



Education

## **Ethernet Technology**

Sunil Ahluwalia, Intel Corporation  
Manoj Wadekar, Intel Corporation

# SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced without modification
  - ◆ The SNIA must be acknowledged as 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.

## ➤ Consolidating I/O over Ethernet

- ◆ The audience will learn how hardware, software, standards and innovation are all needed to derive the vision of a unified fabric utilizing 10GE for the data center of the future. The unification of fabrics is not an all or nothing approach. There is a clearly defined unified fabric sweet spot within a data center. The audience will learn to identify and assess if this approach is right for their data center and business level objectives. In this session you will learn the differences between reactive innovation vs proactive innovation and how it applies to 10G, Acceleration, OS virtualization, HW virtualization and the evolution of storage fabrics.

## ➤ Overview of Ethernet Technology

- ◆ Ethernet Evolution
- ◆ Frame Format

## ➤ 10 Gigabit Ethernet Technology

- ◆ Demand for 10GbE
- ◆ 10GbE PHY technology

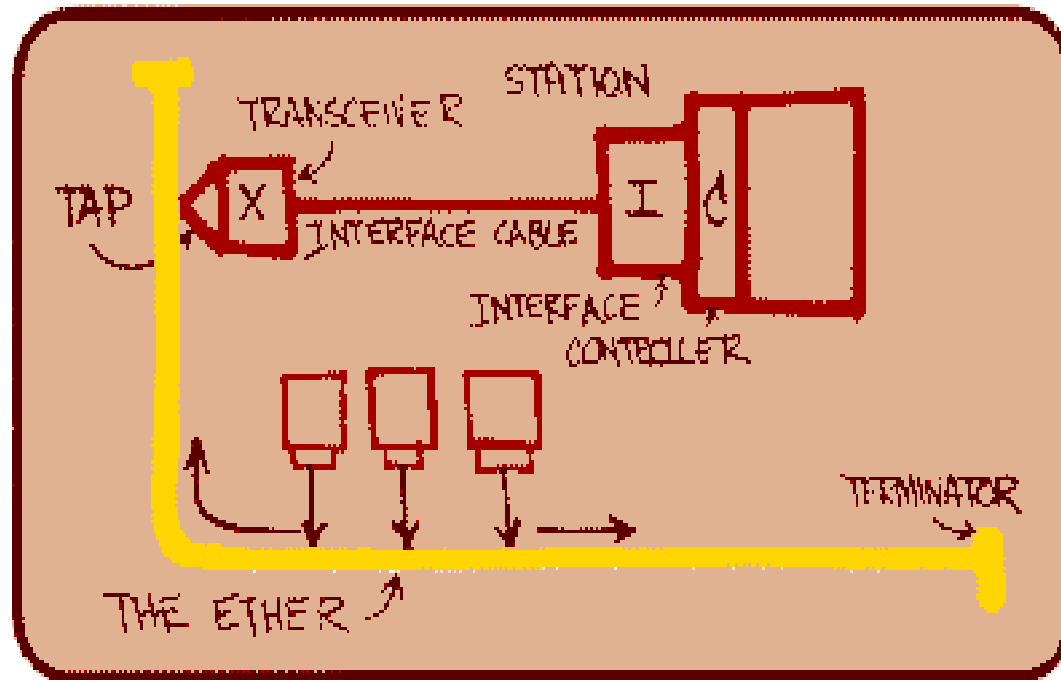
## ➤ I/O Consolidation over Ethernet

- ◆ Ethernet enhancements for “lossless” fabric

## ➤ Next Generation Ethernet

- ◆ 40G / 100G

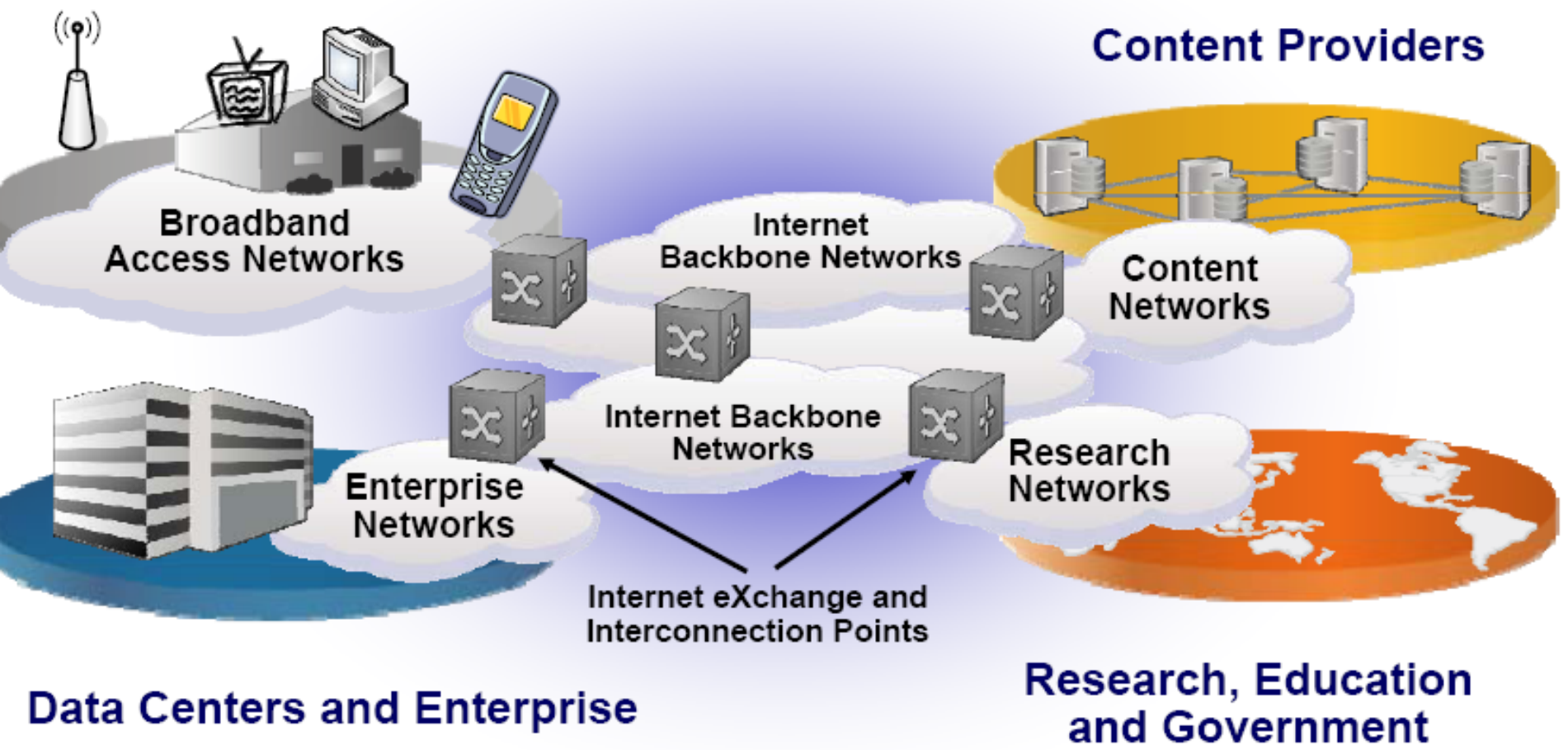
# Original Ethernet



June 1976 National Computer Conference

# Ethernet Everywhere!

## Broadband Access



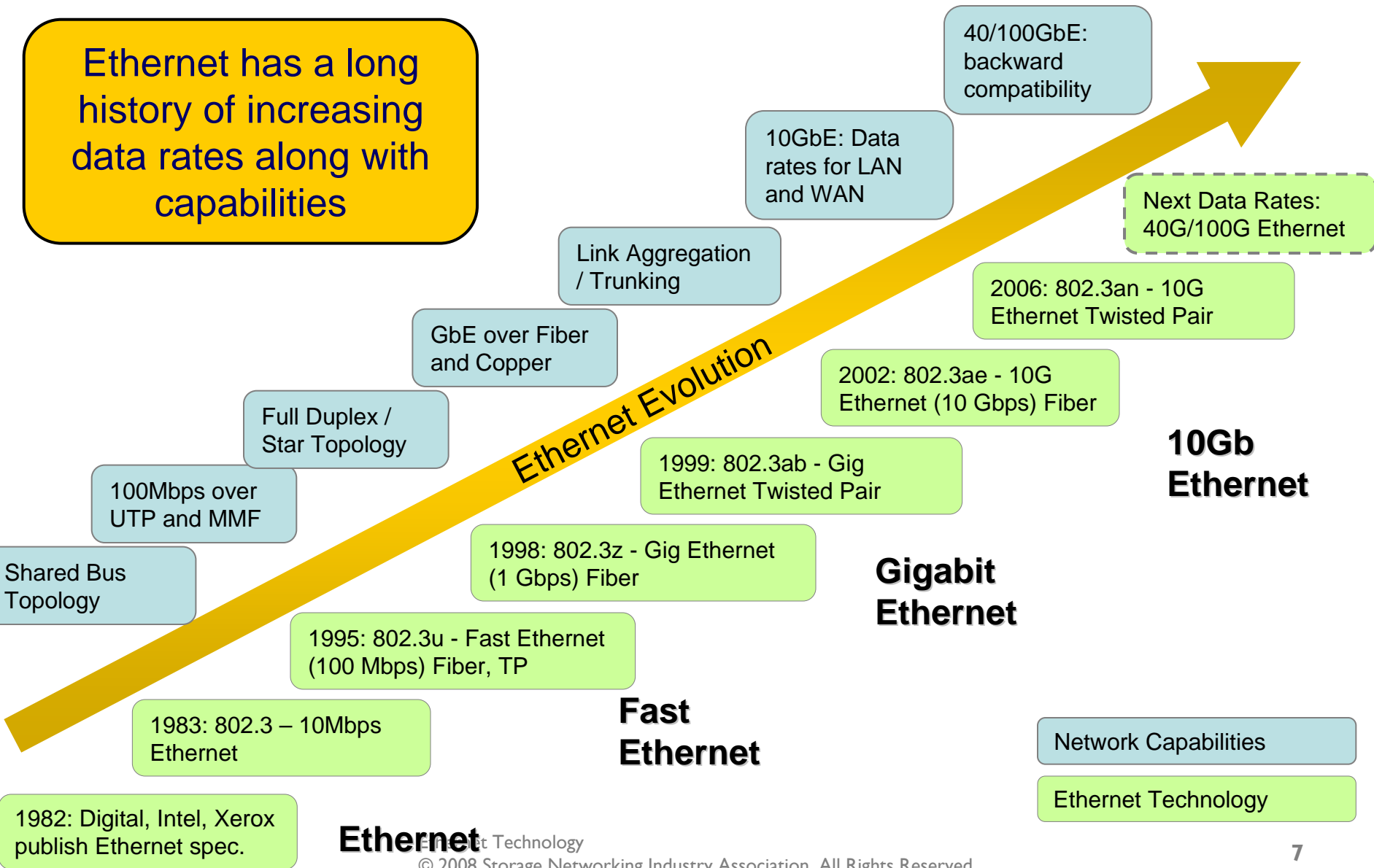
## Data Centers and Enterprise

## Research, Education and Government

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

# Ethernet Evolution

Ethernet has a long history of increasing data rates along with capabilities



# Network Topologies

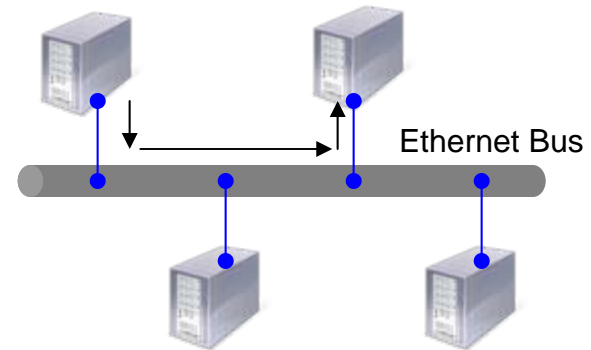
## Original Ethernet – Shared Bus Topology

- Single Coaxial bus
- Half duplex operation
- Media contention managed using CSMA/CD protocol
- Low utilization

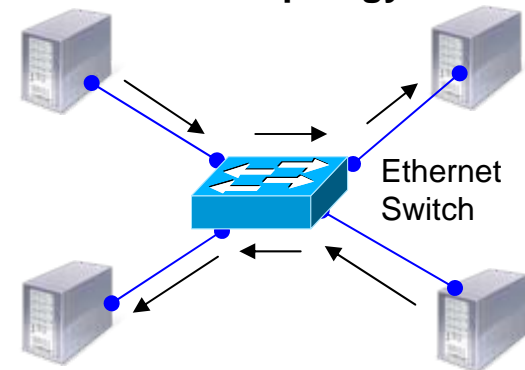
## Today: Star Topology

- Point-to-Point connections
- Full Duplex operation
- No Media contention
- Higher Aggregate bandwidth

**Shared Bus Topology**



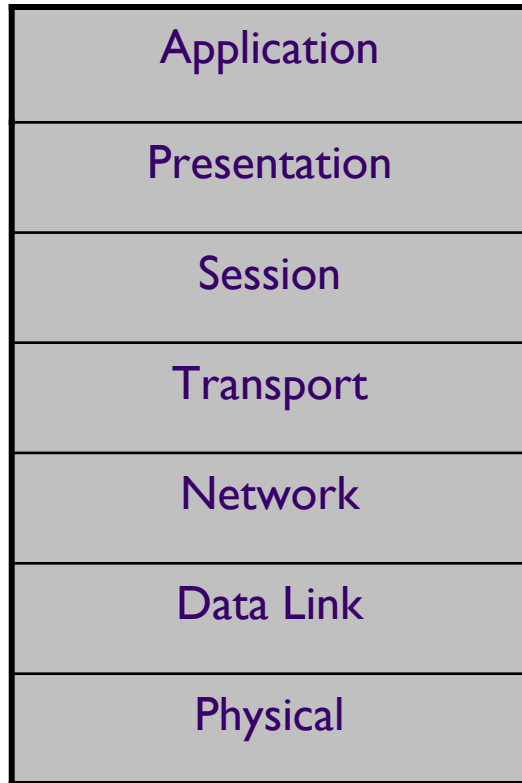
**Star Topology**



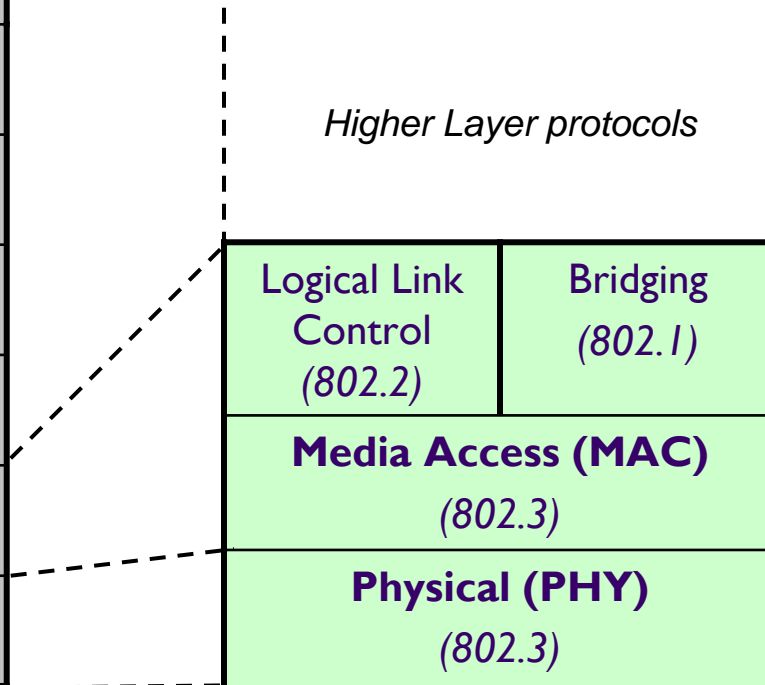
**Modern implementations of Ethernet are all Switched**

