



Education

## **ETHERNET ENHANCEMENTS FOR STORAGE**

Sunil Ahluwalia, Intel Corporation  
Errol Roberts, Cisco Systems Inc.

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced in their entirety without modification
  - ◆ The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

**NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.**

## ➤ Ethernet Enhancements for Storage

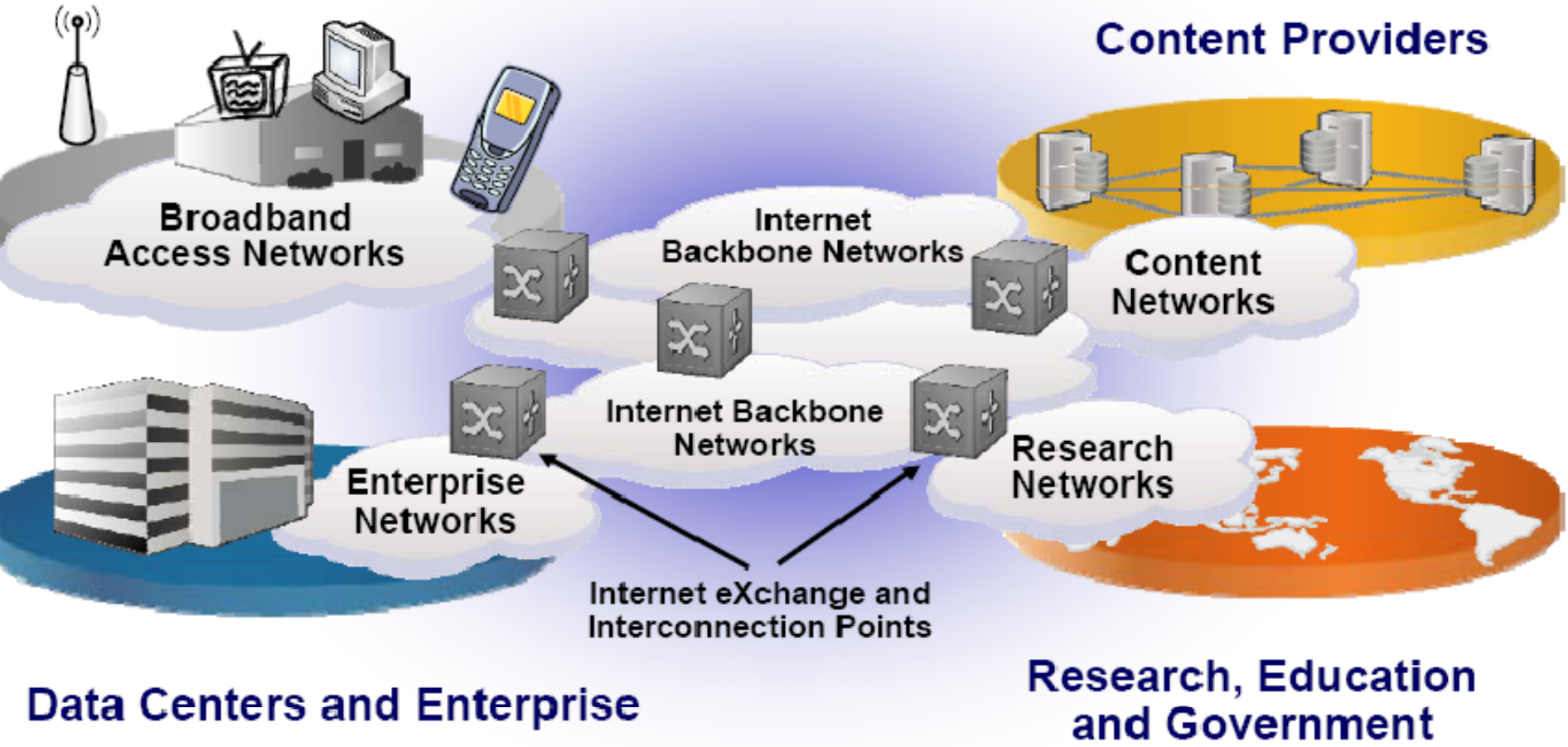
This session discusses the Ethernet enhancements required for storage traffic. It reviews an end-to-end view to evaluate FCoE benefits from a host and switch perspective.

# Agenda

- Ethernet Everywhere!
  
- Data Center Requirements
  
- Ethernet Enhancements
  - ◆ Data Center Bridging
  
- FCoE Deployment

# Ethernet Everywhere!

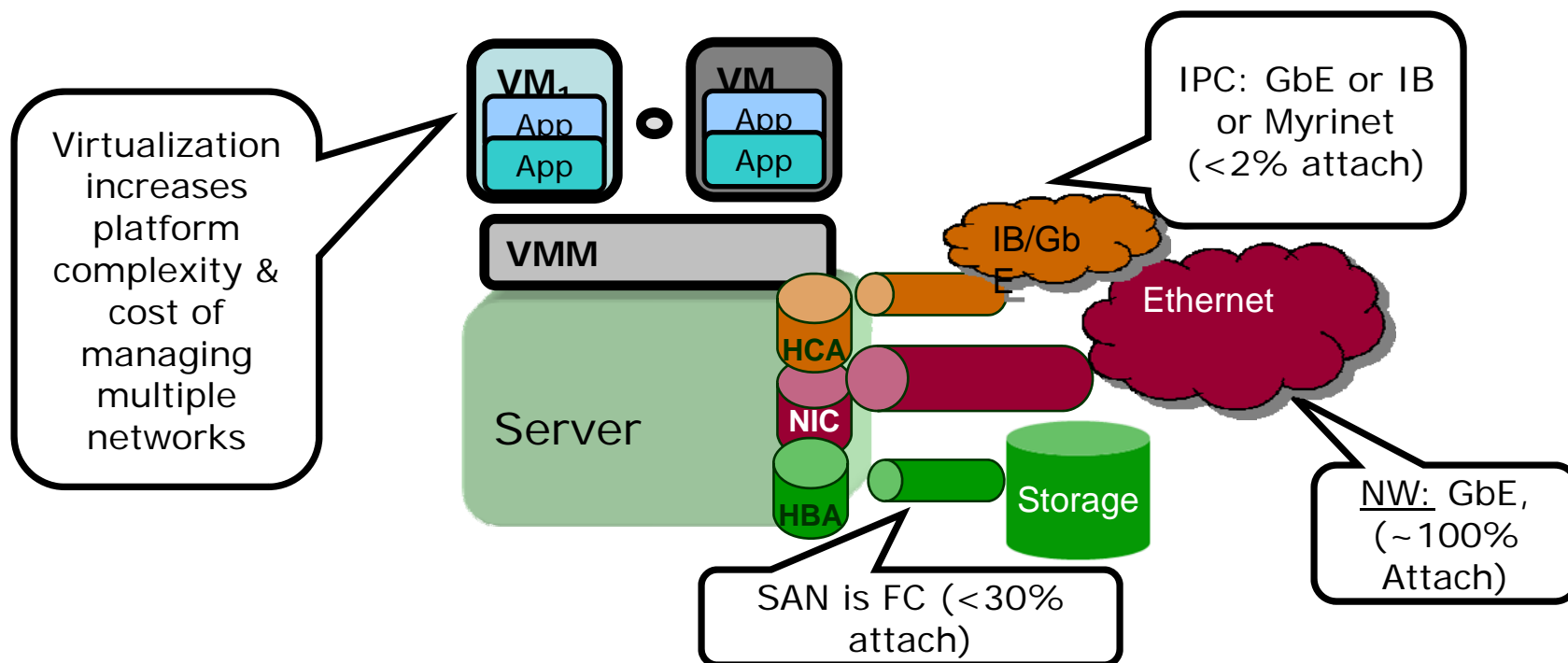
## Broadband Access



## Data Centers and Enterprise

## Research, Education and Government

**Nearly all of the traffic on the Internet either originates or terminates with an Ethernet connection**



## ➤ Multiple networks, one per traffic class

- ◆ IP and other LAN protocols over an Ethernet network
- ◆ SAN over a Fibre Channel network
- ◆ IPC over an InfiniBand network

# Different Network Characteristics

## LAN/IP

- **Must be Ethernet!**
  - Too much investment
  - Too many applications that assume Ethernet
  - Pervasive LAN technology

## Storage

- **FC SAN implementations**
  - lossless requirement over Ethernet
- **IP SAN assumes IP and Ethernet with IP recovery mechanisms**

## IPC (Inter-Process Communication)

- **Transparent to underlying network, provided that**
  - It is cheap
  - It is low latency
  - It supports APIs like OFED, MPI, sockets

# Ethernet Enhancements for Data Center

## ➤ Traffic Differentiation

- ◆ Provides end-to-end traffic differentiation for LAN, SAN and IPC traffic

## ➤ “Lossless” Fabric: Reliable Transport in Ethernet

- ◆ Transient congestion - Priority Based Flow Control
- ◆ Persistent congestion - Congestion Notification

## ➤ Optimal Bridging

- ◆ Allow shortest path bridging within Data Center
- ◆ Eliminates the need to shut off links to prevent loops

## ➤ Configuration management

- ◆ Exchange parameters and work with legacy systems

# Ethernet Enhancements [Data Center Bridging]

# What is Data Center Bridging?

Data Center Bridging is an architectural collection of Ethernet extensions designed to improve Ethernet networking and management in the Data Center.

