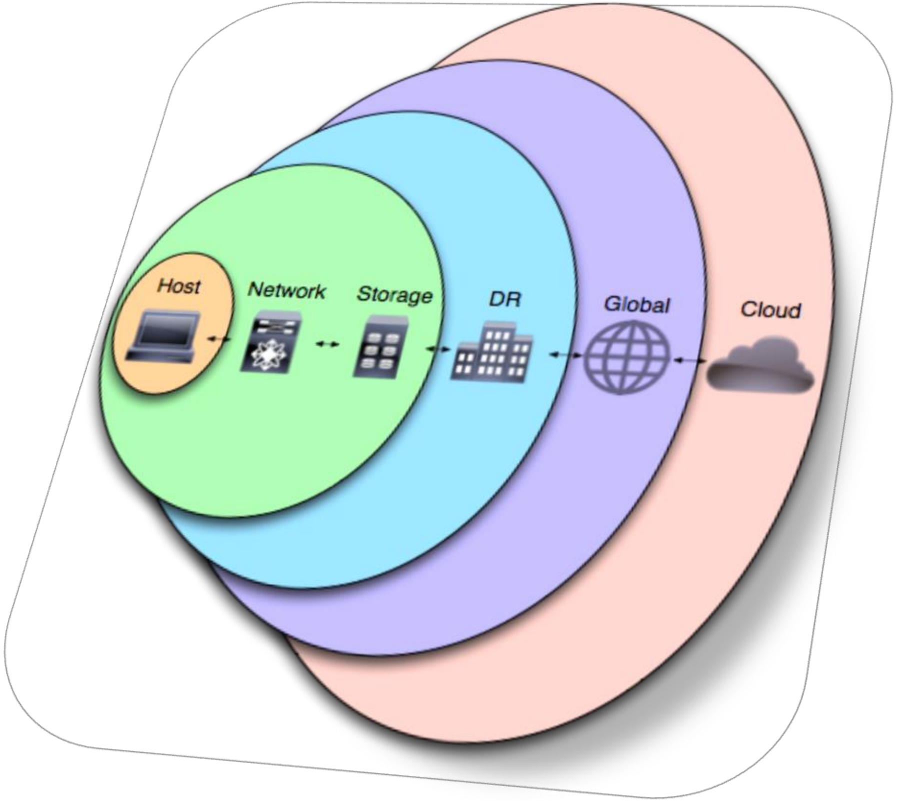


Client/Server Relationships are pre-defined

- 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

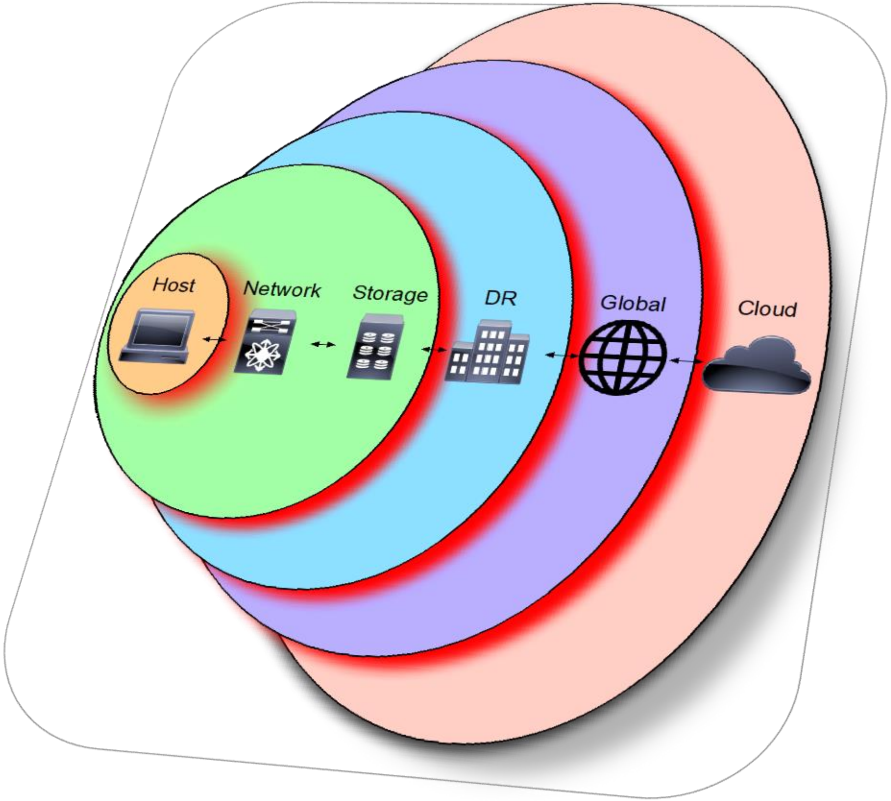
Big Picture

- 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?