# OSI Reference Model

OSI Reference Model



IEEE 802.3 Reference Model



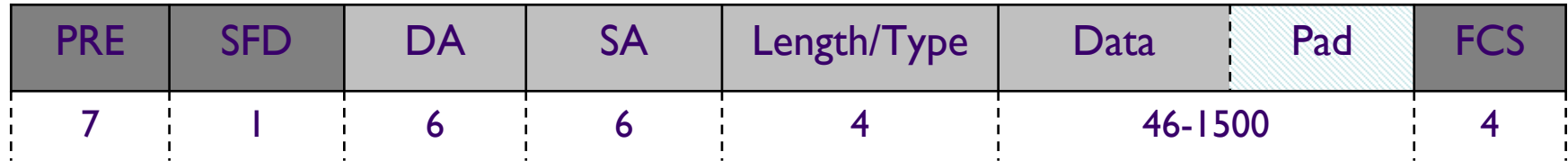
**MAC Function:**

- Data Encapsulation
- Error Detection
- Initiating frame transmission

**PHY Function:**

- Encoding
- Mux / Demux
- Signal Transmitters
- Auto-negotiation

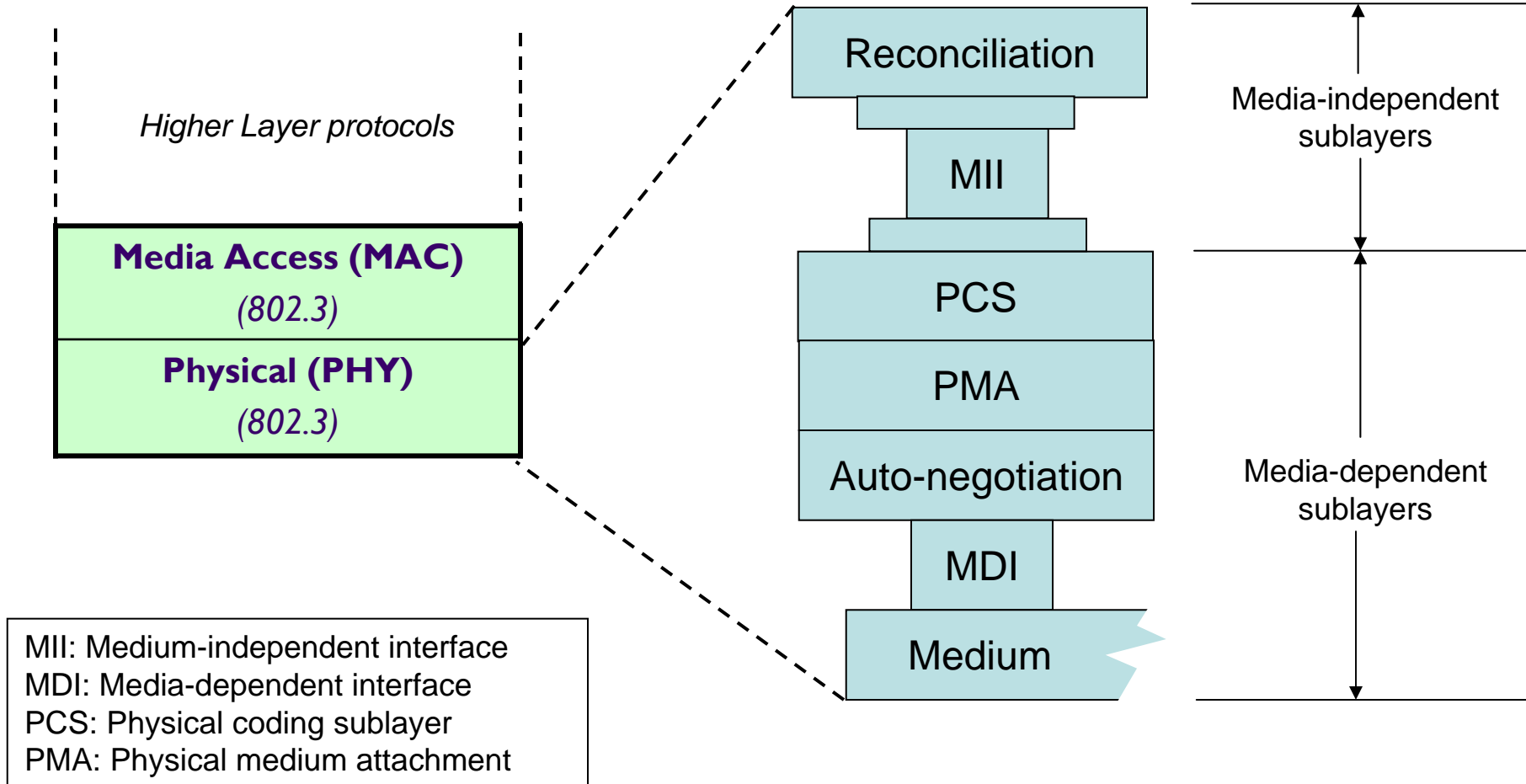
# Ethernet Frame Format



- **Ethernet address: Source / Destination Address**
  - ◆ Unique and controlled by IEEE
  - ◆ High order address bit is 1 for multicast and broadcast
  - ◆ A destination address of only 1s is accepted by all stations
- **Length / Type**
  - ◆ IEEE 802.3 / Ethernet V2. Majority implementations are V2
  - ◆ Field value less than 1500 indicates bytes in data field
  - ◆ Field value greater than 1536 indicates type of frame
- **Data**
  - ◆ Filler or padding added if data is less than 46 bytes
- **FCS**
  - ◆ 32-bit CRC created by source MAC

# Ethernet PHY

## IEEE 802.3 Reference Model



## ➤ Overview of Ethernet Technology

- ◆ Ethernet Evolution
- ◆ Frame Format

## ➤ 10 Gigabit Ethernet Technology

- ◆ Demand for 10GbE
- ◆ 10GbE PHY technology

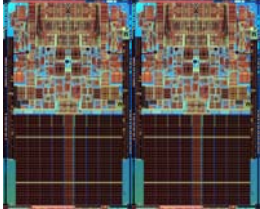
## ➤ I/O Consolidation over Ethernet

- ◆ Ethernet enhancements for “lossless” fabric

## ➤ Next Generation Ethernet

- ◆ 40G / 100G

# 10 Gigabit Ethernet demand



Multi-Core CPU architecture allowing faster execution of multiple applications on the same processor



