iSCSI : A loss-less Ethernet fabric with DCB

Jason Blosil, NetApp
Gary Gumanow, Dell

SNIA Legal Notice

Abstract

◆ A look at different storage technologies and a brief look at each; benefits/tradeoffs, making the case for a SAN. Then looking at making your SAN a lossless, high-performance, predictable, resource for your business. Well look at different storage protocols and how they compare to the OSI model, and the new DCB protocols. And lastly, I'll present some findings of using these technologies.

◆ Learning Objectives

   ◆ A detailed look and comparison between three storage models; iSCSI, FC, and FCoE. Comparing all three to the OSI model. A discussion of the requirements for each layer of the networking stack.

   ◆ A look at the benefits that DCB can provide iSCSI, as a lossless Ethernet fabric. A review of performance and contention issues.

What we'll talk about today…

- ◆ How is iSCSI doing today
- ◆ What is Data Center Bridging
- ◆ What's the roadmap for Ethernet
- ◆ A protocol comparison
- ◆ iSCSI over DCB performance
- ◆ Does DCB really matter for iSCSI

IDC Forecast (Dec 2009)

Revenue ($M)

Total CAGR: 5%

iSCSI CAGR: 22%

■ FC SAN ■ iSCSI SAN ■ NAS ■ DAS

Revenue Share

— FC SAN — iSCSI SAN — NAS — DAS

iSCSI today
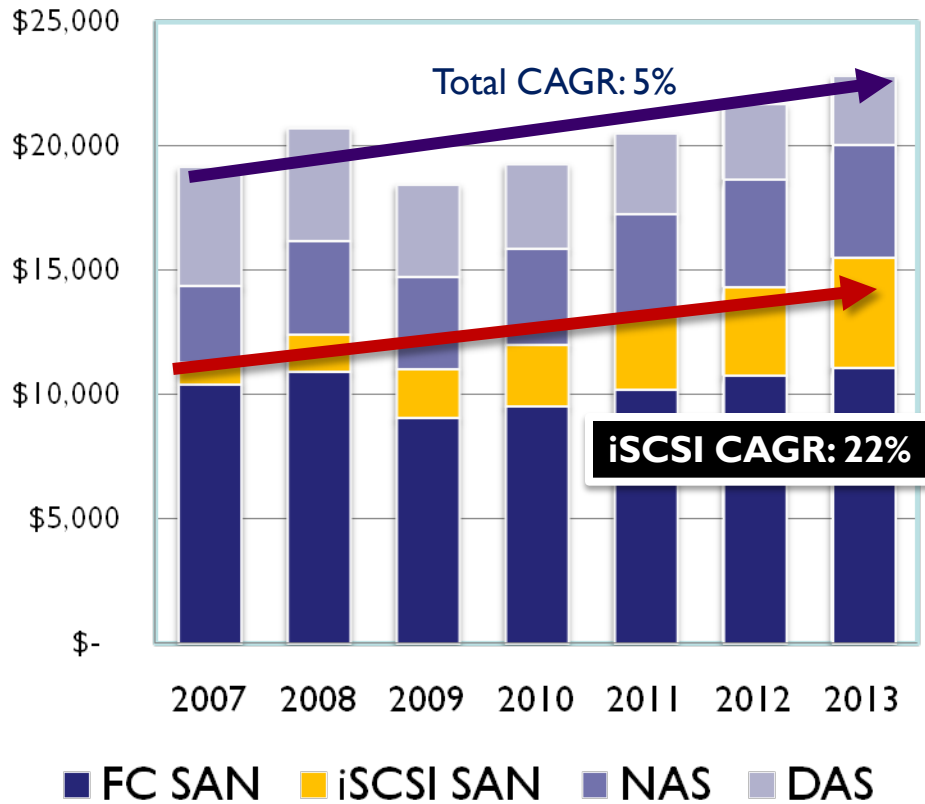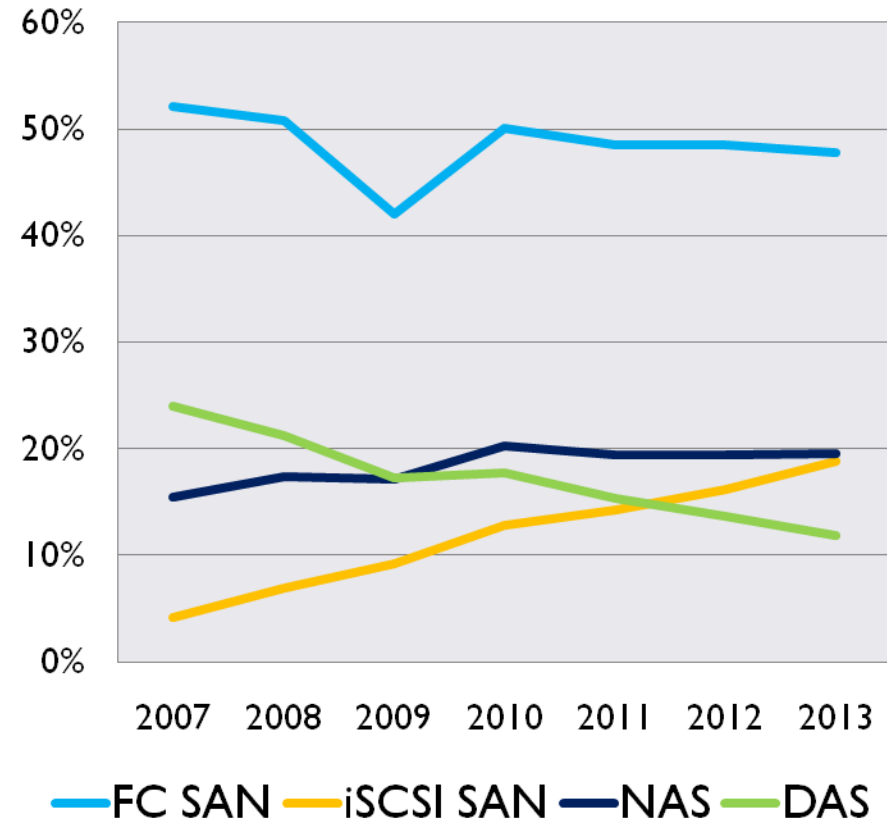
- ❯ **Rich set of tools**
  - ◆ Disaster Recovery / Backup
  - ◆ Data Virtualization
  - ◆ Thin provisioning
  - ◆ Snapshots
  - ◆ Replication
  - ◆ VM integration
  - ◆ Disk de-dupe
- ❯ **SW Initiators available from all major OSes**
  - ◆ MPIO support
  - ◆ Low CPU utilization without HW offload
- ❯ **Architecture choice**
  - ◆ Scale-out & Scale-up
- ❯ **10GbE available today**

Sum of Terabytes	CAGR 2006-2009
iSCSI	43.6%
Storage Growth	29.0%

Source: IDC Storage Tracker Q4 2009

Sum of Terabytes
IDC Storage Tracker Q4 '09

iSCSI: A Lossless Ethernet Fabric with DCB
© 2010 Storage Networking Industry Association. All Rights Reserved.