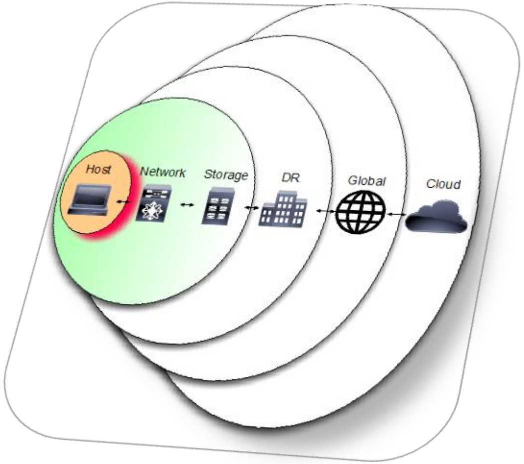


Big Picture

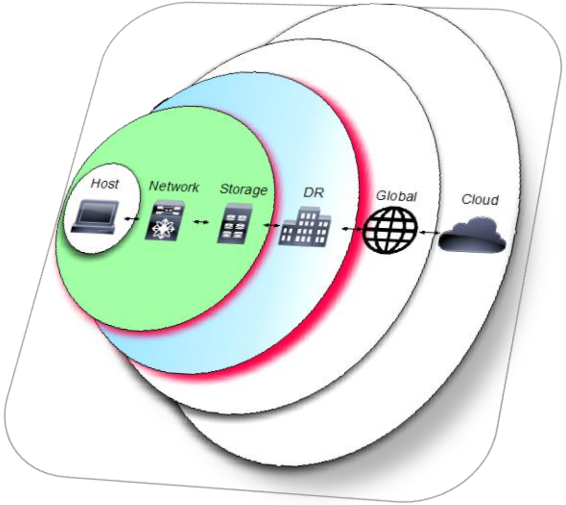
- When you miss the sweet spot, you risk major problems
  - Careful of the “Danger Zones”