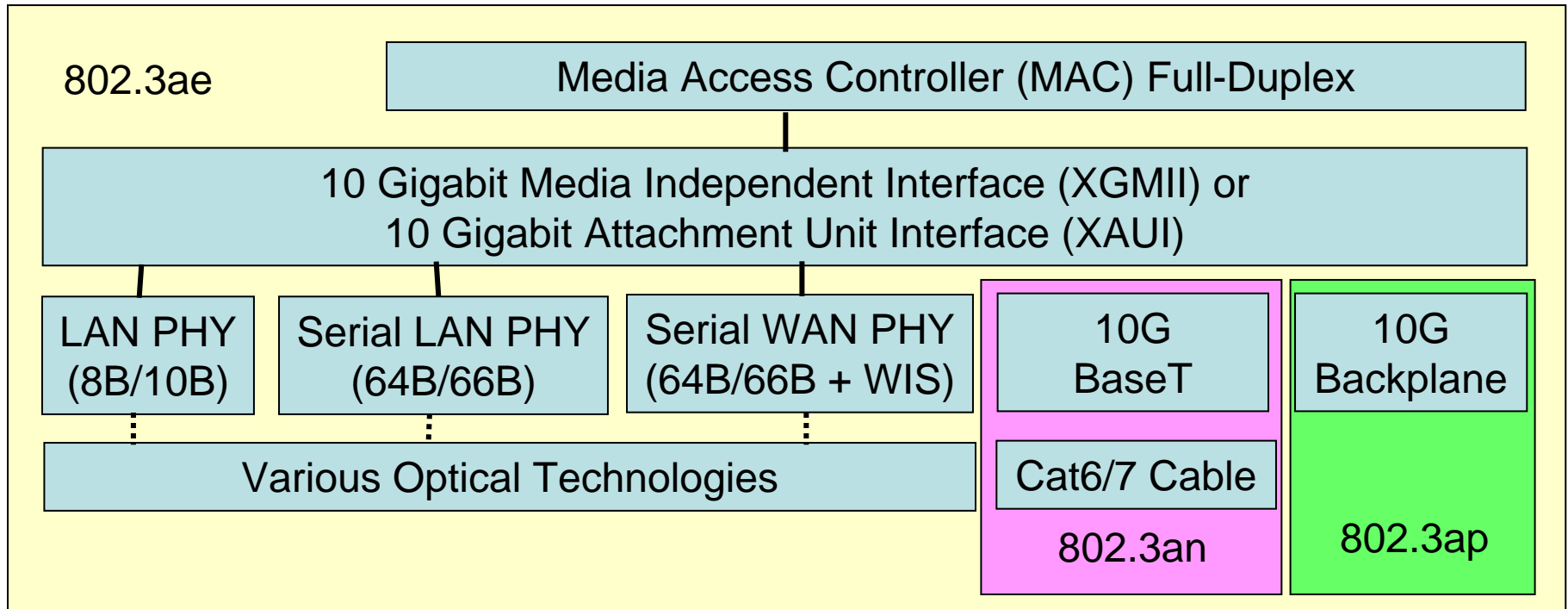
Growing need for storage network is driving the demand for higher bandwidth network



Consolidation of multiple virtual servers on a physical servers demands more network bandwidth per server

**Server Virtualization and Network Storage are driving the need for higher bandwidth network connections**

# 10GbE Standards



802.3an and 802.3ap make 10GbE even more compelling for data center applications

# Changes with 10GbE

## ➤ 802.3ae defines 10GbE

- ◆ 10Gbps Data Rates
- ◆ Full-duplex only; no more carrier-sensing multi-access / collision detection (CSMA/CD)
- ◆ Optical Physical Layer
  - LAN PHY and WAN PHY options
    - WAN PHY compatible with SONET