Sometimes also called

CEE = Converged Enhanced Ethernet

DCB = Data Center Bridging (IEEE)

DCE = Data Center Ethernet (Cisco Trademark)

# IEEE Enhancements for Data Center

- Effort underway to provide DC enhancements in IEEE
  - ◆ 25+ companies actively championing in IEEE
  - ◆ Work is called Data Center Bridging (DCB)
- IEEE projects necessary for I/O Consolidation in Data Center
  - ◆ Congestion Notification: Approved project IEEE 802.1Qau
  - ◆ Shortest Path Bridging: Approved project IEEE 802.1aq
  - ◆ Enhanced Transmission Selection: Approved project IEEE 802.1Qaz
  - ◆ Priority based Flow Control: Approved project in IEEE 802.1Qbb
  - ◆ DCB Capability Exchange Protocol: Part of various projects above
- DCB Standards trending for ratification in 2009/10

## ➤ Link Sharing (Transmit)

- ◆ Different traffic types may share same queues/links
- ◆ Large burst from one traffic should not affect other traffic types

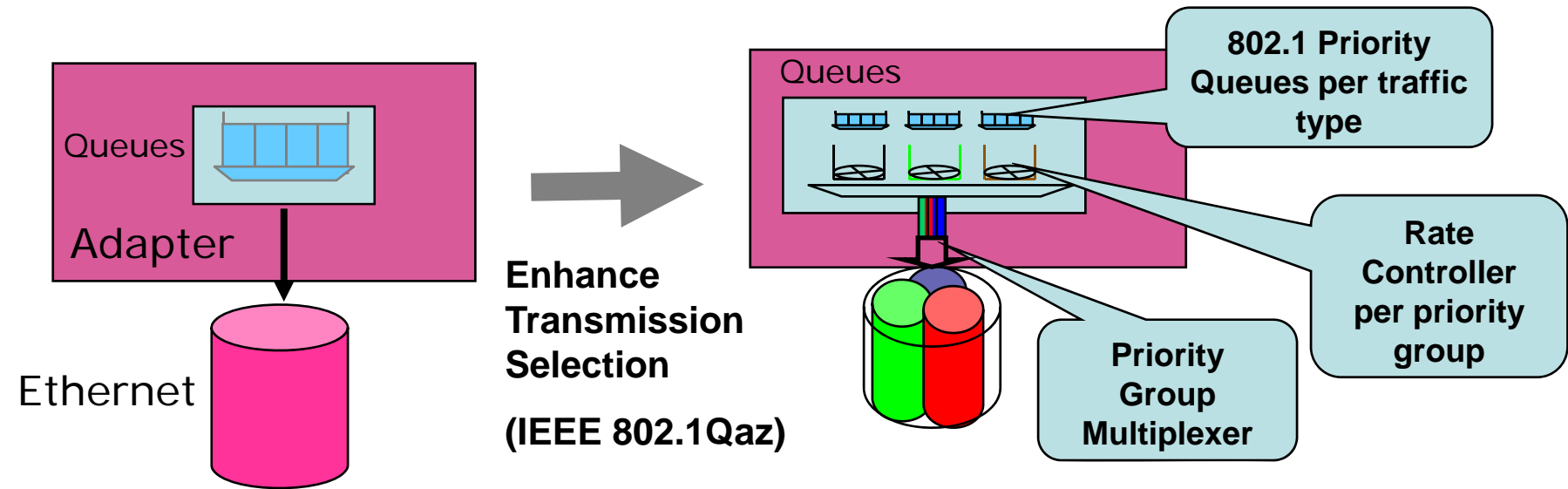
## ➤ Resource Sharing

- ◆ Different traffic types may share some resources (buffers)
- ◆ Large queued traffic for one traffic type should not starve other traffic types out of resources

## ➤ Receive Handling

- ◆ Different traffic types may need different receive handling (eg. interrupt moderation)
- ◆ Optimisation for CPU utilisation for one traffic type should not create large latency for small messages for other traffic types

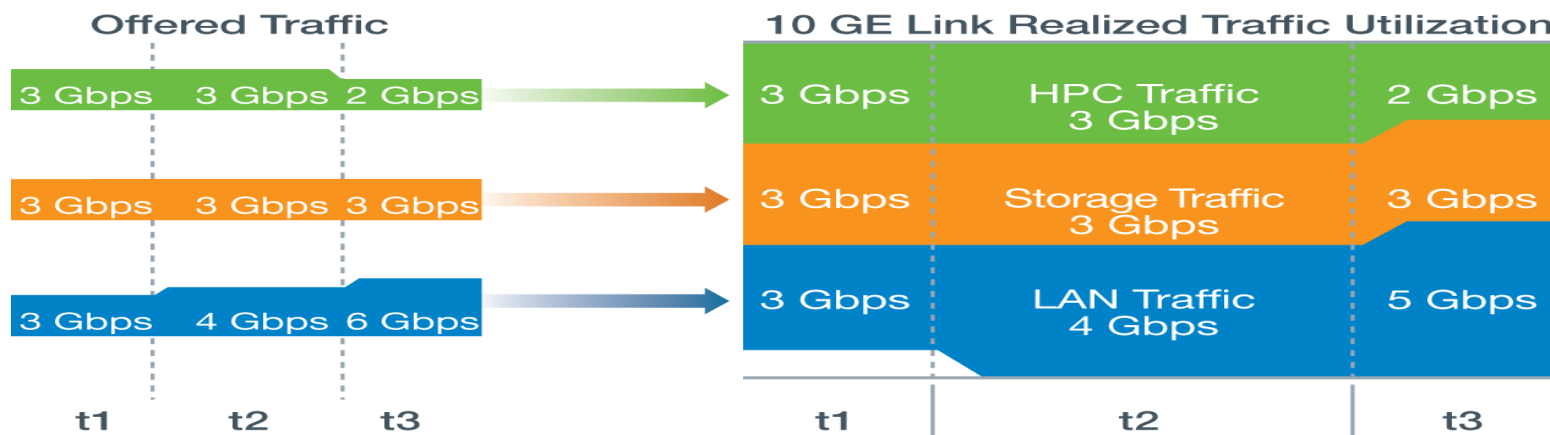
# Traffic Differentiation



- Multiple Link Partitions, One per traffic class
- Resource allocation and association
- Provisioning “aggregate flow bundles”

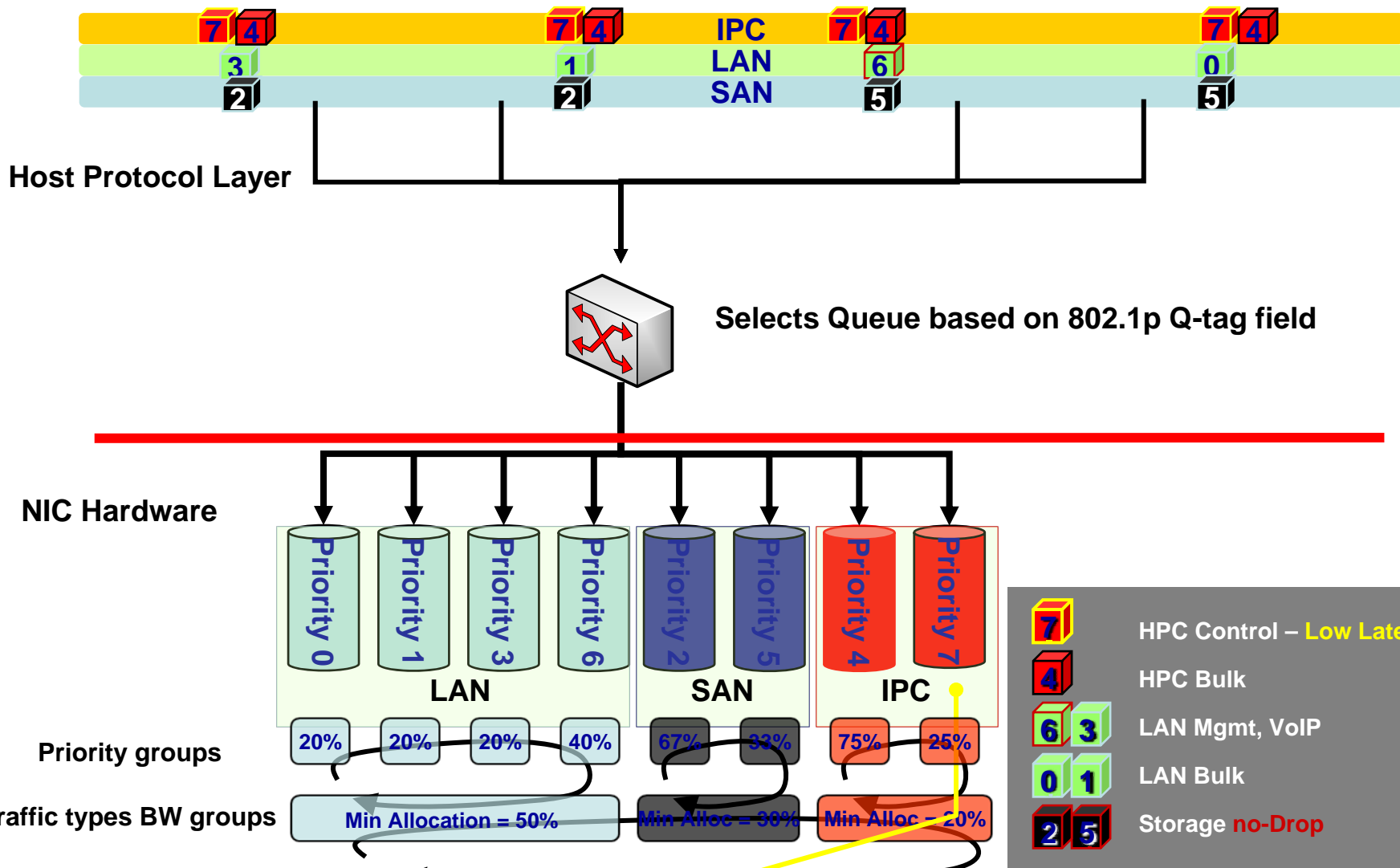
# Enhanced Transmission Selection (IEEE 802.1Qaz)