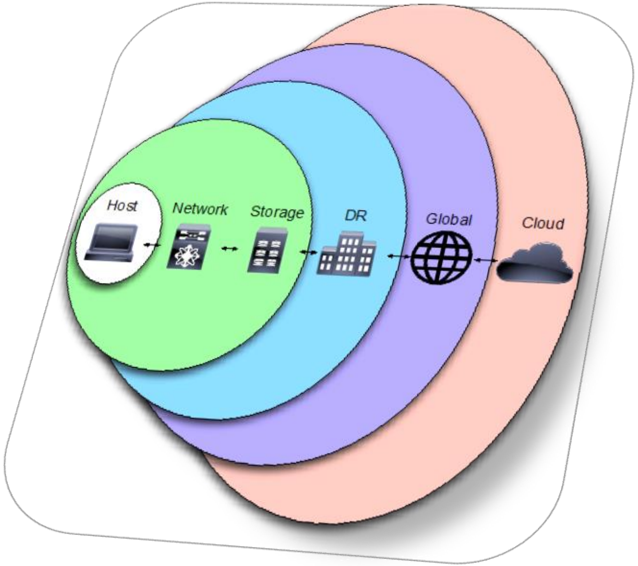
Scope Comparison



PCIe

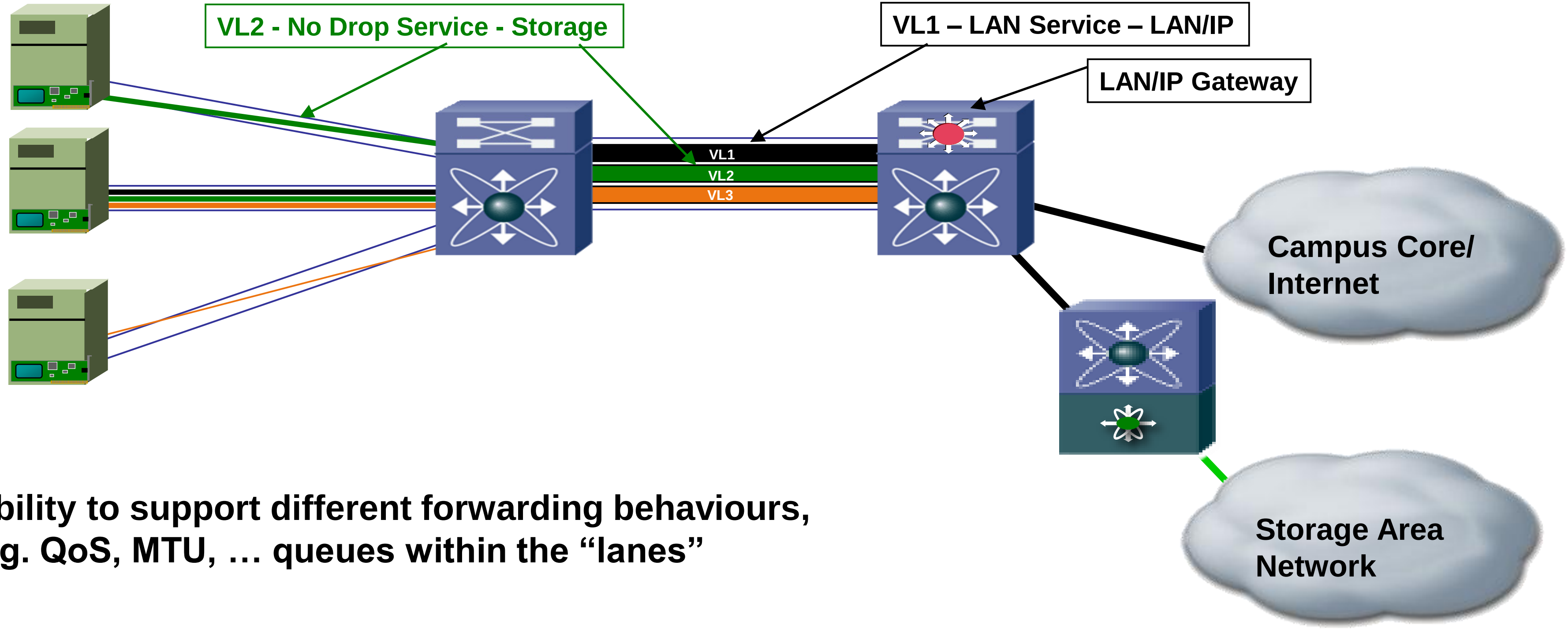


Fibre Channel
Ethernet (FCoE, iSCSI,iSER, NVMe-oF)
InfiniBand



Ethernet (NFS, SMB, Object)