## ➤ 802.3an adds twisted pair cabling

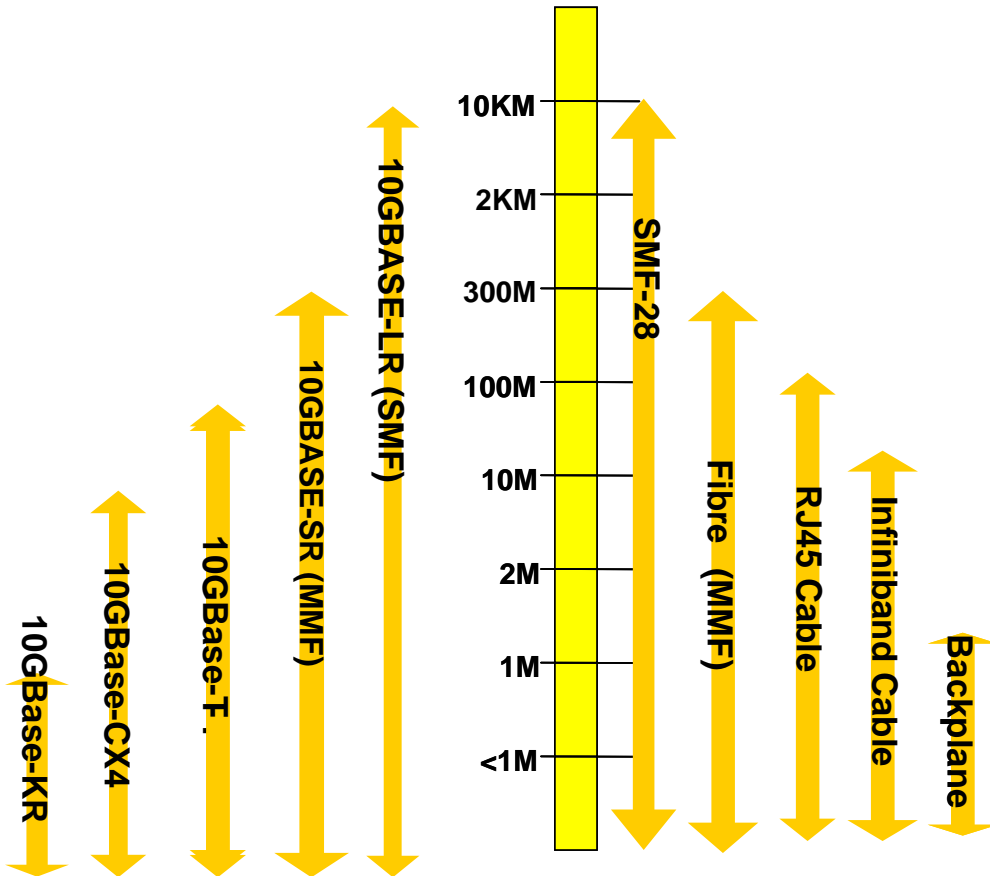
- ◆ Cat 6 and Cat 7

## ➤ 802.3ap adds backplane specifications

- ◆ Blade servers and communications equipment

# 10Gig PHY Technology

Technology	Length	Media
------------	--------	-------



➤ Several options targeted for specific application, span length, and media

➤ **Fibre (Optical)**

- ◆ Single Mode Fibre
- ◆ Multi-mode Fibre
- ◆ Many choices

➤ **Copper**

- ◆ 10G Base-T
- ◆ Serial 10G Copper Cable

➤ **Backplane**

- ◆ XAUI (4 x 3.125)
- ◆ 10GBase-KX4
- ◆ 10G Base-KR

## ➤ Overview of Ethernet Technology

- ◆ Ethernet Evolution
- ◆ Frame Format

## ➤ 10 Gigabit Ethernet Technology

- ◆ Demand for 10GbE
- ◆ 10GbE PHY technology

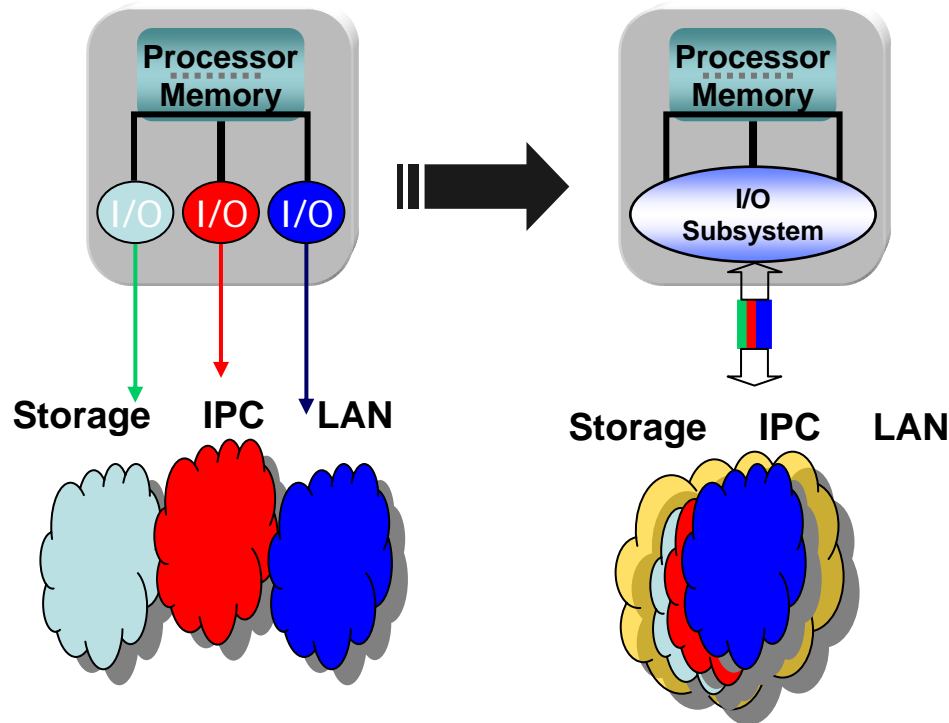
## ➤ I/O Consolidation over Ethernet

- ◆ Ethernet enhancements for “lossless” fabric

## ➤ Next Generation Ethernet

- ◆ 40G / 100G

# I/O Consolidation over Ethernet



- Reduced CapEx: Fewer ports from server, fewer fabrics, fewer cables
- Reduced OpEx: Simplified management, less power/thermal, reduced installation costs
- I/O Consolidation for Storage:
  - ◆ iSCSI and FCoE
- 10GbE drives I/O convergence

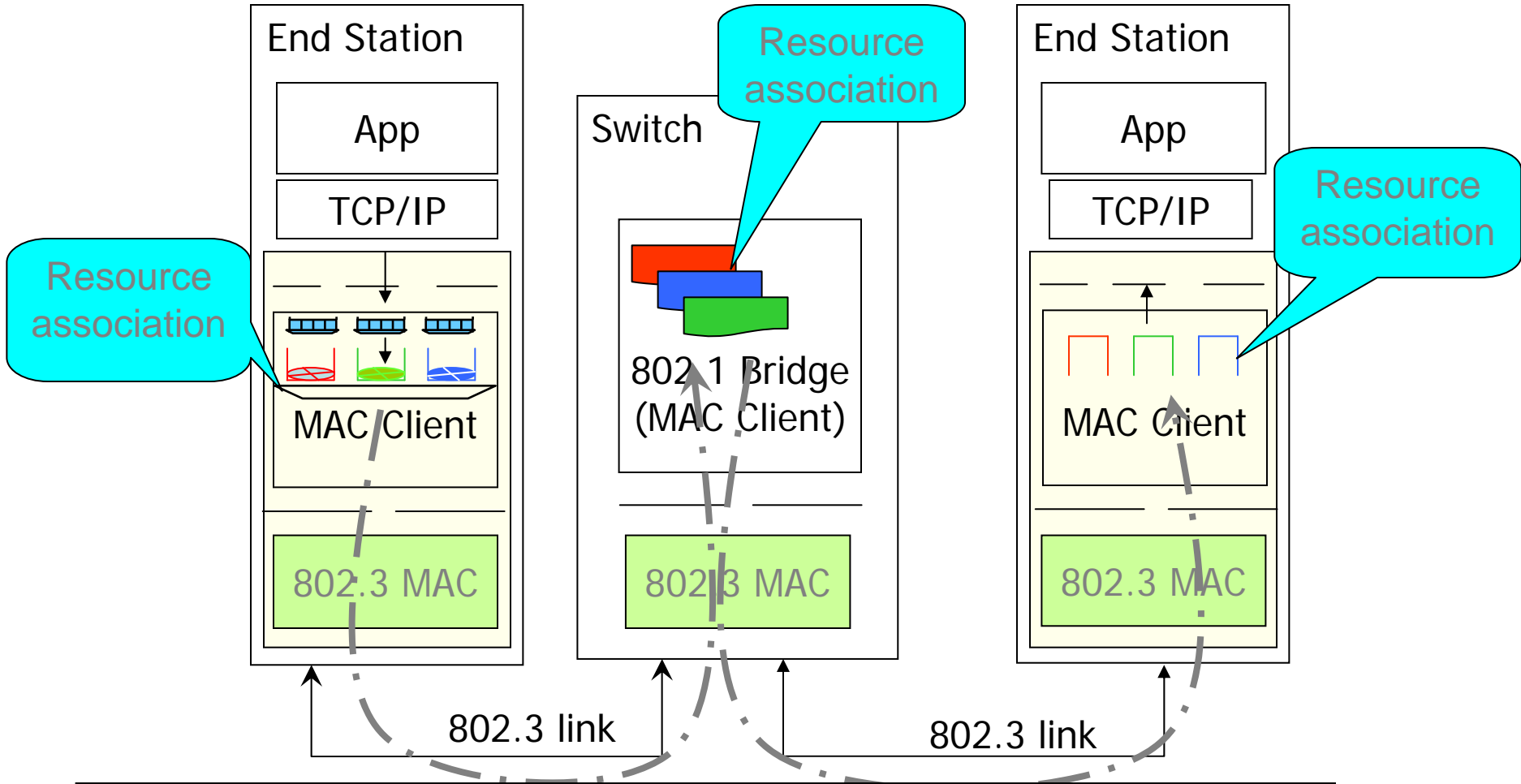
IO consolidation delivers lower TCO

# Ethernet Enhancements for Data Center

- **Traffic Differentiation: Priority Groups**
  - ◆ 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 - Backward Congestion Notification
  
- **Bi-sectional Bandwidth: Shortest-Path Bridging**
  - ◆ Allow L2-Multipathing within Data Center

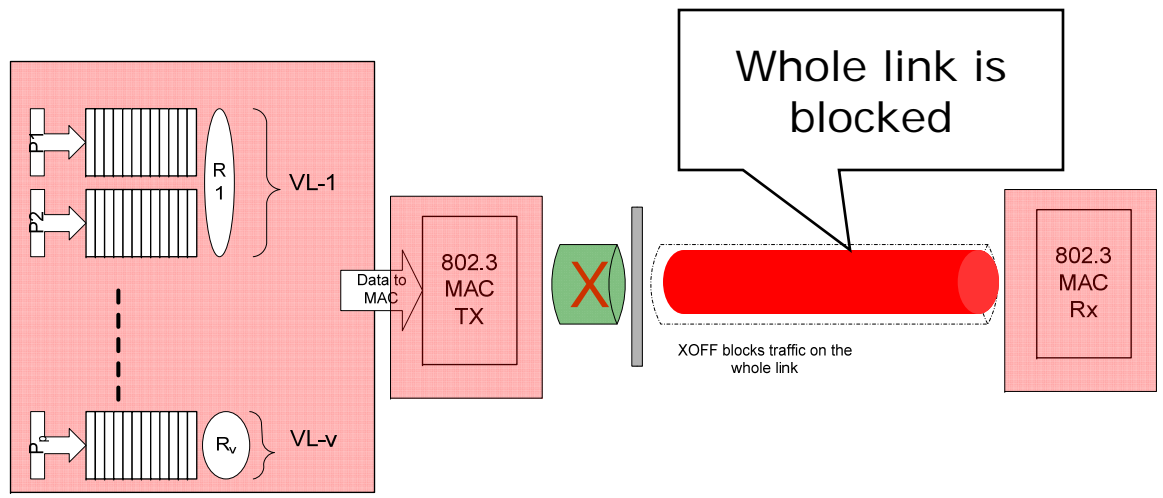
Moving Ethernet from Best Effort to a deterministic fabric