## ***Priority based Bandwidth Management***

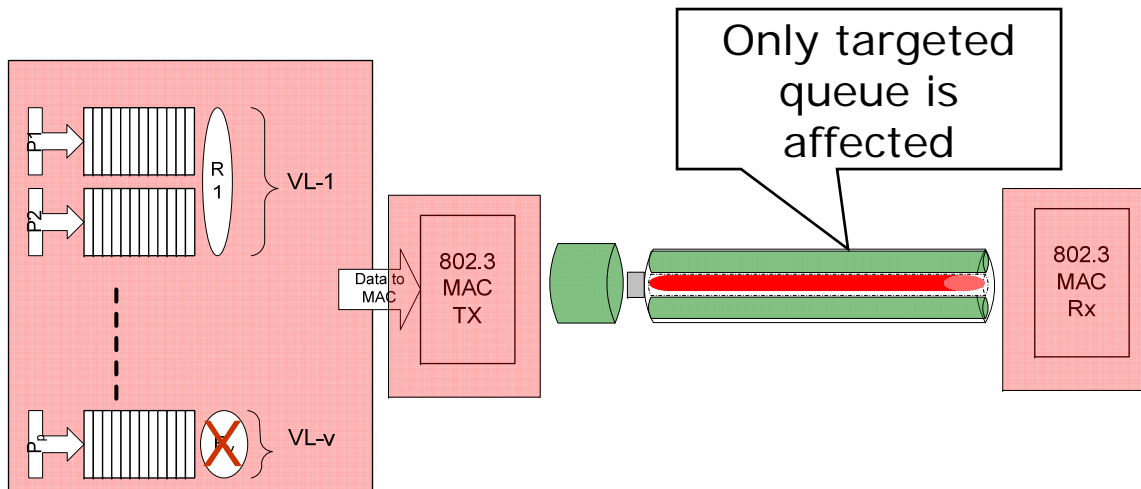
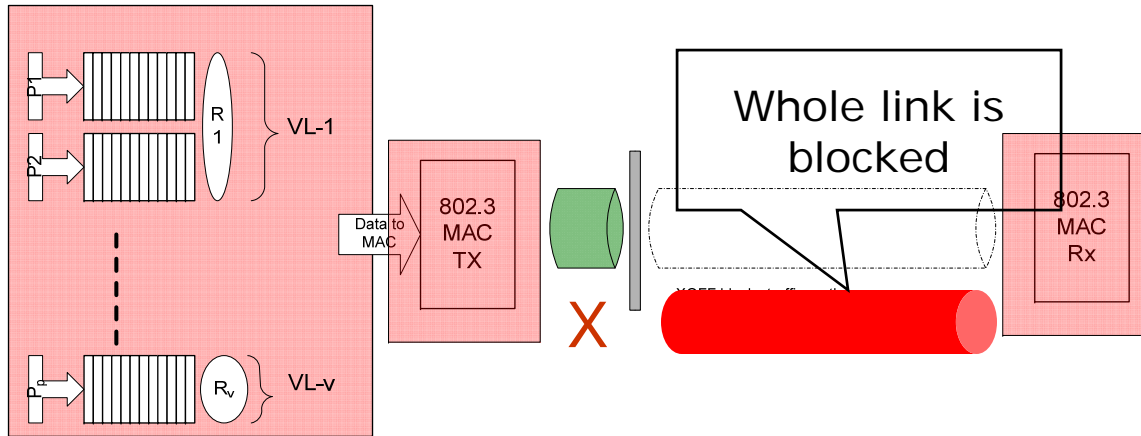


Enables Intelligent sharing of bandwidth between traffic classes control of bandwidth

# Packet Flow



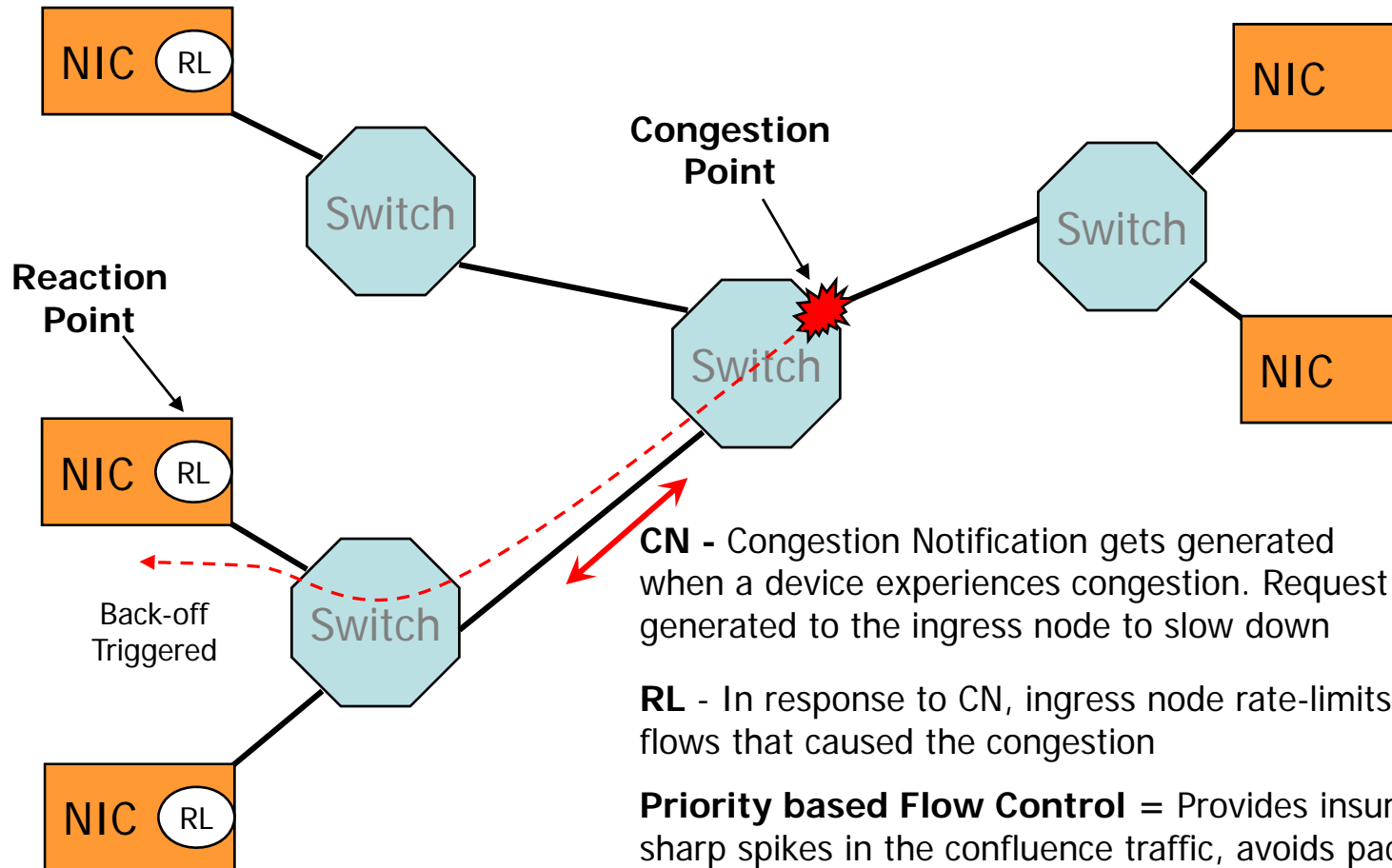
# Priority-based Flow Control (IEEE 802.1Qbb)



# PFC and BB\_Credits

- IEEE 802.3x Pause provides no drop flow control
  - ◆ similar to BB credits for FC
- Priority Flow Control is a finer grained mechanism of flow control over standard pause or link level BB credits
- Priority Flow Control uses .Ip CoS value mapping to a system class to send appropriate pause to previous hop
- The Pause frame is handled by the MAC layer
  - ◆ Similar to the R\_RDY handling by the FC-I level
- The BB\_Credit mechanism allows to not lose frames over any link
  - ◆ Under-utilizing a link if the credits are not enough
  - ◆ Requiring to handle the buffer in maximum frame size units

# Congestion Notification (IEEE 802.1Qau)



**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

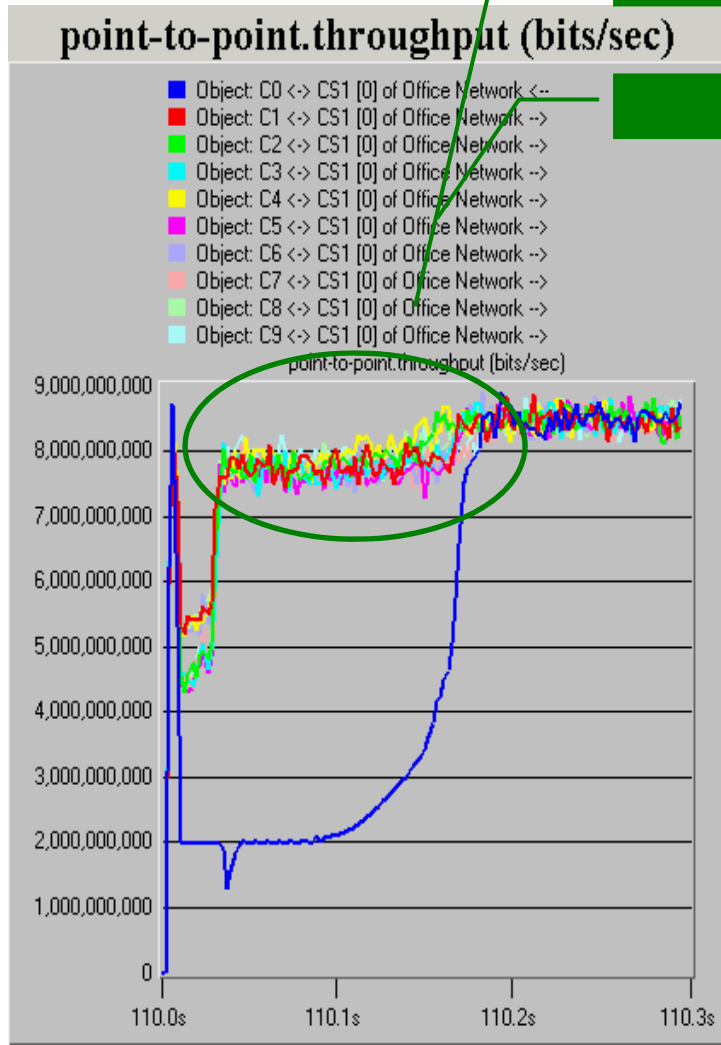
**Priority based Flow Control** = Provides insurance against sharp spikes in the confluence traffic, avoids packet drops