Trends in the datacenter
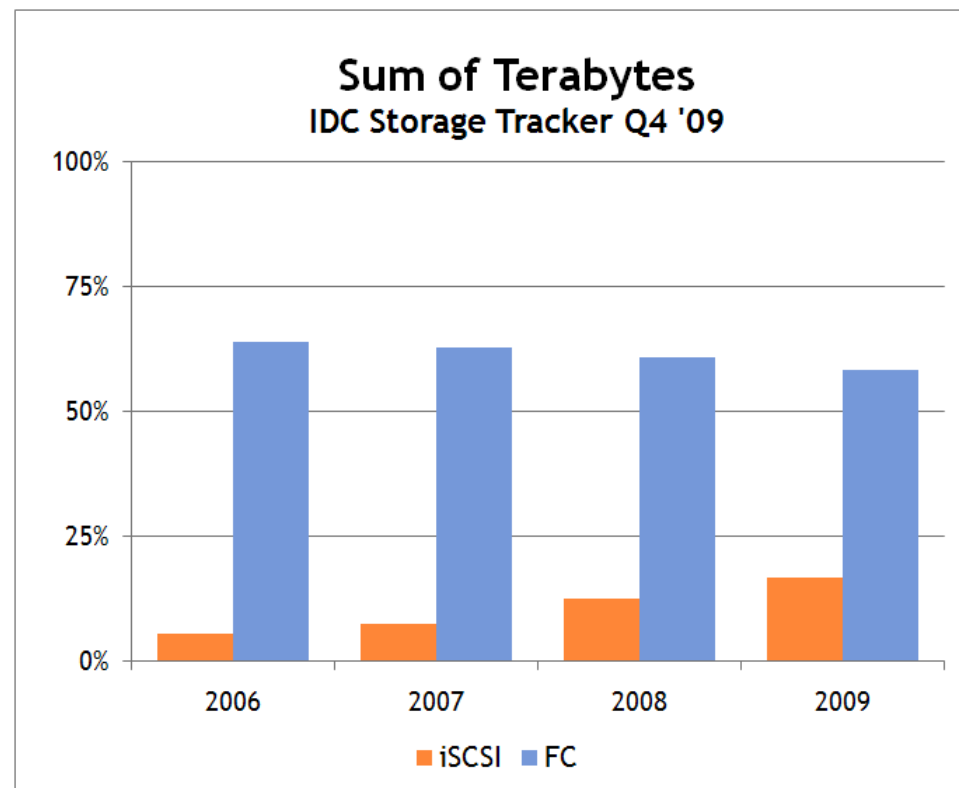
- ## Networking
  - ◆ Data Center Bridging (DCB)
  - ◆ 10 Gigabit Ethernet
  - ◆ Top of Rack switching
- ## Storage
  - ◆ Increased focus on Ethernet-based storage
- ## Server
  - ◆ PCIe allows server architectures to support > 50Gbps
  - ◆ Increased density with virtualization

Virtualization Drives Bandwidth

From

To

App App App App App . . .

OS OS OS OS OS . . .

VM VM VM VM VM . . .

Hypervisor

App

OS

Server

2-4Gb
Sufficient

Ethernet Switch

Server

10-20 Gb
Needed

Ethernet Switch
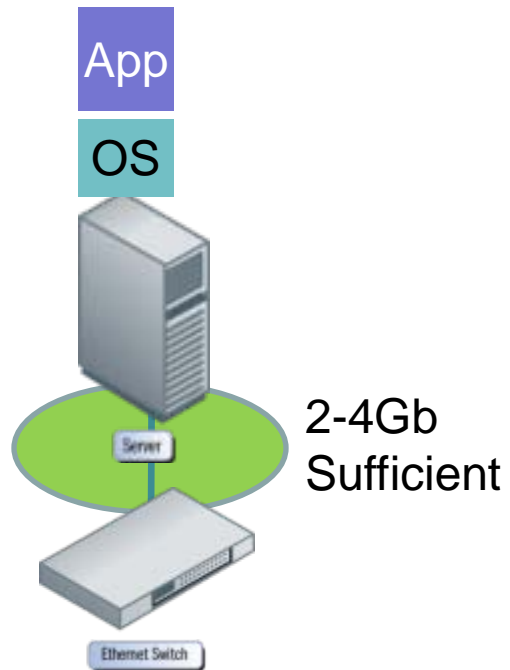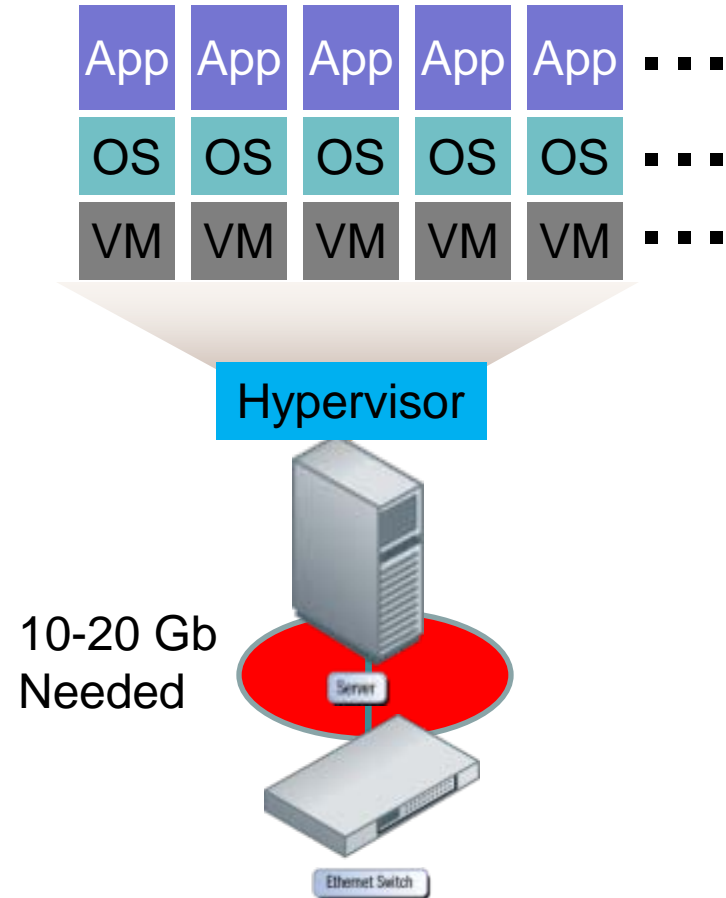
Ethernet Flexibility – Shared Network

App	App	App	App	App
OS	OS	OS	OS	OS
VM	VM	VM	VM	VM

Hypervisor

Server

Ethernet Switch

Ethernet

… Voice, LAN, etc.

