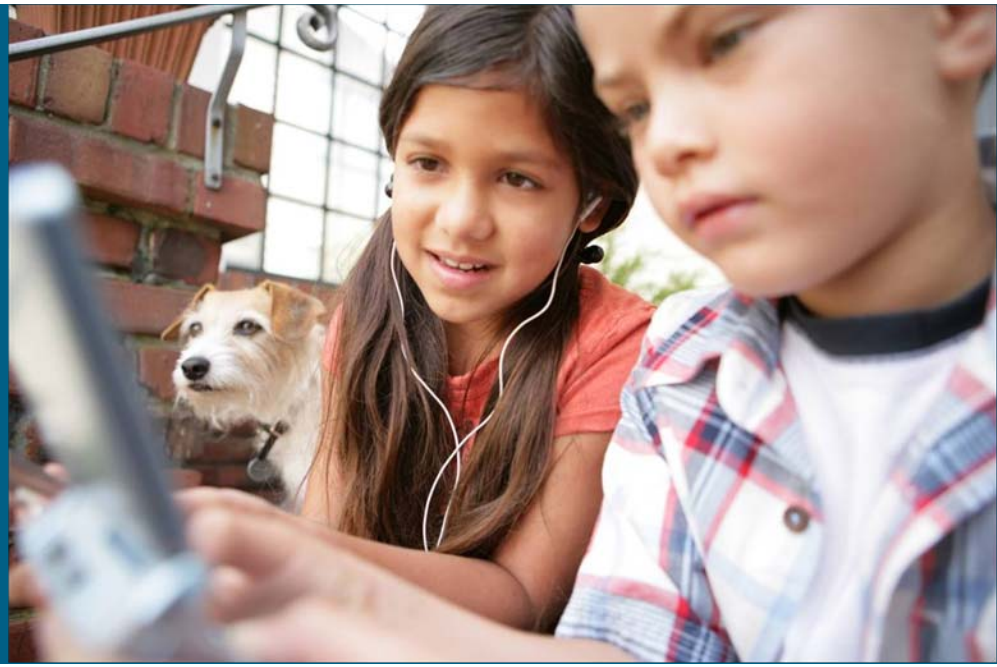
Data Center Ethernet
The Framework for I/O Consolidation and Unified Fabric

Bjørn R. Martinussen
brm@cisco.com
Consulting Systems Engineer
Data Centre Technical Marketing