# Congestion Management

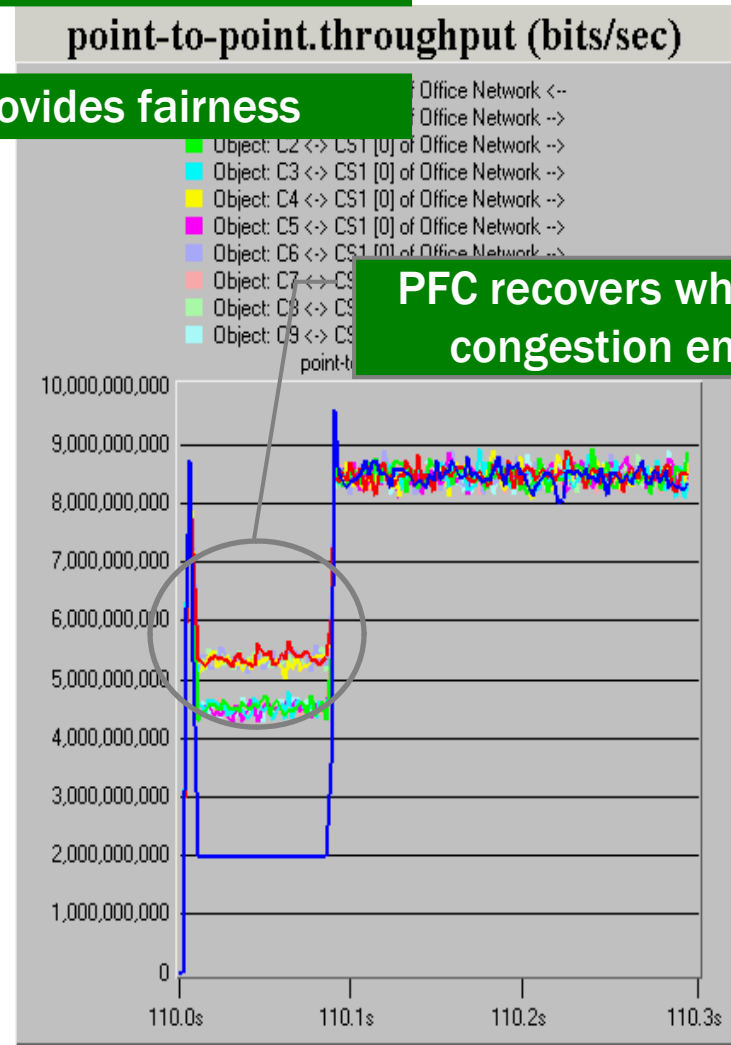
- **PFC is good for transient congestion**
  - ◆ Reacts to avoid packet loss but doesn't diminish congestion.
  - ◆ Congestion spreads upstream which can affect innocent flows.
  - ◆ Unfair low-priority latencies when higher priorities are culprit flows.
- **QCN works on persistent congestion**
  - ◆ Rates reduced to eliminate congestion.
  - ◆ Increased aggregate throughput.
  - ◆ Fairness
  - ◆ Reduced egress buffer usage limits congestion spreading.

# Hotspot Throughput

**QCN rapidly recovers throughput during congestion**



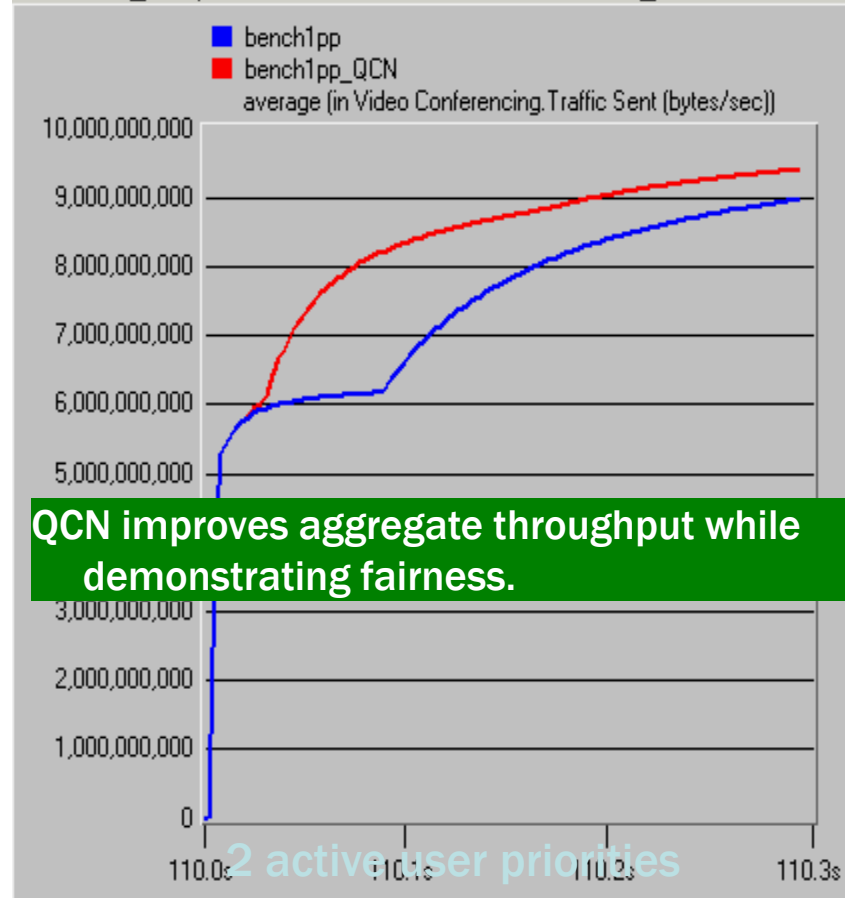
**QCN provides fairness**



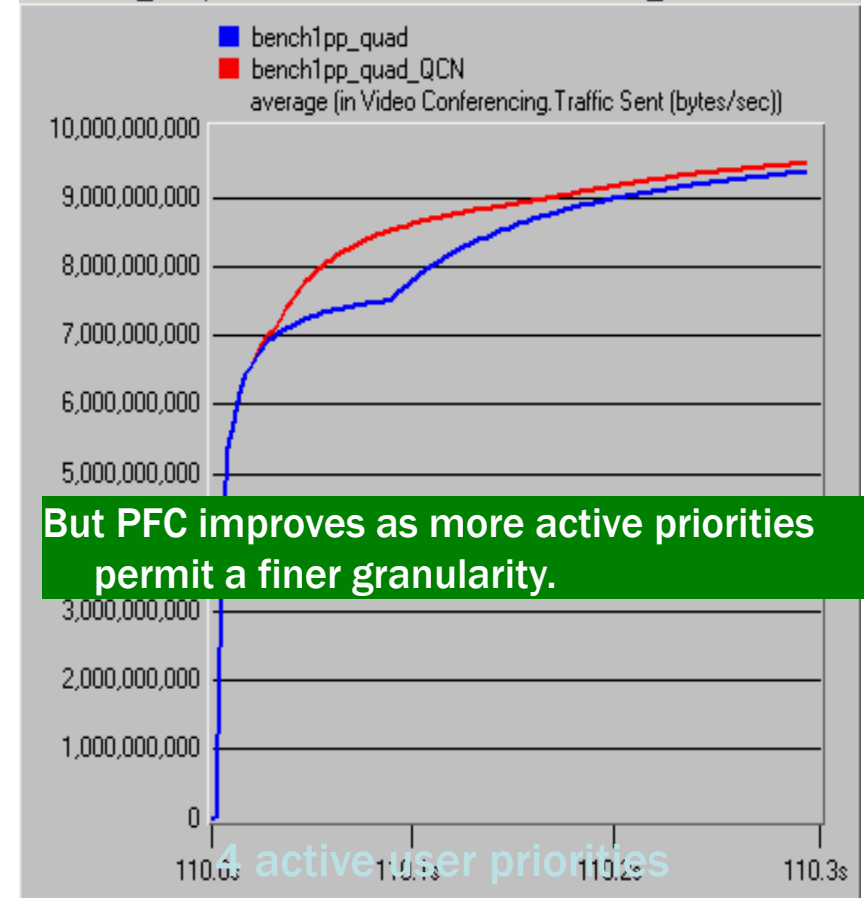
**PFC recovers when congestion ends**

# Aggregate Throughput

average (in Video Conferencing.Traffic S

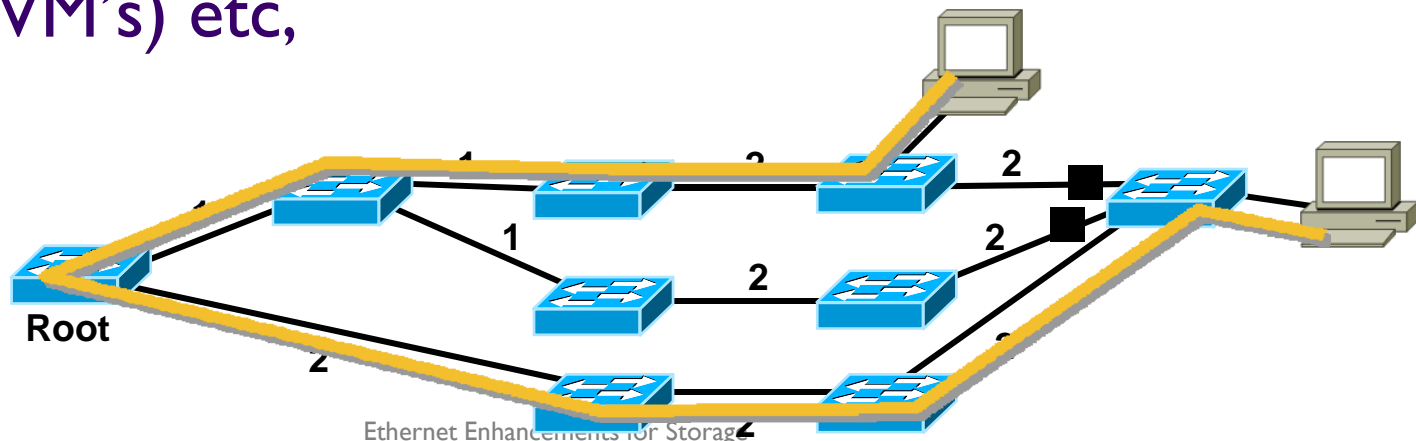


average (in Video Conferencing.Traffic S



## Why?

- Spanning Tree Protocol (STP) and its variants have a bad reputation with customers
  - ◆ Non-optimal forwarding
  - ◆ Parallel paths cannot be leveraged
- These problems can be solved at L3
- But L3 cannot be deployed in many scenarios such as clusters, metro Ethernet, virtualized servers (VM's) etc,