NFS   CIFS   iSCSI

The need for 10GbE
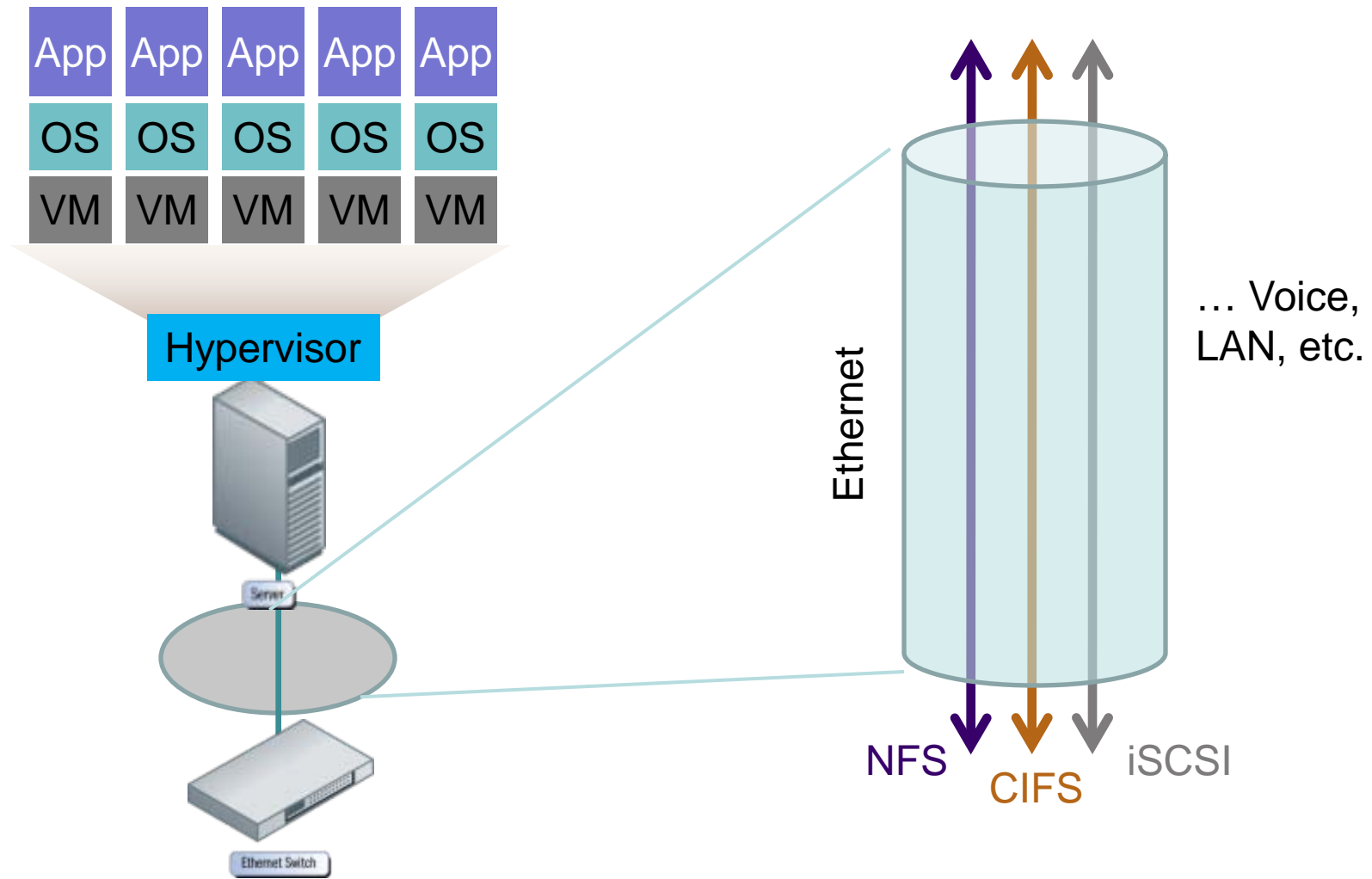
Virtualization

- VMs requires more network capacity
- 80% CPU increases demand on bandwidth
- 10GbE & iSCSI offers the bandwidth needed today
- 10GbE w/ DCB enables enhanced fabric virtualization by offering enhanced manageability and QoS

Consolidation

- Enabling server virtualization allows for increasing levels of server consolidation
- Unifying LAN & SAN consolidates network fabric
- Enables reductions in HW, power and OpEX

Application Performance

- Instant access to data anytime anywhere
- Larger database, streaming media and other high I/O applications demand data center performance
- Enables thin client and desktop virtualization

Fabric Roadmaps
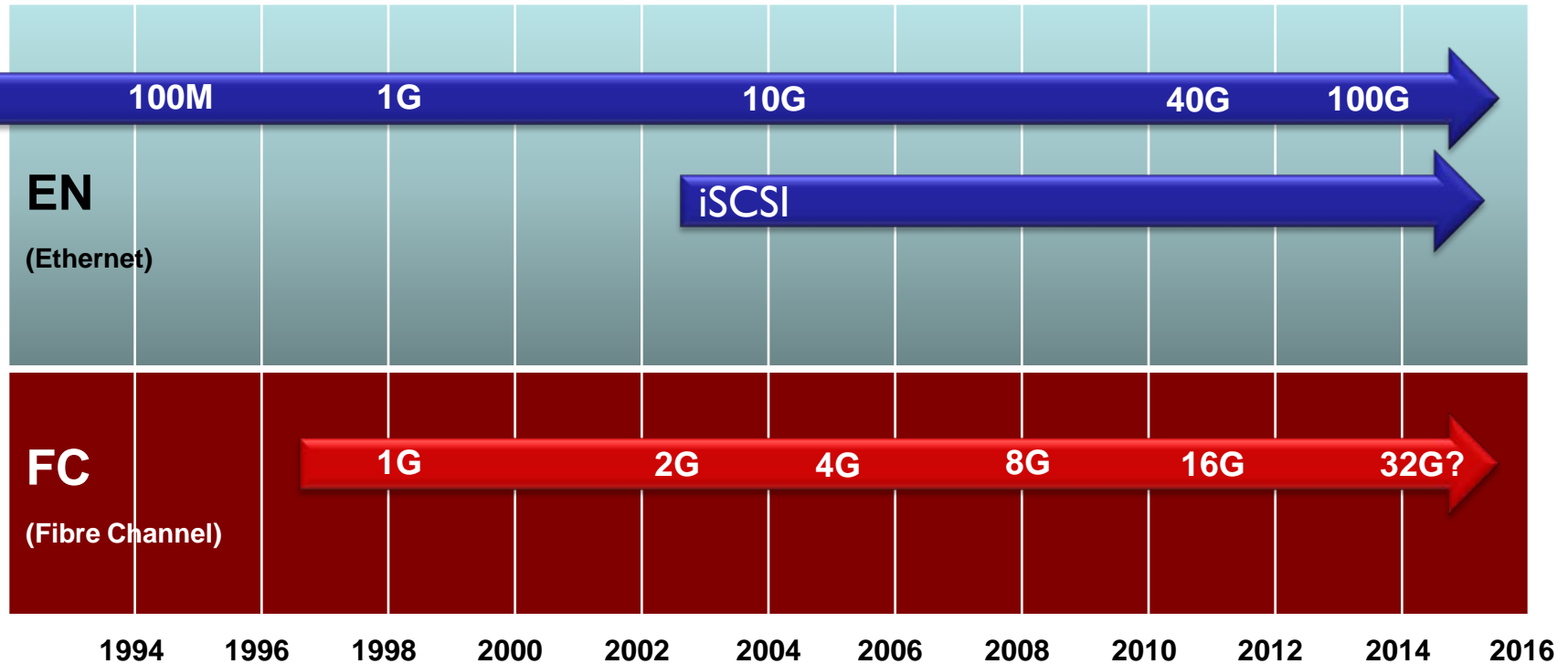
	1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014	2016
EN (Ethernet)	100M		1G			10G				40G	100G	
						iSCSI						
FC (Fibre Channel)			1G		2G	4G		8G		16G		32G?