I/O Consolidation, Unified Fabric



What is Server I/O Consolidation

- IT Organizations operate multiple parallel networks

  IP Applications (including NFS, NAS,…) over a Ethernet network

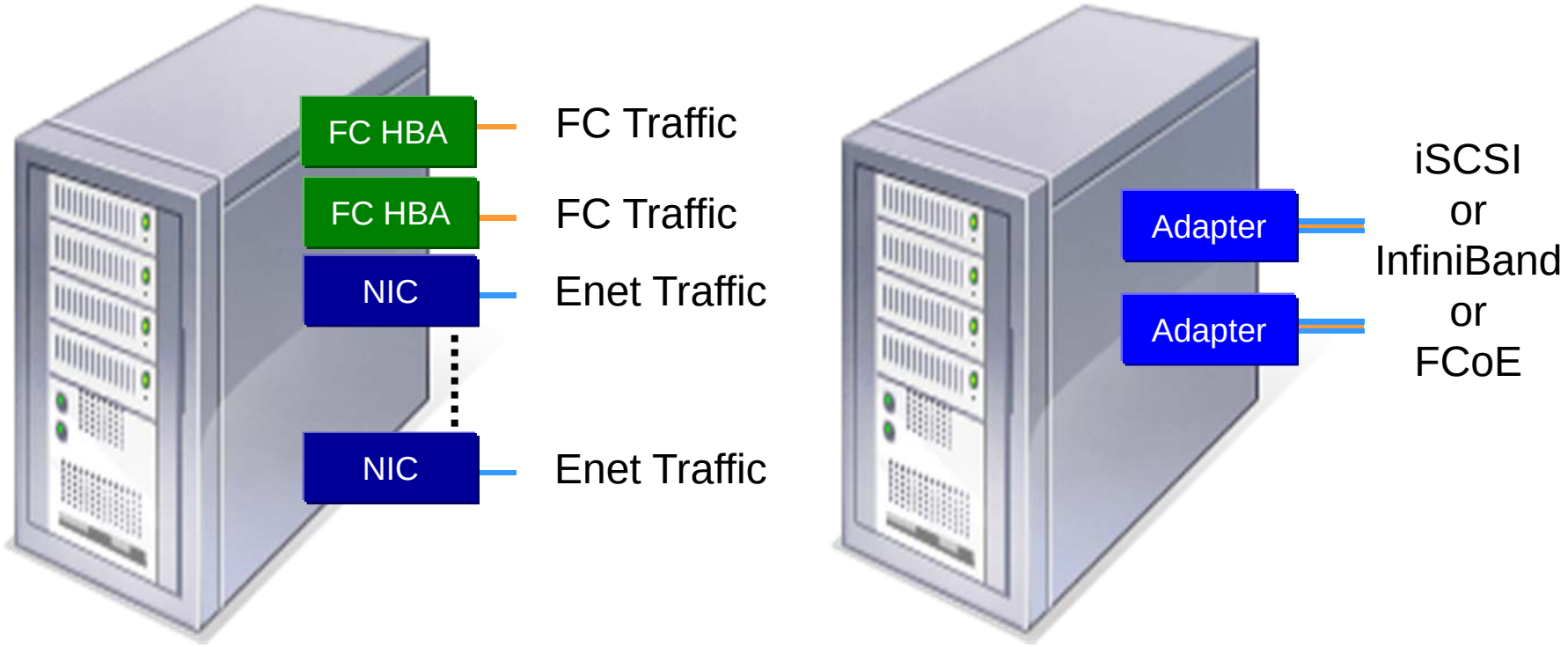
  SAN over a Fibre Channel network

  HPC/IPC over an InfiniBand network *)

- Server I/O consolidation combines the various traffic types onto a single interface and single cable
- Server I/O consolidation is the first phase for a Unified Fabric (single network)

***) for lowest latency requirements, InfiniBand is the best and most appropriate technology**

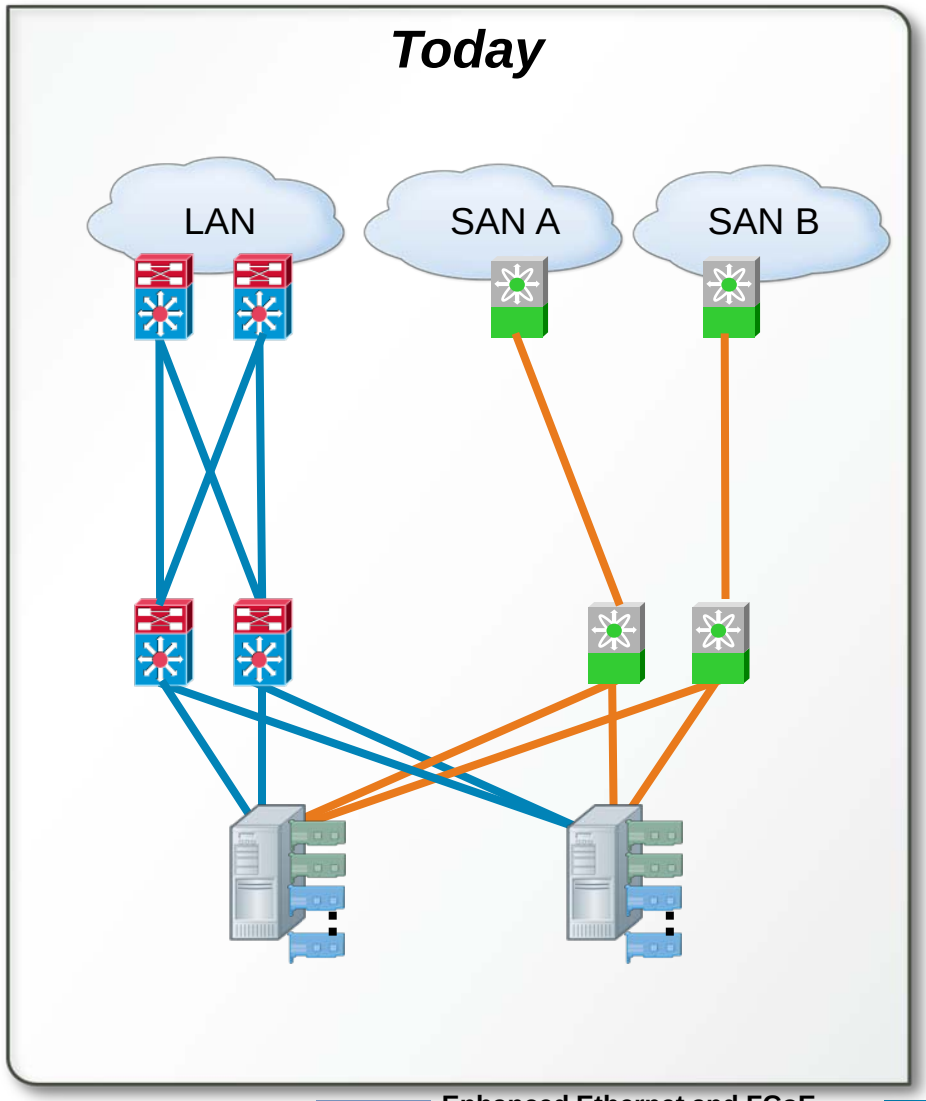
I/O Consolidation Benefits



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)