# Shortest Path Bridging - 802.1aq

## Another Approach to Shortest Path Bridging

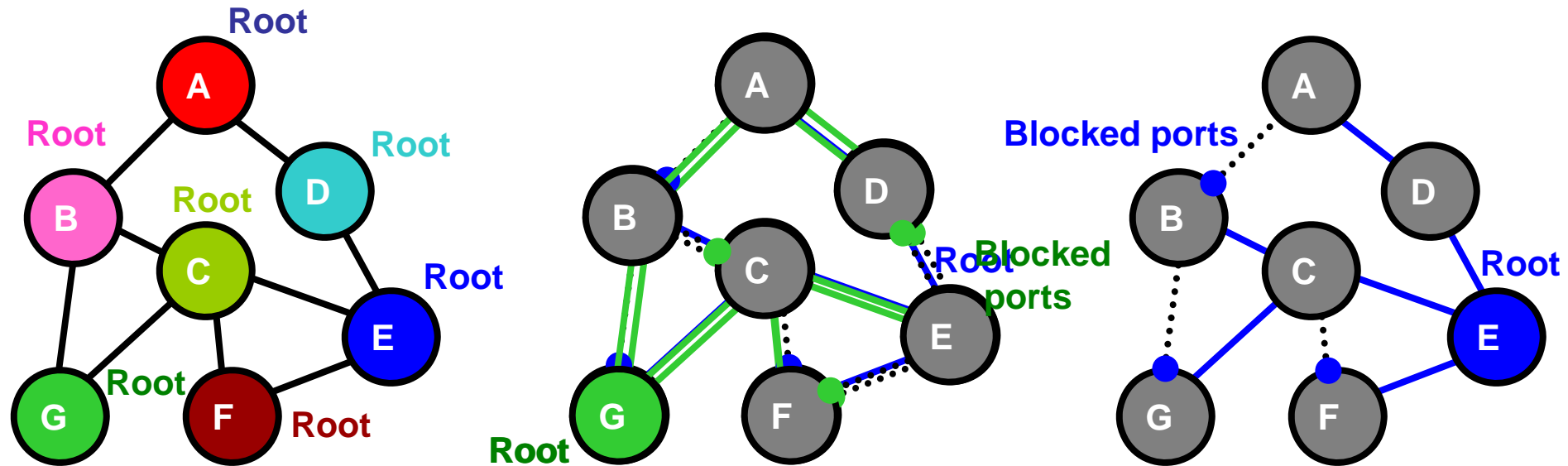
- What is it –
  - ◆ Enhancement to 802.1Q to provide Shortest Path Bridging (Optimal Bridging) in L2 Ethernet topologies
  - ◆ Provides for each bridge to be the root of its own topology and hence uses the “best” path to any destination
- Benefits –
  - ◆ Resolves issues related to root disappearance
  - ◆ Fast convergence – no count to infinity
- Does not require a link state protocol (unlike TRILL)
- Resources -

<http://www.ieee802.org/1/files/public/docs2005/aq-nfinn-shortest-path-0905.pdf>

<http://www.ieee802.org/1/files/public/docs2006/aq-nfinn-shortest-path-2-0106.pdf>

# 802.1aq - Spanning Tree per bridge

How does it work



Each bridge is the root of a separate spanning tree instance.

Bridge G is the root of the green tree

Bridge E is the root of the blue tree

Both trees are active at all times

## ➤ What is it –

- ◆ “Transparent Interconnection of Lots of Links” Internet Drafts (also called “Routing Bridges” or “RBridges”)
- ◆ IETF effort to solve L2 STP forwarding limitations
- ◆ TRILL is a solution intended for data centers (and campuses) to provide connectivity among end stations with ease of current bridges but without using spanning tree protocol
- ◆ Replaces STP with a link-state routing protocol to discover the topology

## ➤ Benefits –

- ◆ Shortest-Path Frame routing in multi-hop 802.1-compliant networks
- ◆ Permits Load Splitting among multiple paths
- ◆ Forwarding based on destination bridge-id – smaller tables than conventional bridge systems

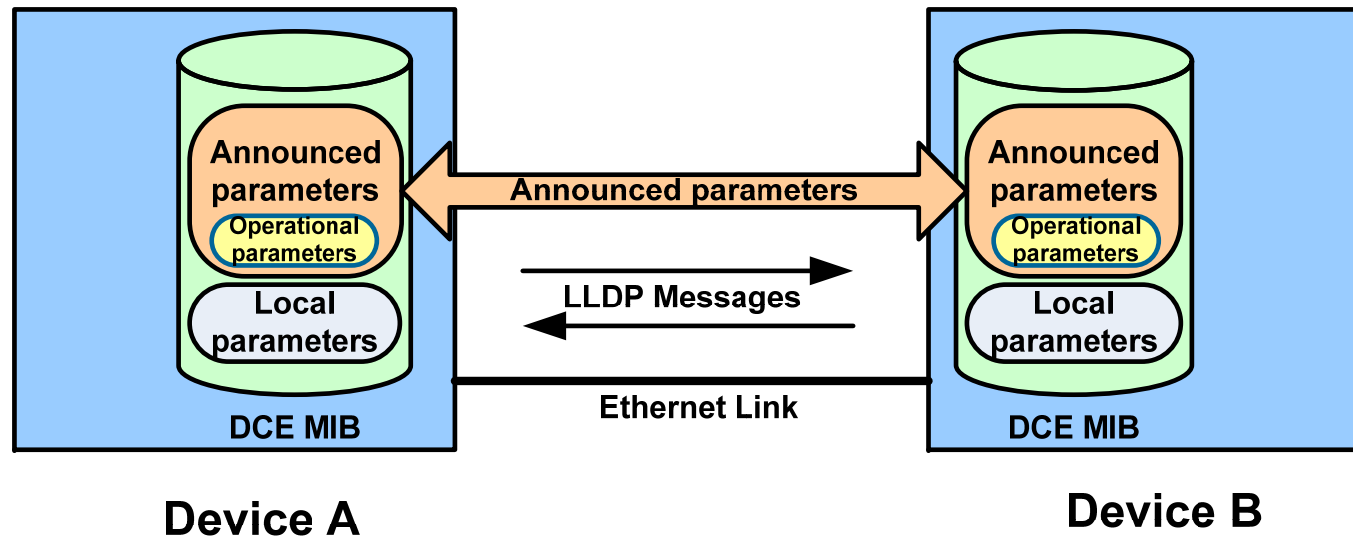
## ➤ Must be backward compatible with 802.1d – inter-working at the edge

## ➤ Resources -

- ◆ <http://www.ietf.org/html.charters/trill-charter.html>
- ◆ <http://www.ietf.org/internet-drafts/draft-ietf-trill-rbridge-protocol-03.txt>
- ◆ Dinesh Dutt (Cisco)

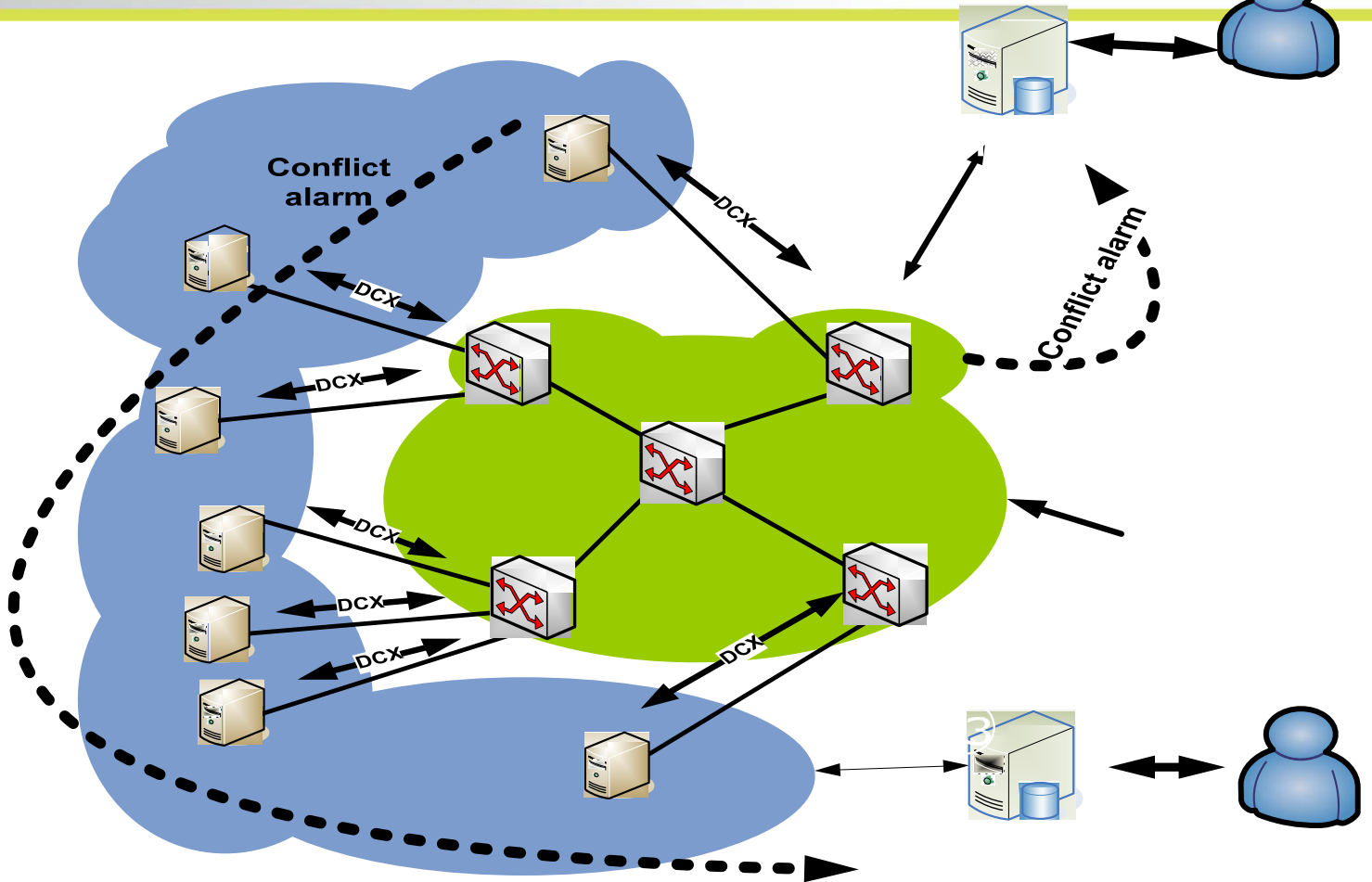
- Link level capability and configuration exchange
  - ◆ Similar to FLOGI and PLOGI in Fibre Channel
  - ◆ Allows either full configuration or configuration checking
- Based on LLDP (Link Level Discovery Protocol)
  - ◆ Added reliable transport
  - ◆ Link partners can choose supported features and willingness to accept configuration from peer
- Feature TLVs
  - ◆ Priority Groups (Link Scheduling)
  - ◆ Priority-based Flow Control
  - ◆ Congestion Management (Backwards Congestion Notification)
  - ◆ Application (frame priority usage)
  - ◆ Logical Link Down