◆ Infrastructure migrates to Ethernet irrespective of storage protocols

◆ Convergence and virtualization are driving enhancements to Ethernet

iSCSI: A Lossless Ethernet Fabric with DCB

Fabric Roadmaps

DCB

EN
(Ethernet)

100M 1G 10G △ 40G 100G

iSCSI

FCoE

FC
(Fibre Channel)

1G 2G 4G 8G 16G 32G?

1994	1996	1998	2000	2002	2004	2006	2008	2010	2012	2014	2016
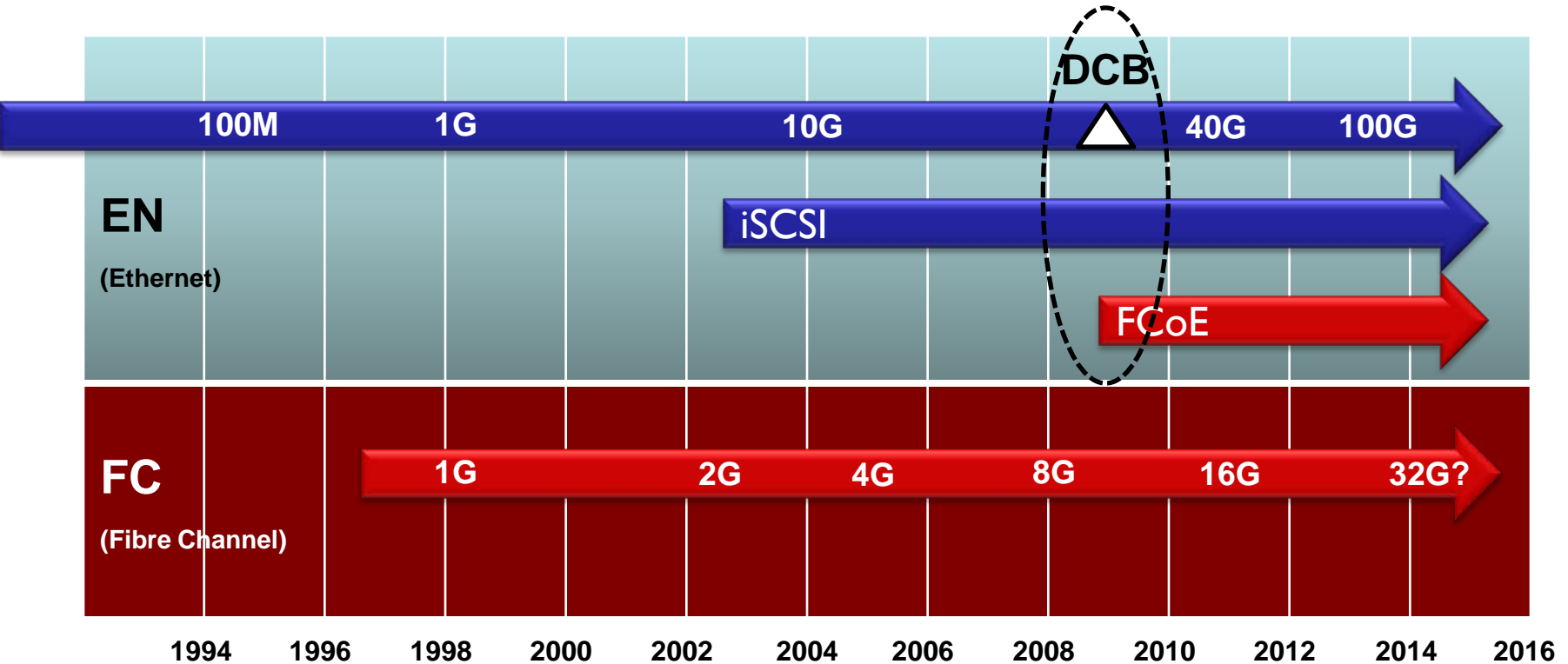
◆ Infrastructure migrates to Ethernet irrespective of storage protocols

◆ Convergence and virtualization are driving enhancements to Ethernet
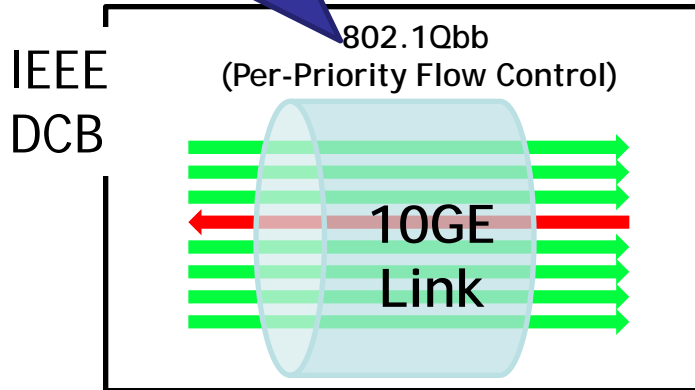
Data Center Bridging

- ### Enhancements to Ethernet
  - Provides enhanced QoS support to Ethernet
  - Not just storage, but all applications

- ### What constitutes DCB Standards?
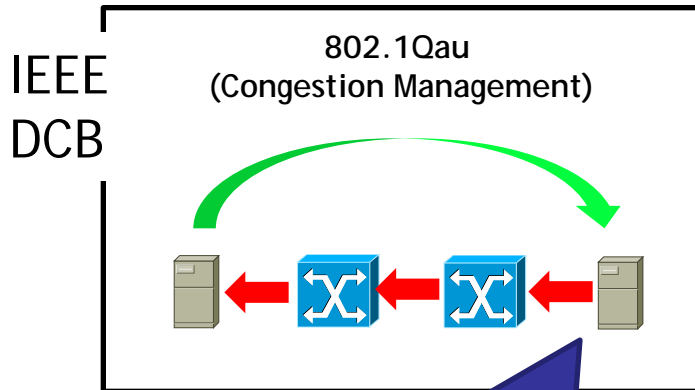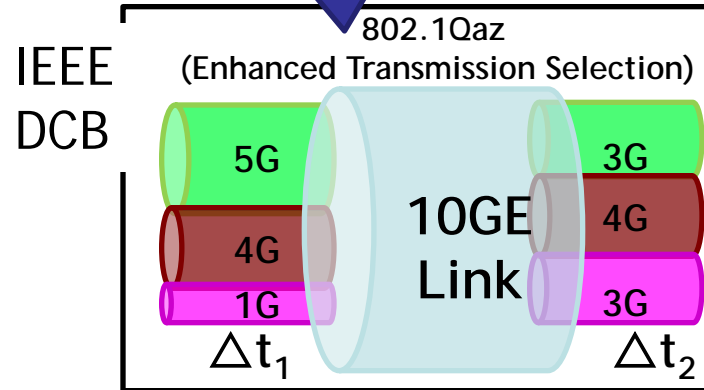 - **PFC** aka Priority based Flow Control (802.1Qbb)
  - **ETS** aka Enhanced Transmission Selection (802.1Qaz)
  - **CN** aka Congestion Notification (802.1Qau)
  - **DCBX** aka Data Center Bridging capability eXchange
    - › LLDP vs. DCBX
      - LLDP: Primarily a link level information exchange protocol
      - DCBX: Neighbors can configure parameters based on info exchange and state machine

DCB Components

HALT an individual lane, but NOT all of them!