# Priority Groups (IEEE 802.1Qaz)

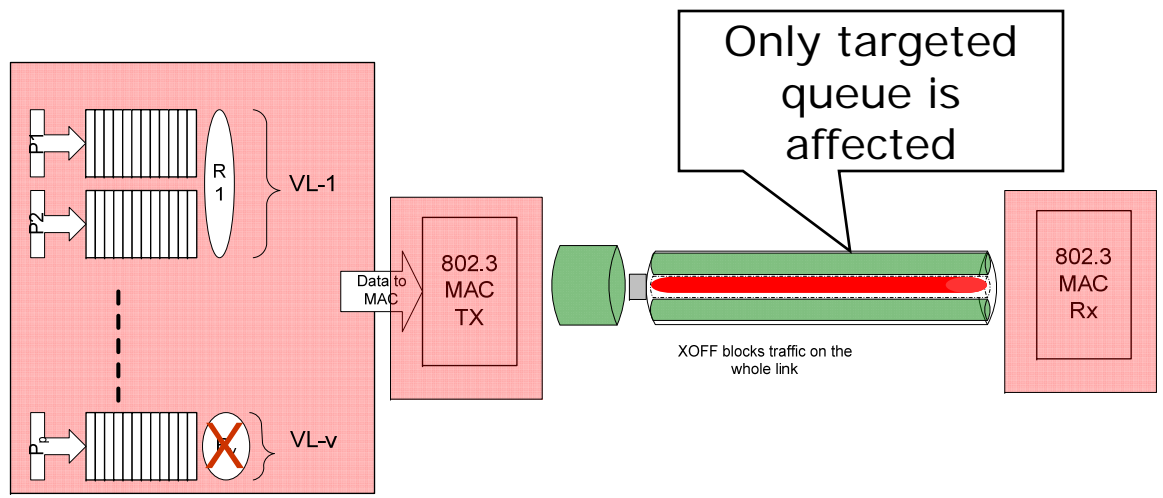


PG's allow latency optimization for one application while allowing throughput optimization for other application