# Configuration Management



- ◆ **DCBX is a protocol between link peers to exchange DCB parameters and capabilities**
- ◆ **It uses LLDP (Link Layer Discovery Protocol)**
- ◆ **Announced DCB Parameters**
  - ◆ *Bandwidth Group ID*: Link bandwidth percentage
  - ◆ *User Priority*: Bandwidth Group ID, Bandwidth Group BW percentage, and QCN capabilities
  - ◆ Administrative and Operational Modes

# Example DCBX Deployment Model

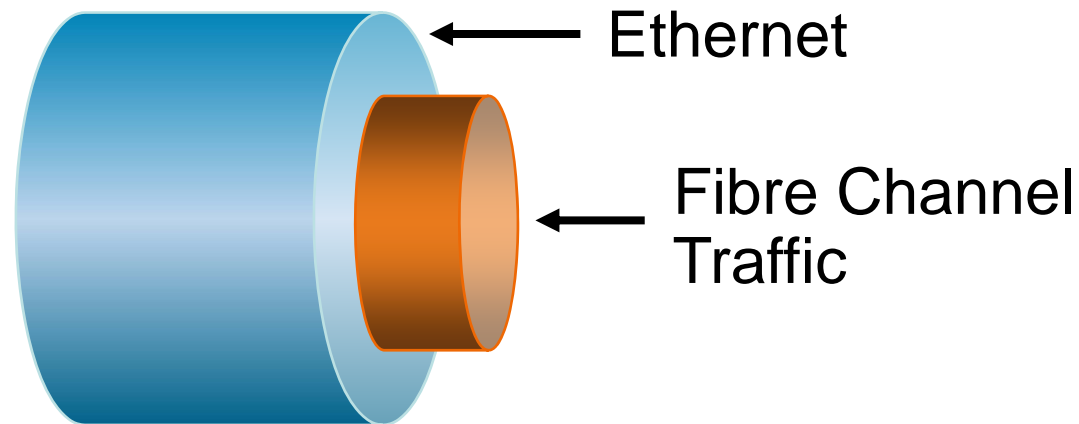


- Detects configuration mismatches between link peers and notifies Management
- Discovers DCB related peer capability
- Detect boundaries of congestion management

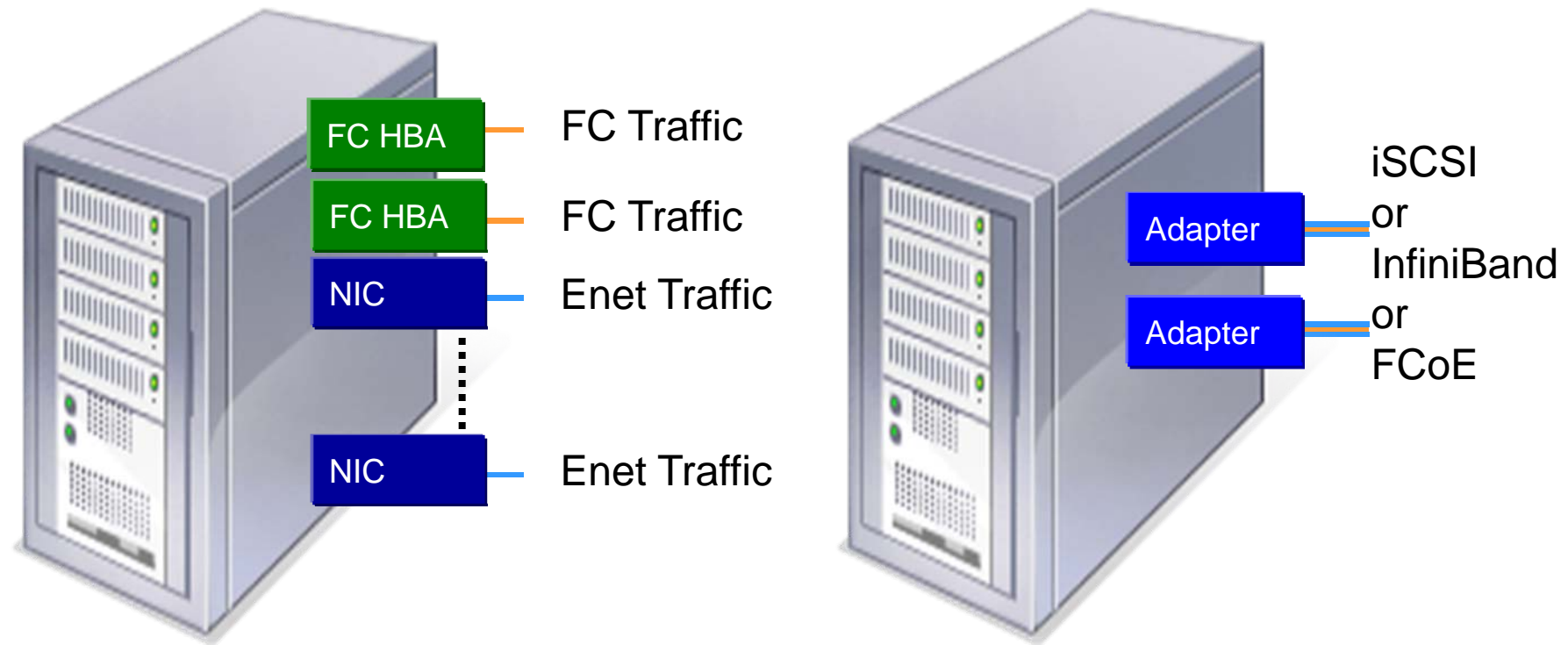
# **FCoE and I/O Consolidation [Server Perspective]**

# FCoE: FC over Ethernet

- FCoE is I/O consolidation of FC storage traffic over Ethernet
  - ◆ FC traffic shares Ethernet links with other traffics
  - ◆ Requires a lossless Ethernet fabric



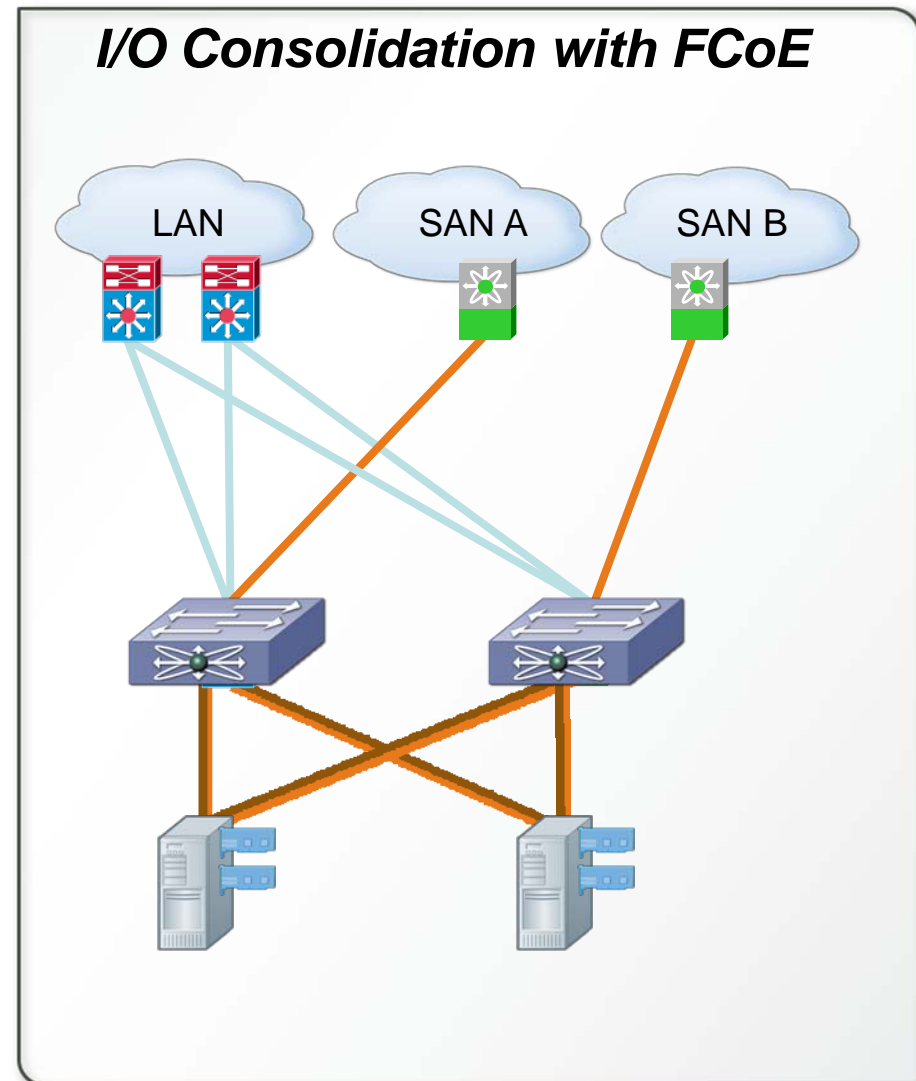
# Server I/O Consolidation



- ◆ **Adaptor:** NIC for Ethernet/IP, HCA for InfiniBand, Converged Network Adaptor (CNA) for FCoE
- ◆ **Customer Benefit:** Fewer NIC's, HBA's and cables, lower CapEx, OpEx (power, cooling)