Allocate bandwidth based upon predetermined classes of traffic

IEEE DCB

802.1Qbb
(Per-Priority Flow Control)

10GE Link

IEEE DCB

802.1Qaz
(Enhanced Transmission Selection)

5G

4G

1G

3G

4G

3G

10GE Link

$\triangle t_1$

$\triangle t_2$

IEEE DCB

802.1Qau
(Congestion Management)

IEEE DCB

Data Center Bridging Exchange

Upper Layers

Upper layer driver

Data Center Bridging Parameter Exchange

Auto-negotiation

Driver initialization

MAC

Local Node

4 Declare link UP

3 DCB Parameter Exchange

2 Speed negotiation

1 Ethernet link

Upper Layers

Upper layer driver

Data Center Bridging Parameter Exchange

Auto-negotiation

Driver initialization

MAC

Remote Node

End-to-End Communication between end-points. Tells the end-point to BACK OFF!
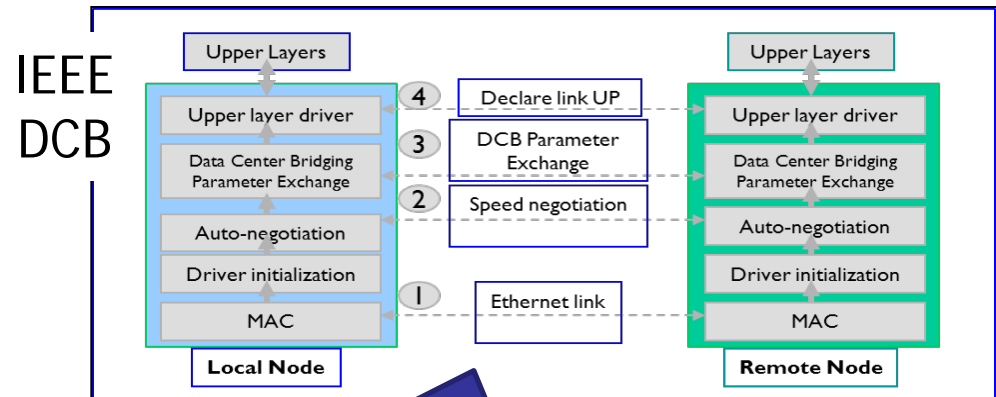
End-nodes exchange DCB capabilities

s Ethernet Fabric with DCB
Networking Industry Associat
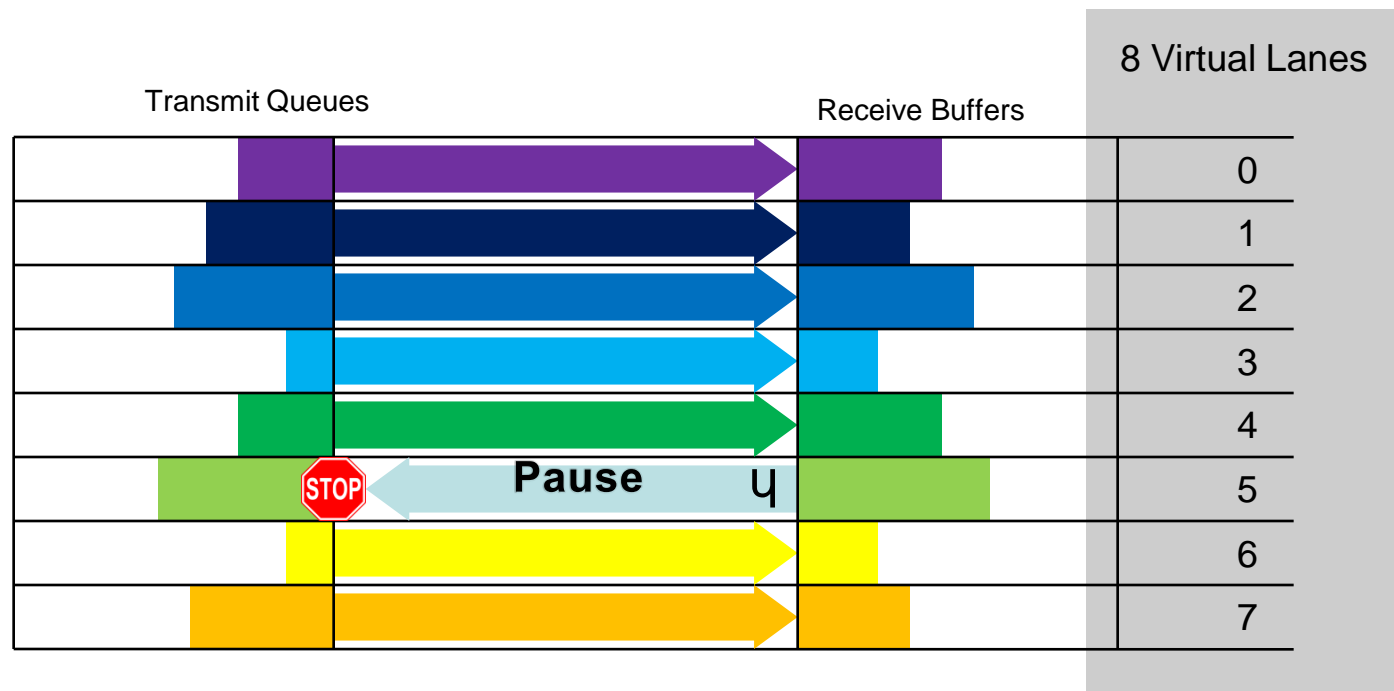
14

Priority-based Flow Control
IEEE 802.1Qbb – PFC

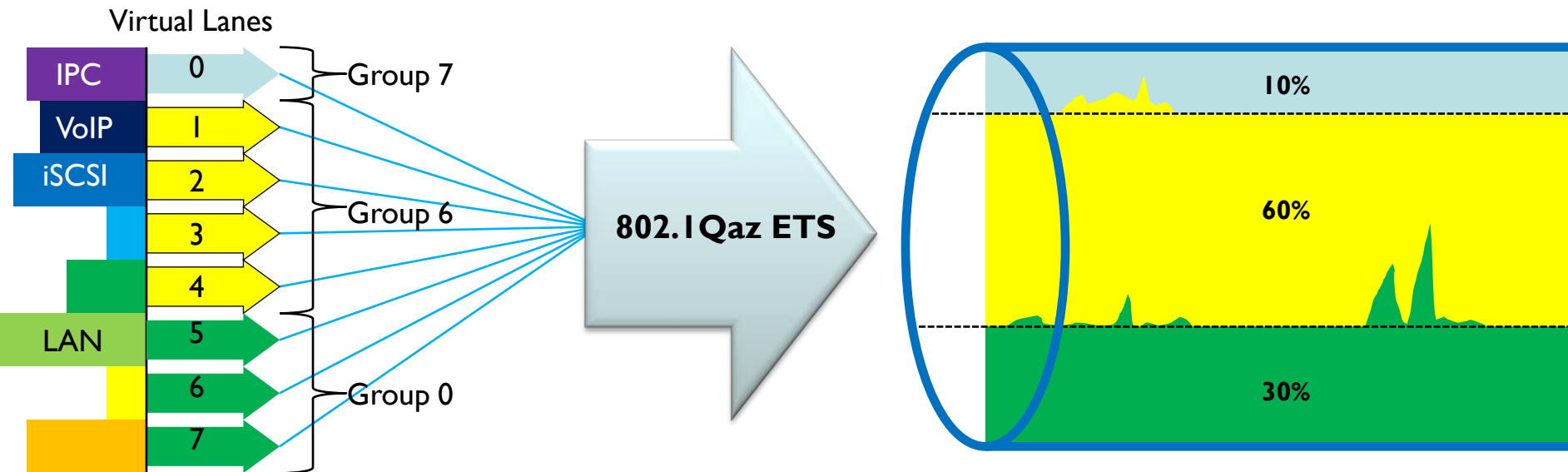- **During periods of heavy congestion**
  - Ensures delivery of critical data without loss
  - Latency sensitive traffic continues normal operation