# Priority-based Flow Control

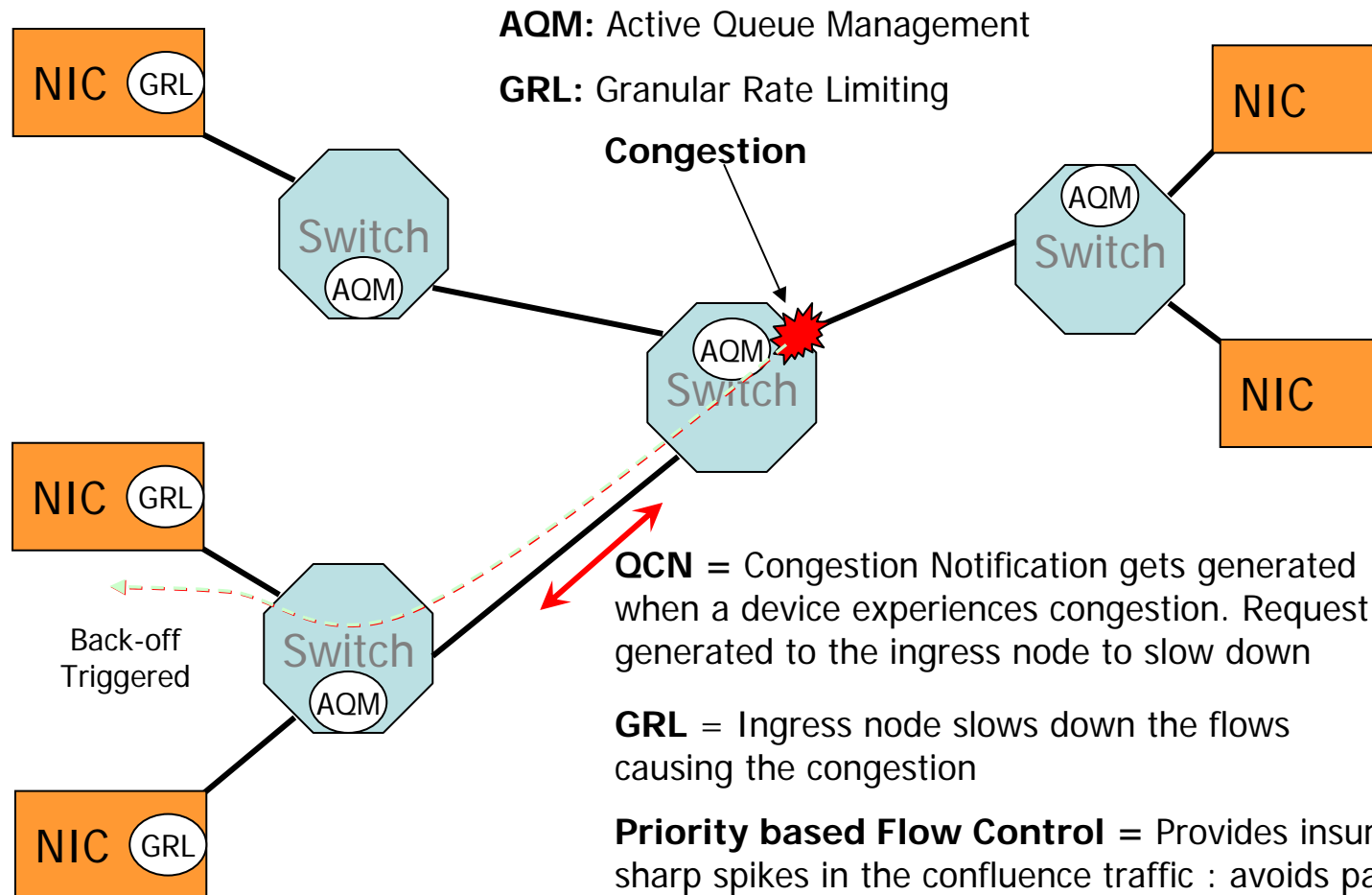


Link Pause

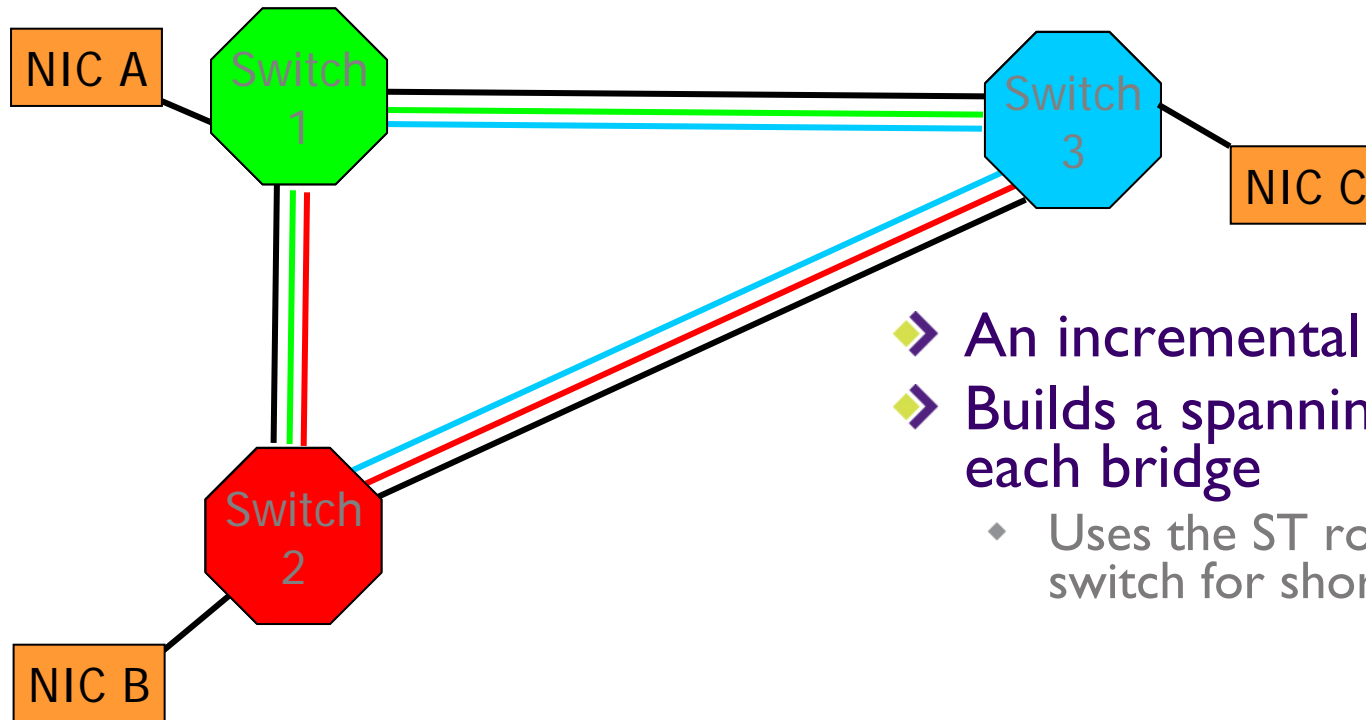


Priority Pause

# Congestion Notification (IEEE 802.1Qau)



# Shortest Path Bridging (IEEE 802.1Qaq)

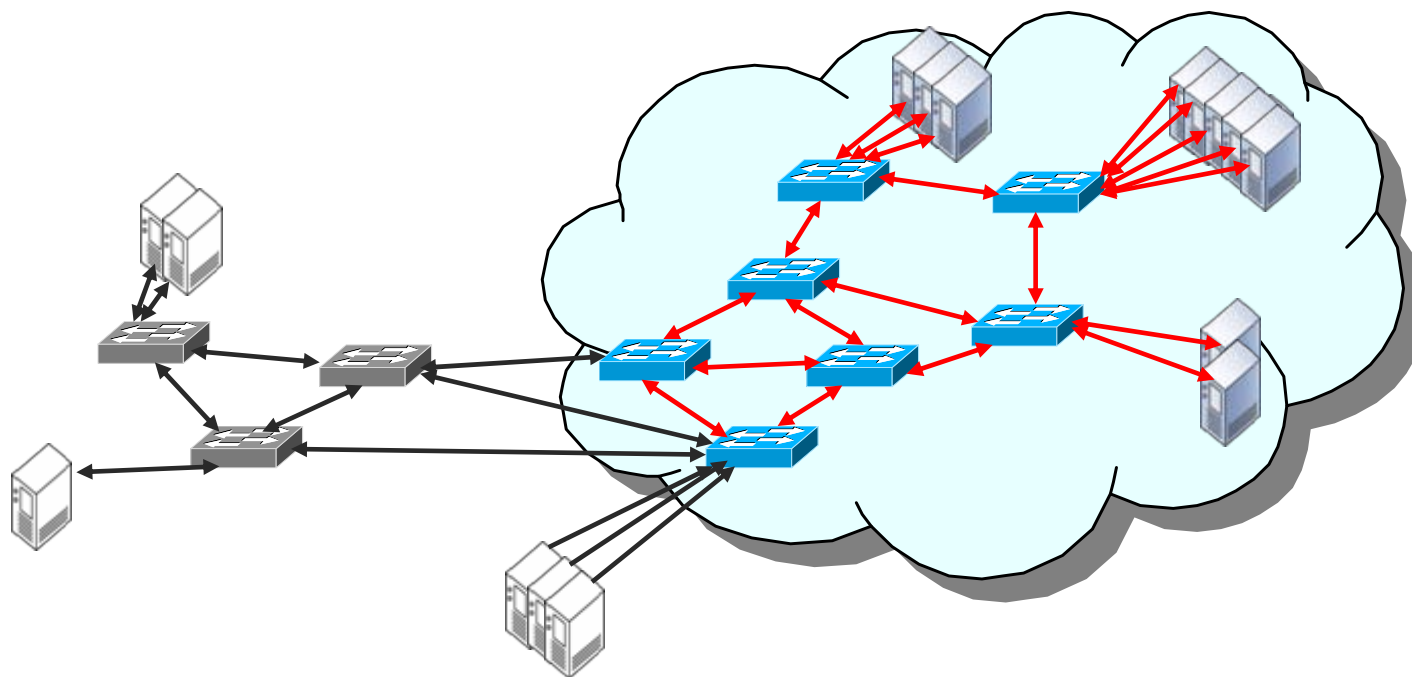


- An incremental advance to MSTP
- Builds a spanning tree (ST) for each bridge
  - ◆ Uses the ST rooted at the source switch for shortest path bridging

## ➤ Ensures forward and reverse paths are aligned

- ◆ Reflection Vector and other ideas being investigated to “align” spanning trees

# Gluing it all together



- Capability Exchange Protocol allows discovery of compliant devices, capabilities
- Allows formation of cloud of compliant devices
- Allows incremental deployment – rack at a time