## ➤ I/O consolidation

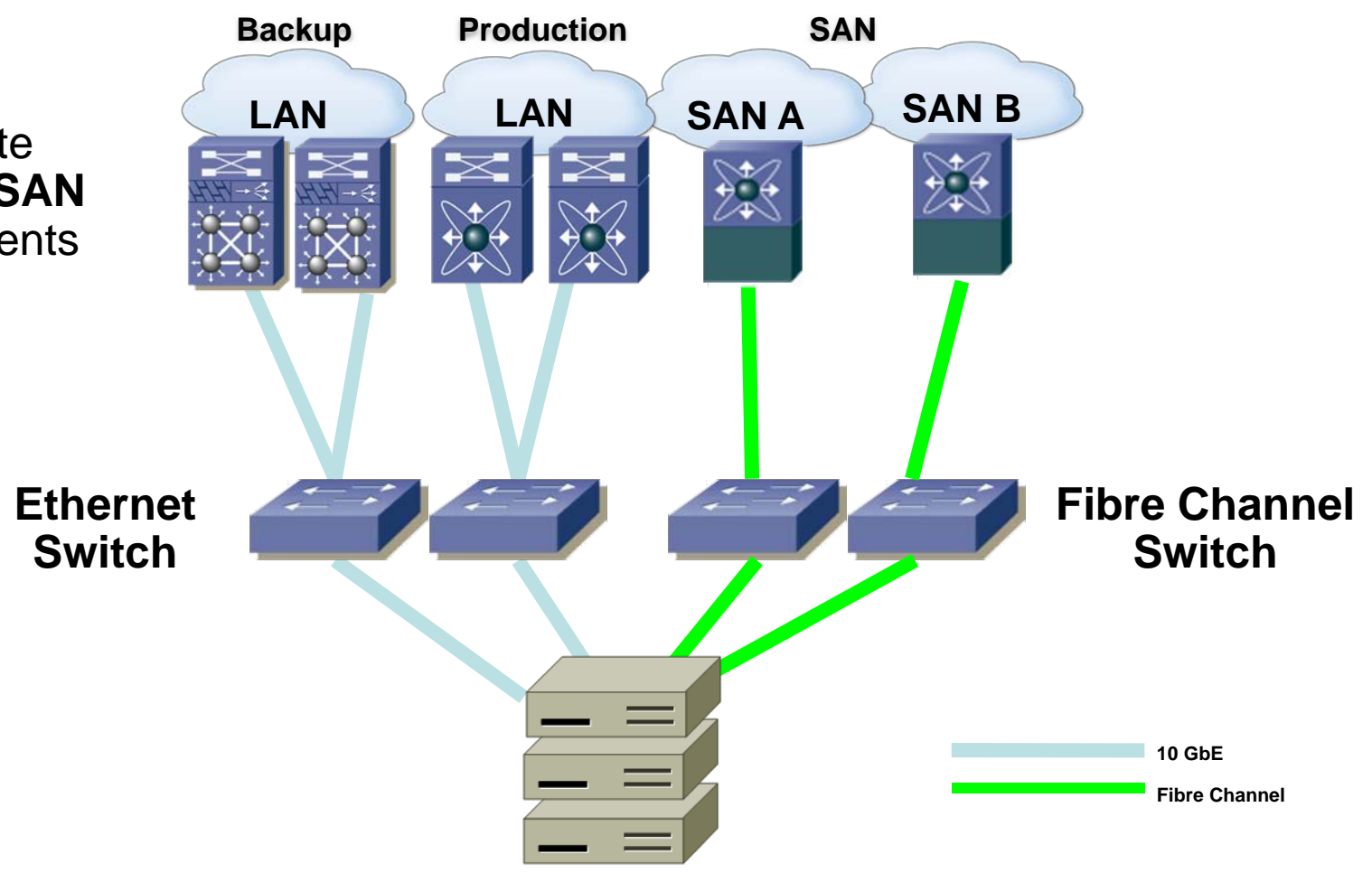
- ◆ Reduction of server adapters
- ◆ Simplification of access layer & cabling
- ◆ Gateway free implementation – fits in installed base of existing LAN and SAN
- ◆ L2 Multipathing Access – Distribution
- ◆ Lower TCO
- ◆ Fewer Cables
- ◆ Investment Protection (LANs and SANs)
- ◆ Consistent Operational Model



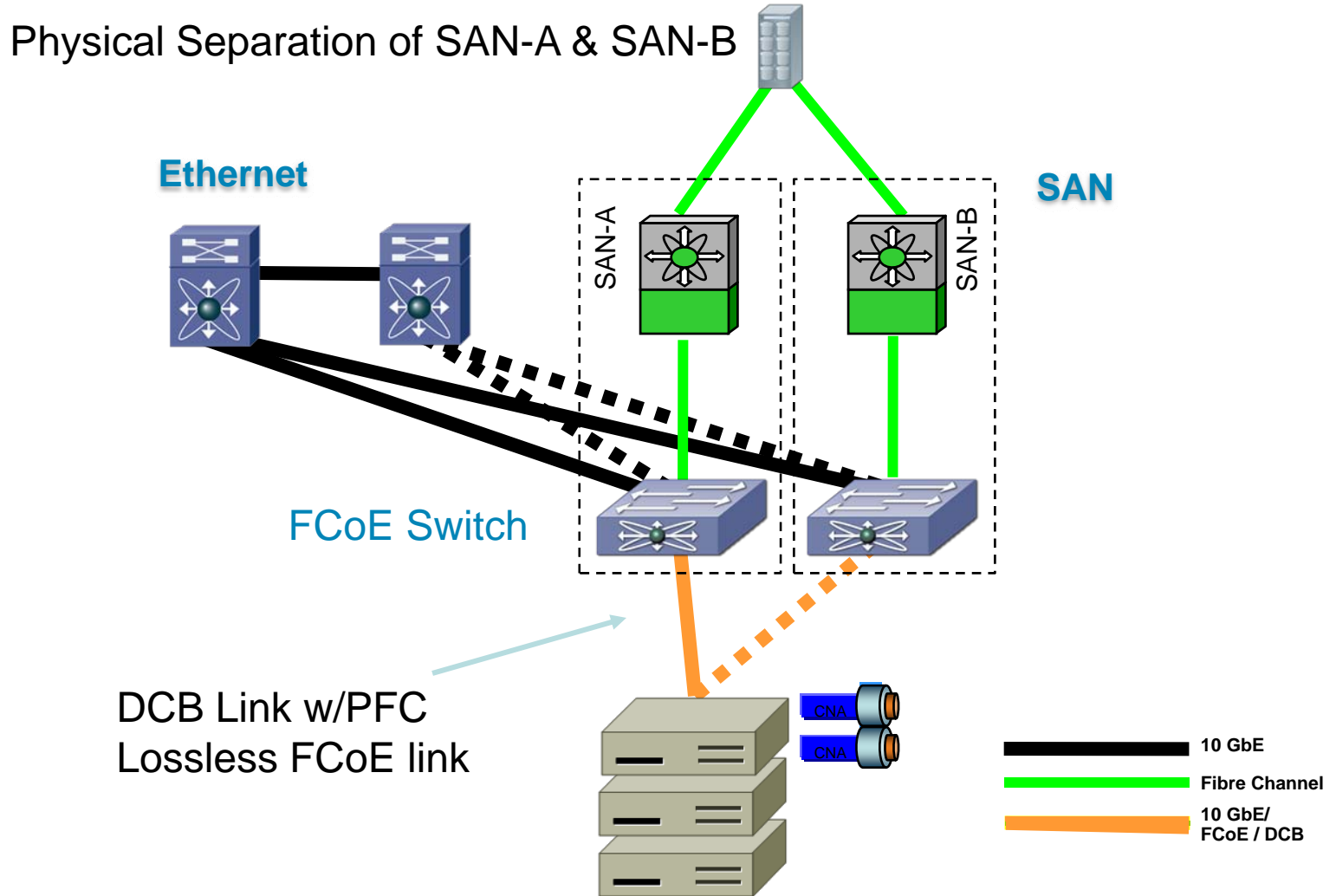
# FCoE Deployment

# Typical Data Center Server Access Layer Topology

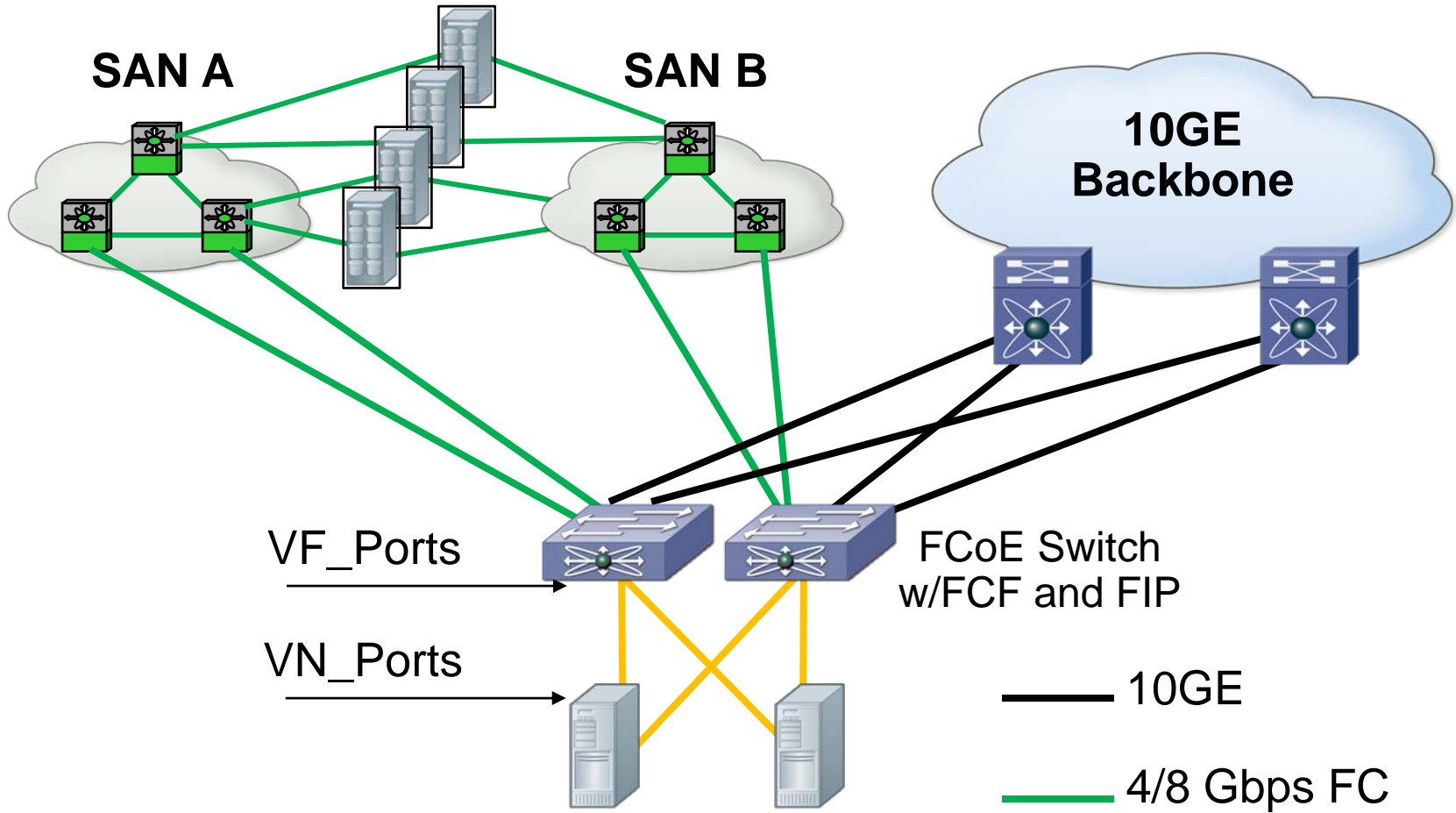
Separate  
**LAN** and **SAN**  
Environments