8 Virtual Lanes

Transmit Queues

Receive Buffers

	0
	1
	2
	3
	4
STOP Pause ц	5
	6
	7

- Capability to apply differentiated treatment to different traffic within the same traffic class enabled by ETS

Congestion Notification
(IEEE 802.1Qau)

Congestion Point

Reaction Point

NIC RL

Switch

NIC RL

Back-off Triggered

Switch

NIC RL

Switch

Switch

NIC

NIC

CN - Congestion Notification gets generated when a device experiences congestion. Request is generated to the ingress node to slow down

RL - In response to CN, ingress node rate-limits the flows that caused the congestion

Priority based Flow Control = Provides insurance against sharp spikes in the confluence traffic, avoids packet drops

Data Center Bridging eXchange (DCBX) Overview
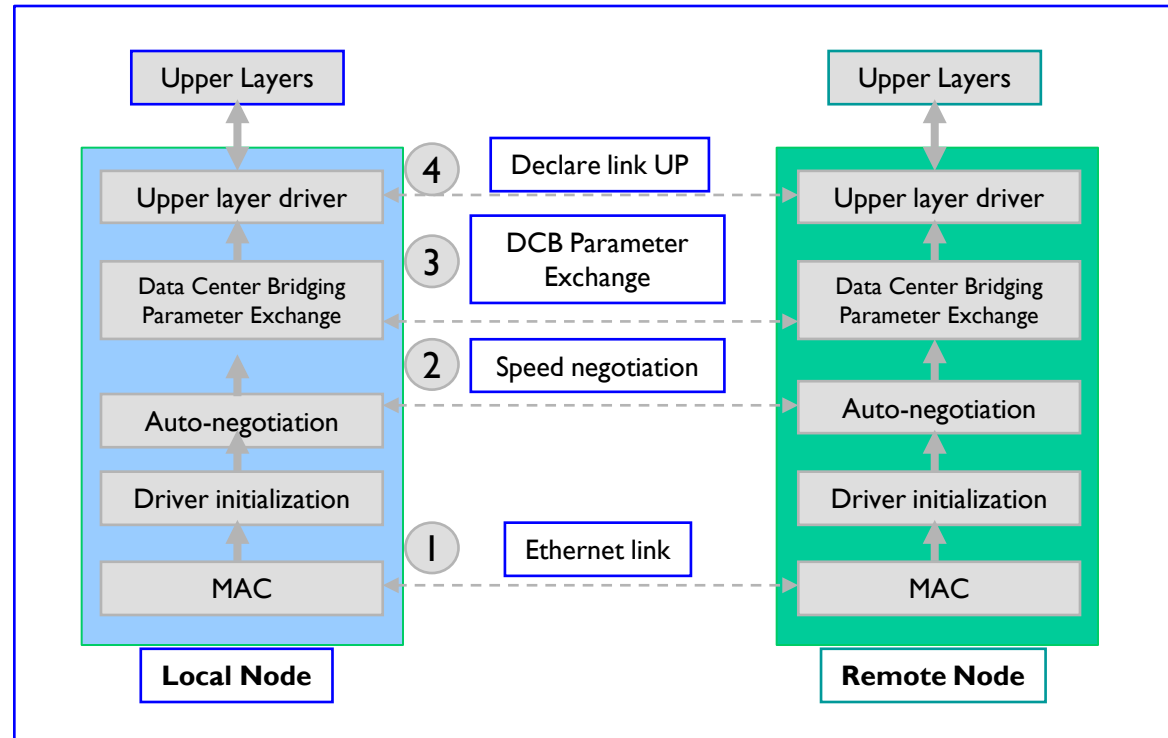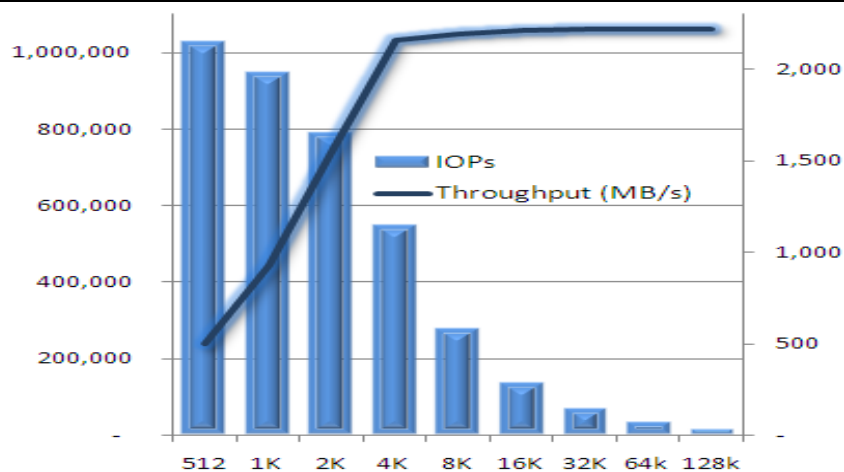
Discovery & Negotiating Capabilities

- DCBX allows two sides to exchange PFC and ETS attributes
- It also allows exchange of Application capabilities like iSCSI
- It uses LLDP as transport to exchange the above capabilities
- DCBX only works for point-to-point interfaces

Transparent configuration of end-nodes

Preliminary iSCSI DCB Results

iSCSI WITH DCB

10G
DCB SWITCH

1

2

10G

10G

10G

10G

10GbE ARRAY

WINDOWS SERVER 2008 x64
10GbE CNA

DCB LINKs
(PFC)