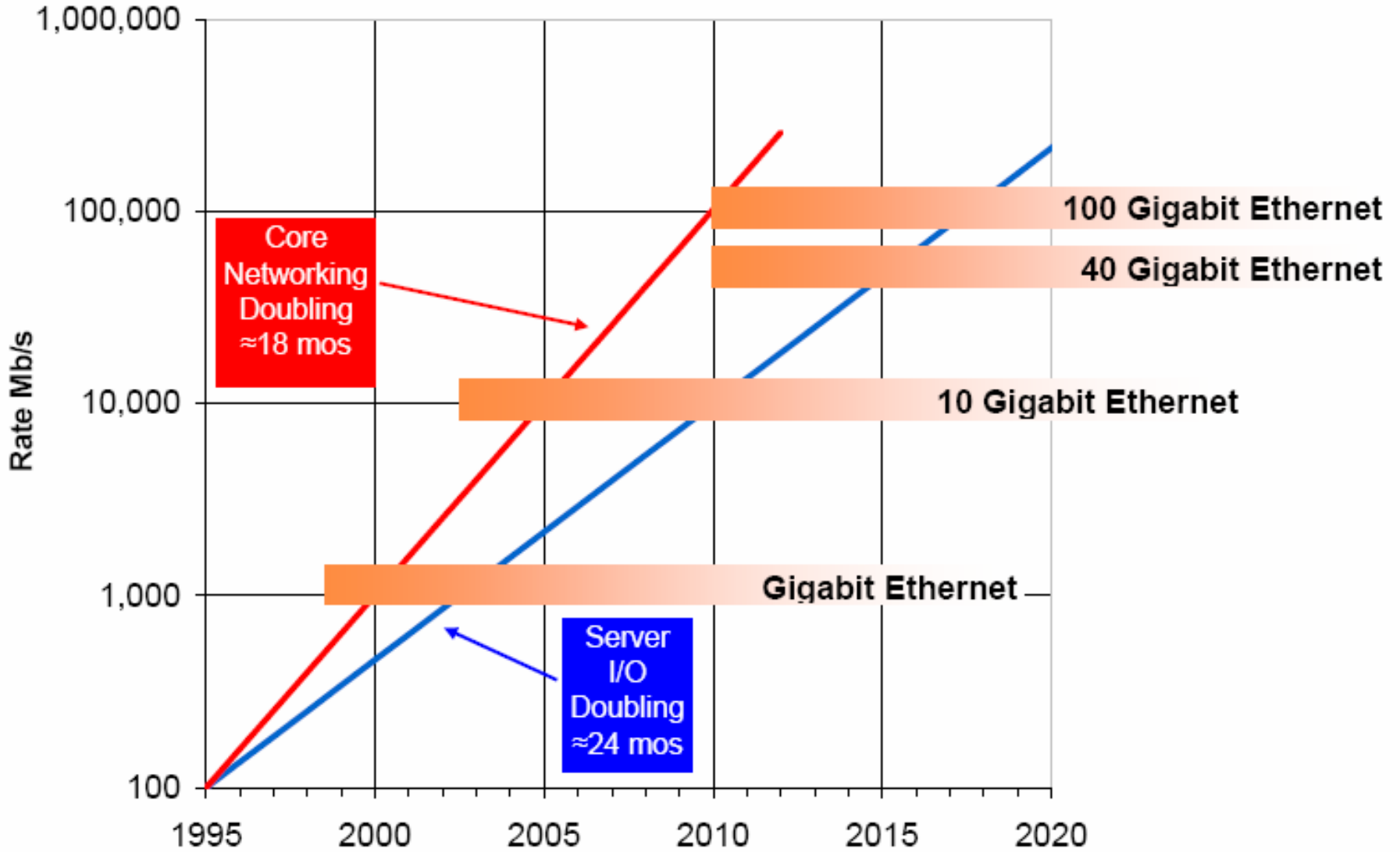
# IEEE Enhancements for Data Center

- 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 IO Consolidation in Data Center
  - ◆ Congestion Notification: Approved project IEEE 802.1Qau
  - ◆ Shortest Path Bridging: Approved project IEEE 802.1Qaq
  - ◆ Virtual Links (Priority Groups): Approved project IEEE 802.1Qaz
  - ◆ Priority based Flow Control: In consideration in IEEE 802.1
  - ◆ DCB Capability Exchange Protocol: Part of various projects above
- DCB Standards trending for ratification in ~2009

# Agenda

- Overview of Ethernet Technology
  - ◆ Ethernet Evolution
  - ◆ Frame Format
  
- 10 Gigabit Ethernet Technology
  - ◆ Demand for 10GbE
  - ◆ 10GbE PHY technology
  
- I/O Consolidation over Ethernet
  - ◆ Ethernet enhancements for “lossless” fabric
  
- Next Generation Ethernet
  - ◆ 40G / 100G

# 40/100 GbE Networking



# Ethernet Evolution continues ...

- Higher Speed study group objectives
  - ◆ Full duplex operation only
  - ◆ Preserve 802.3 frame format and size
  - ◆ Support BER better than or equal to 10<sup>-12</sup>
  - ◆ Support 40 Gb/s data rate (different cable types)
  - ◆ Support 100 Gb/s data rate (different cable types)
- Target completion - 2010

- 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.**

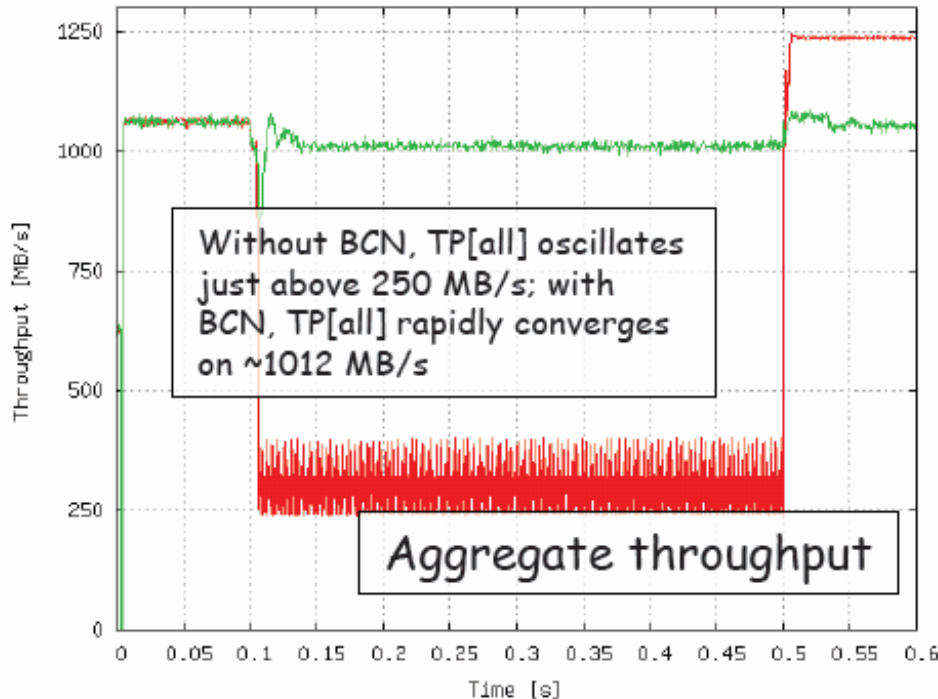
**- SNIA Education Committee**

**Sunil Ahluwalia (Author)**  
**Manoj Wadekar (Author)**

**Howard Goldstein (Tutorial Review)**  
**Walter Dey (Tutorial Review)**

# Backup

# Congestion Management in DCB



Ref: IBM presentation in IEEE 2007

- Attached simulation shows how DCB congestion management (QCN+PAUSE) improves aggregate throughput in the network
- Red line shows throughput collapse without DCB
- Green line shows throughput with DCB

## ➤ Carrier Sense Multiple Access / Collision Detect

sis – muh, see - dee

- Listen for the media to be quiet
- If quiet, you can transmit
- Listen while you transmit to make sure it is just what you are sending
- If not, then someone else is sending at the same time –  
**COLLISION**
- Backoff and repeat process



# New MSA for Next Generation 10G Optical Modules SFP+

- The SFP+ MSA has been kicked off to develop a low cost and low power 10G serial optical interface and leverages existing SFP standard
  - ◆ The SFP+ MSA Rev 1.0 released next week and expect stability by next Rev.
  - ◆ ~30 companies actively participating in this effort
  - ◆ ~65 companies monitoring activities
- Key difference versus existing XFP optical modules
  - ◆ SFP+ removes a 10G retiming stage (saving size, cost, and power)
  - ◆ SFP+ is a smaller form factor (~46% Area Savings)
  - ◆ SFP+ Power target for up to 10km optical links (>800mW)
- Mechanical Specification – SFF-8432 improved but compatible with existing SFP

# Quad 2x4 SFP+ Concept

➤ The demand for XFP and SFP+ are for port density and lower cost.

◆ Rack spacing and port density equates to real states

Screw attach to faceplate

- establishes consistent compression of EMI gasket between housing & face plate of system
- Prevents oil-canning of face plate