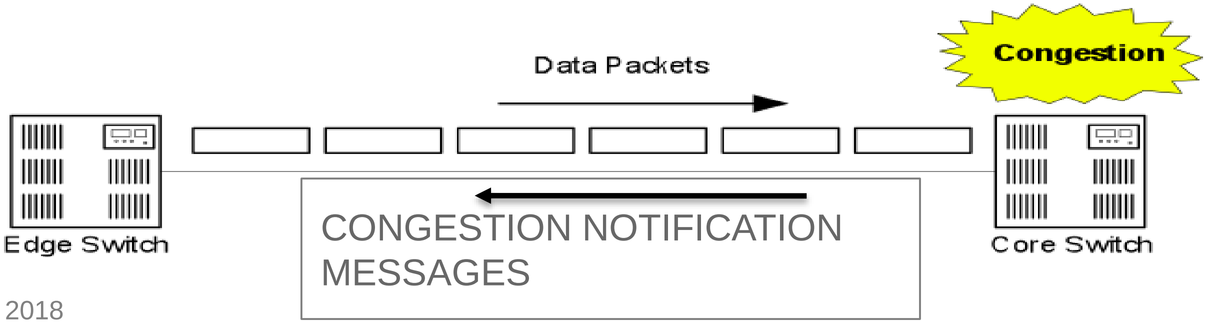
Ethernet Enhancements



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

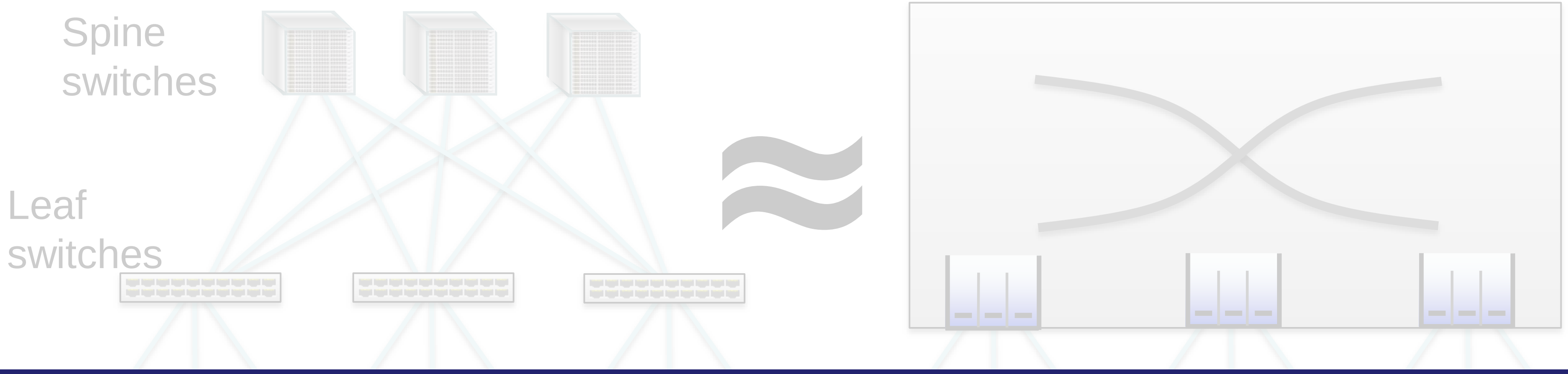
- Congestion indicated quantitatively (reduce load prior to packet loss)
- React in proportion to the extent of congestion, not its presence
  - Reduces variance in sending rates, lowering queuing requirements

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

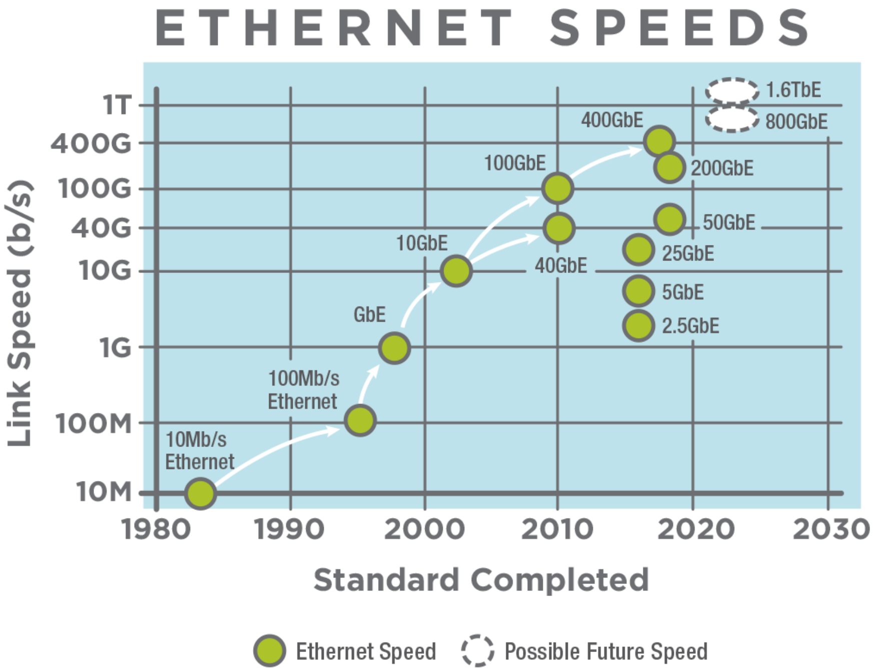
Approximates ideal output-queued switch



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

Flash Memory Summit

Ethernet Roadmap



- How to go faster
  - Different modulation techniques
  - Different data rate/lanes chosen
- New Signaling methods
  - Pulse Amplitude Modulation 4 vs. Non Return to Zero (NRZ)
- New Form Factors
  - Multi-lane interfaces

Source: NEXT ETHERNET ERA ethernet alliance

ethernetalliance.org

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