Balanced iSCSI throughput (600MB/s, 600MB/s)
Steady packet streams (no TCP burstiness)

iSCSI WITHOUT DCB

NON-DCB
10G SWITCH

1

2

10G

10G

10G

10GbE ARRAY

WINDOWS SERVER 2008 x64
10GbE CNA

Unbalanced iSCSI throughput (1100MB/s, 100MB/s)
Typical TCP burstiness

iSCSI performance at 10GbE

Read/Write IOPs and Throughput Test

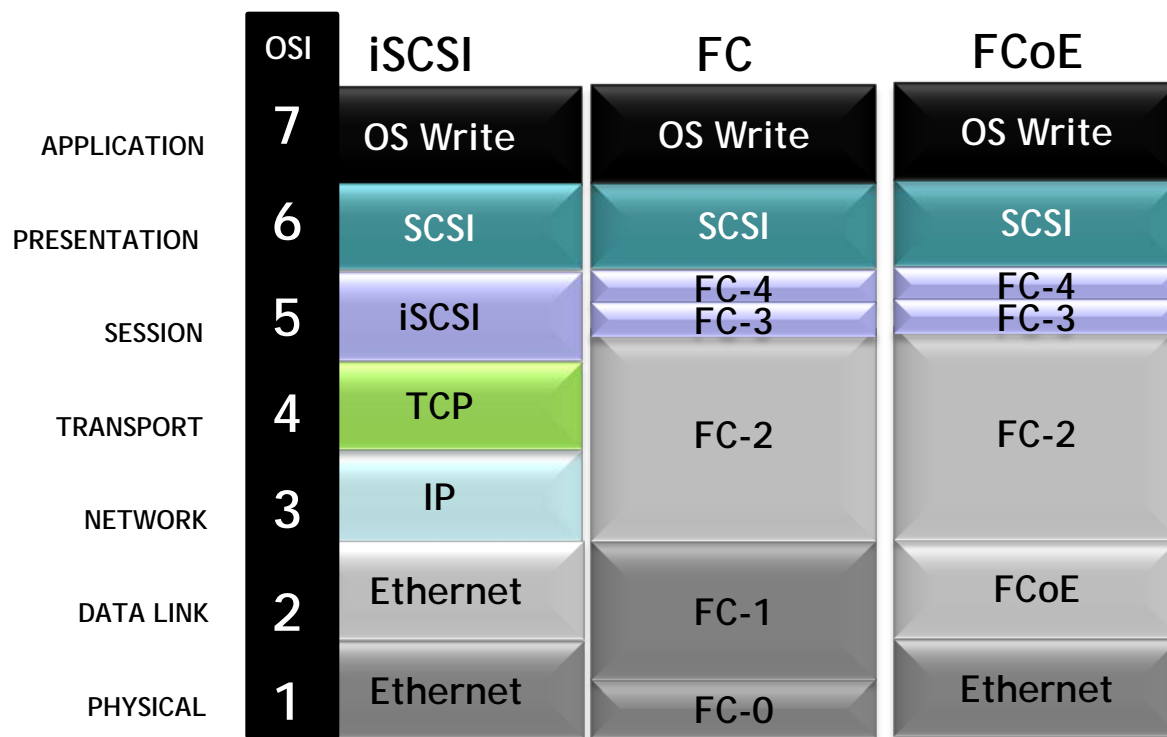**Read/Write IOPs and CPU Test**
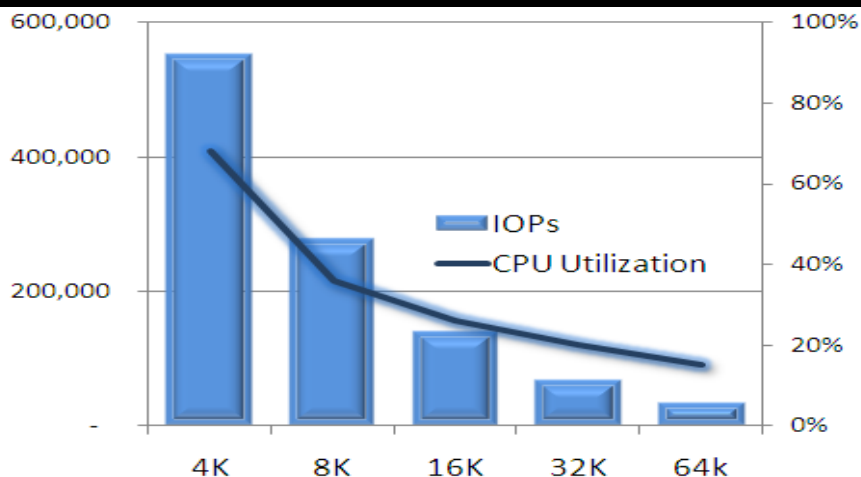
1,030,000 IOPs

- Single Port
- 10GbE line rate
- 10k IOPs per CPU point
- Performance for real world apps
- Future ready: Performance Scales

552k IOPs at 4k represents

- 3,100 Hard Disk Drives
- 400x a demanding database workload
- 1.7m Exchange mailboxes
- 9x transactions of large eTailers
- Jumbo frames: >30% CPU decrease is common for larger IO size (jumbo frames not used here)

Storage protocols

OSI	iSCSI	FC	FCoE
7 APPLICATION	OS Write	OS Write	OS Write
6 PRESENTATION	SCSI	SCSI	SCSI
5 SESSION	iSCSI	FC-4 / FC-3	FC-4 / FC-3
4 TRANSPORT	TCP	FC-2	FC-2
3 NETWORK	IP	FC-2	FC-2
2 DATA LINK	Ethernet	FC-1	FCoE
1 PHYSICAL	Ethernet	FC-0	Ethernet

Fibre Channel is a layered protocol. It consists of 5 layers, namely:

- **FC0** The physical layer, which includes cables, fiber optics, connectors, pinouts etc.
- **FC1** The data link layer, which implements the 8b/10b encoding and decoding of signals.
- **FC2** The network layer, defined by the FC-PI-2 standard, consists of the core of Fibre Channel, and defines the main protocols.
- **FC3** The common services layer, a thin layer that could eventually implement functions like encryption or RAID.
- **FC4** The Protocol Mapping layer. Layer in which other protocols, such as SCSI, are encapsulated into an information unit for delivery to FC2.
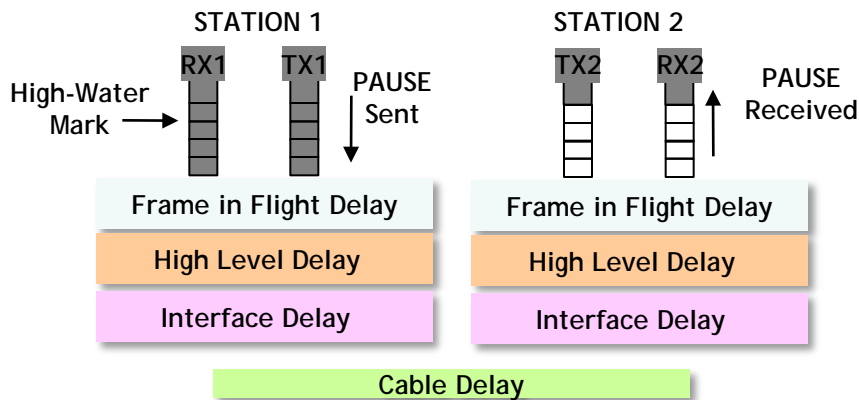
Comparing iSCSI To FCoE