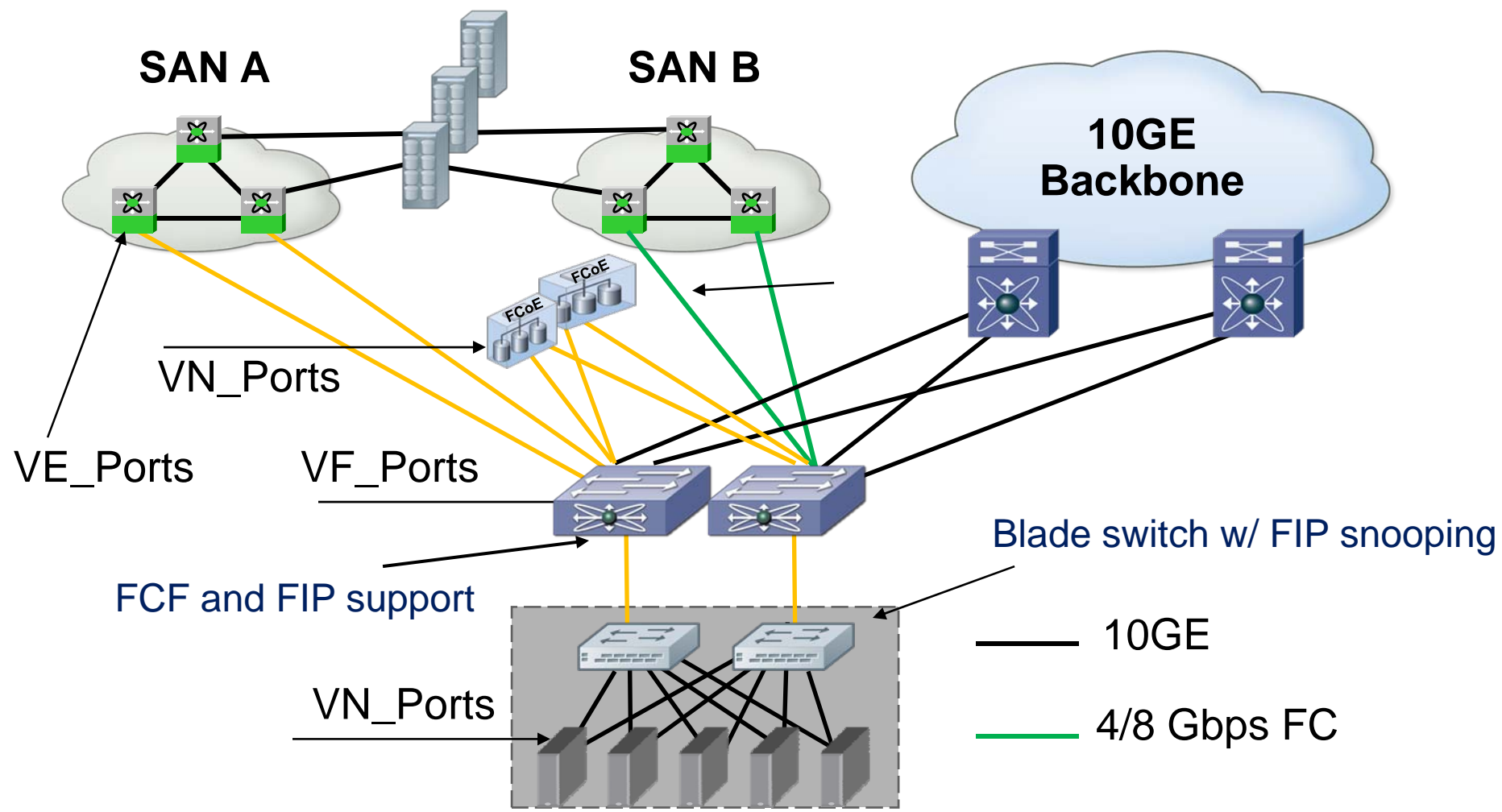
# FCoE Access Model



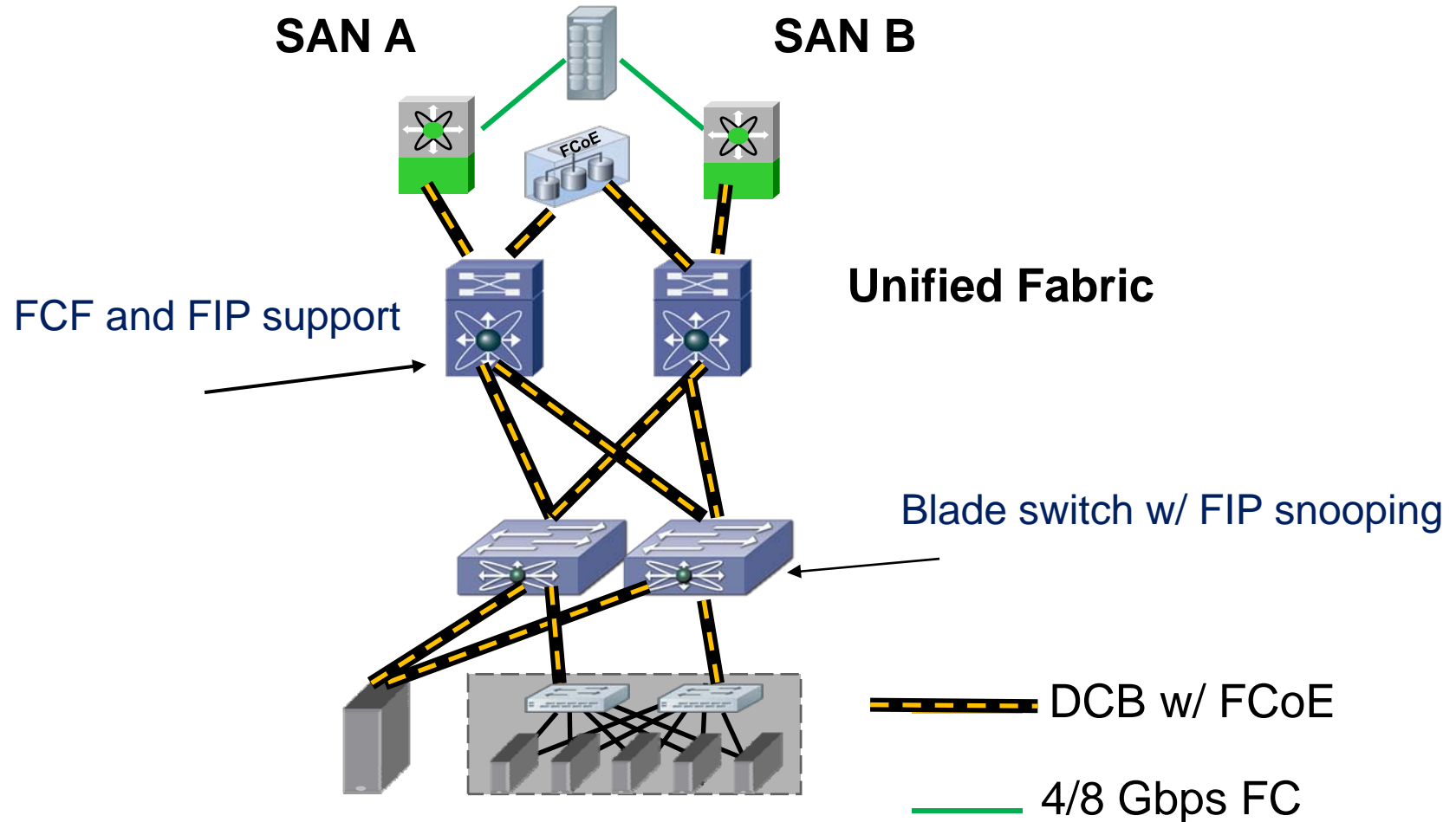
# FCoE: Initial Deployment



# FCoE: Adding Native FCoE Storage



# FCoE: Multi Tiered



# Summary

- Data Center Bridging standards are driving Ethernet Enhancements for multiple traffic types
- Lossless 10GbE is the fabric for I/O consolidation
- Early adoption of FCoE is in the access layer

- Please send any questions or comments on this presentation to SNIA: [tracknetworking@snia.org](mailto:tracknetworking@snia.org)

**Many thanks to the following individuals  
for their contributions to this tutorial.**

**- SNIA Education Committee**

**Rob Peglar  
Walter Dey  
Steve Wilson  
Joe White**