Feature	iSCSI	FCoE
Routable (across LANs and WANs)	Layer 3 & 4	Layer 2 ONLY
MTU	1500 But 9K (Jumbo) Preferred	2500 (Baby Jumbo Frame)
Works over Standard Ethernet	YES	NO
Requires CEE/DCB	**No (but equally benefits)**	**YES in a converged network**
Leverages FC Management Tools	NA	Yes
Requires FC Gateway for Legacy FC	NA	Yes
Flow Control Type	Rate based	Rate based

iSCSI: A Lossless Ethernet Fabric with DCB

Confidential

Source: Q309 – IDC Research

FLOW Control Types

Reactive-Time/Link Dependent

Proactive-Time/Link Independent

PAUSE

STATION 1

RX1　TX1

High-Water Mark →

PAUSE Sent

Frame in Flight Delay

High Level Delay

Interface Delay

STATION 2

TX2　RX2

PAUSE Received

Frame in Flight Delay

High Level Delay

Interface Delay

Cable Delay

CREDITS

STATION 1

TX1

STATION 2

RX2

Counter Incremented

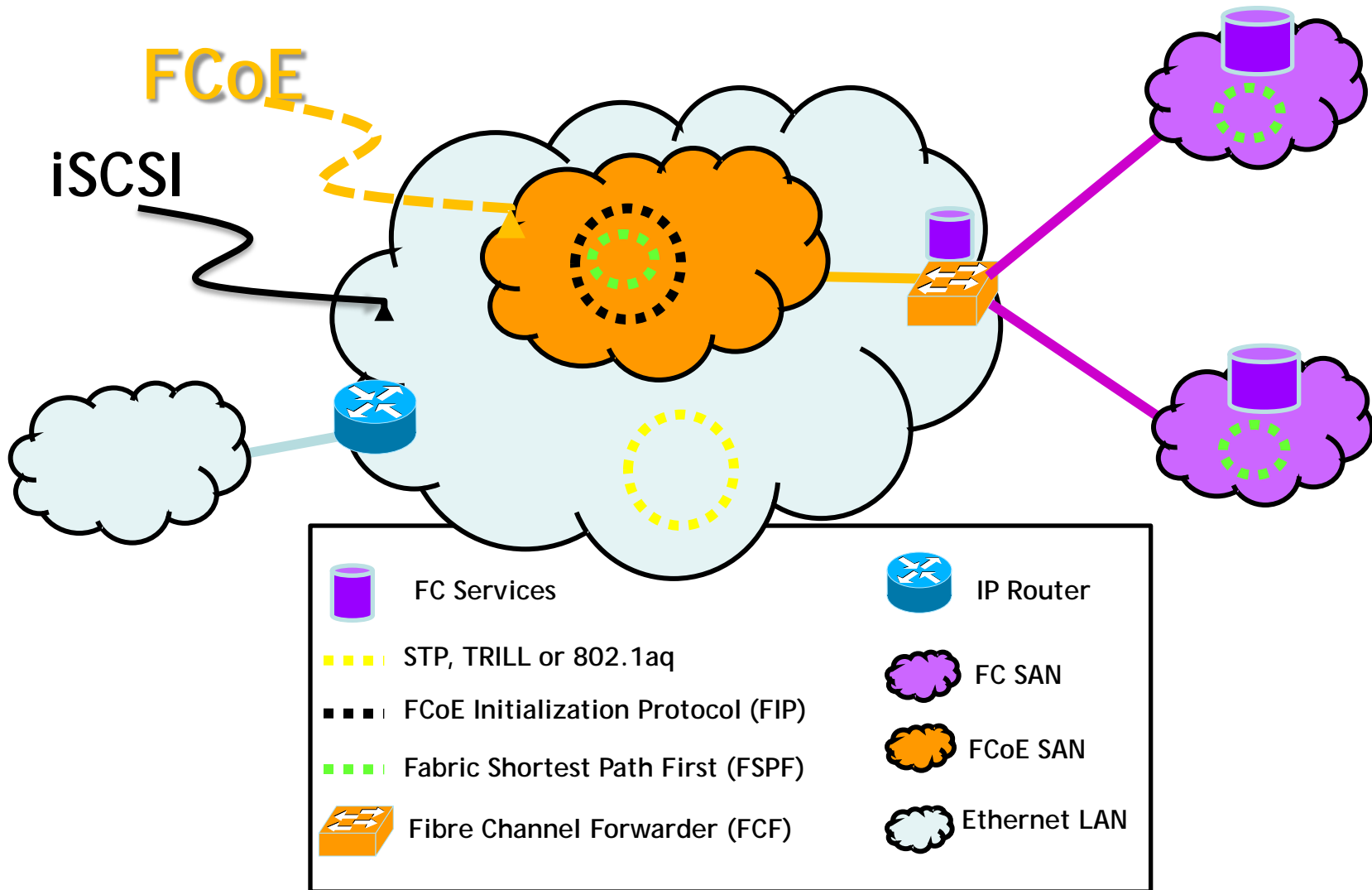Counter Incremented

Buffer Credit Available Message Sent

③

④

②
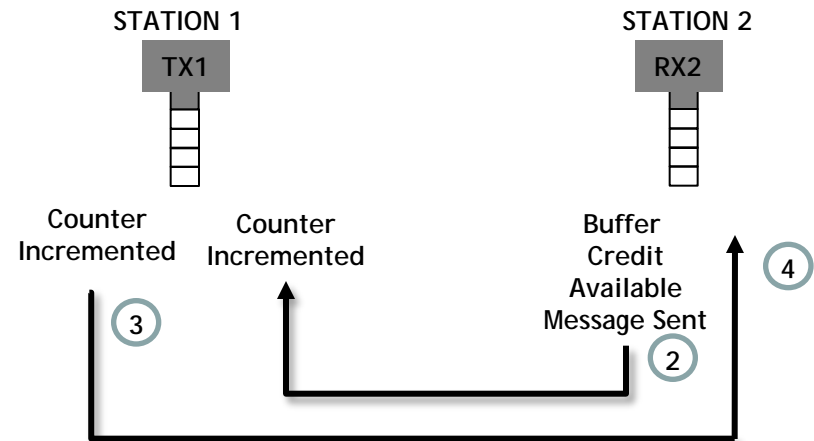
① Ex: Buffer_Credit_CNT= 4

◆ **PAUSE: Pause command sent to sender to halt I/O traffic to avoid overflow of receiving buffers**

◆ **Buffer-Credit: Source and destination negotiate credit count to regulate I/O traffic flow**

Storage Network Components

FCoE

iSCSI

Legend:

Symbol	Description	Symbol	Description
FC Services		IP Router	
STP, TRILL or 802.1aq		FC SAN	
FCoE Initialization Protocol (FIP)		FCoE SAN	
Fabric Shortest Path First (FSPF)		Ethernet LAN	
Fibre Channel Forwarder (FCF)			

Is DCB relevant to iSCSI?

Improves application responsiveness

Provides QoS improvements in mixed traffic environments

Increases robust of Ethernet as a storage fabric

Improves performance of some IP Protocols including iSCSI

Questions?

Q&A / Feedback

SNIA Education

Please send any questions or comments on this presentation to SNIA: tracknetworking@snia.org

> **Many thanks to the following individuals
> for their contributions to this tutorial.**
> **- SNIA Education Committee**
>
> **Jason Blosil**
> **Gary Gumanow**
> **Jordan Plawner**