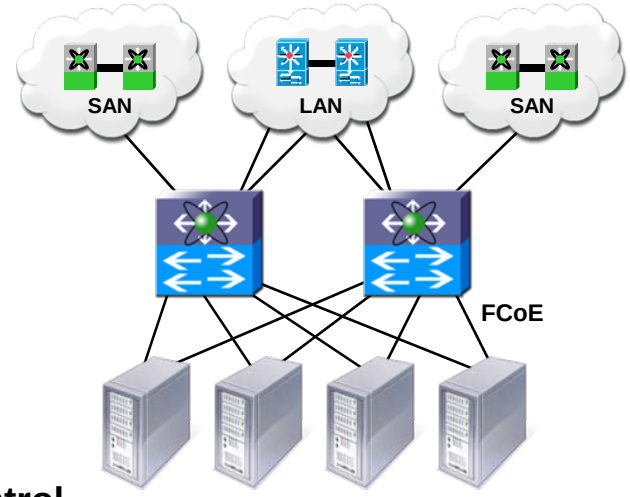
I/O Consolidation



- **Today:**
  - **Parallel LAN/SAN Infrastructure**
  - **Inefficient use of Network Infrastructure**
  - **5+ connections per server – higher adapter and cabling costs**
    - **Adds downstream port costs;**
      cap-ex and op-ex
    - **Each connection adds additional points of failure in the fabric**
  - **Longer lead time for server provisioning**
  - **Multiple fault domains – complex diagnostics**
  - **Management complexity**

Fibre Channel Over Ethernet

- Reduce cost of Fibre Channel transport
  - Leverage NICs, switches, cables, etc
  - Integrate into Fibre Channel infrastructure
    - FCoE appears as FC to the host and the network
    - Preserves current FC infrastructure and management
    - FC frame is unchanged
- Active in T11 FC-BB-5
  - Functional model and frame format approved
- Initial adoption “first 30 meters” from end point
  - End point directly connected to FCoE switch
    - Expected to be most common initial deployment
  - End point connected to FCoE switch through an Ethernet access switch
- Enablers
  - Switch support for 2.5K frames
  - PCI-Express
  - 10G Ethernet
  - Enhancements for Standard Ethernet, eg Priority Flow Control



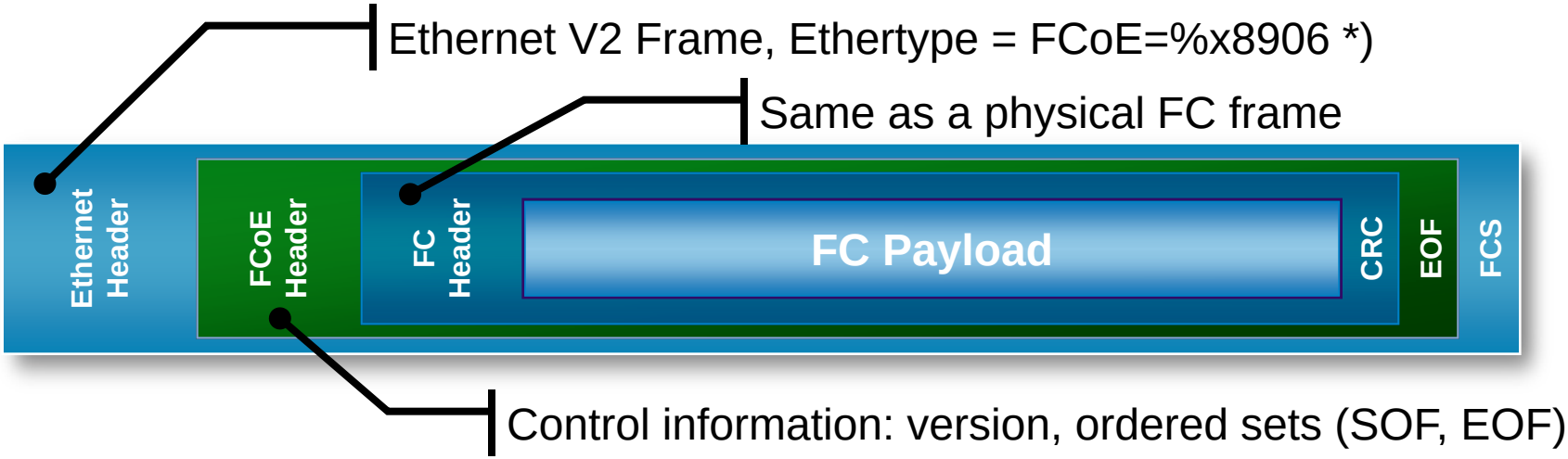
FCoE Enablers

- 10Gbps Ethernet
- PCI-Express
- Lossless Ethernet (Data Center Ethernet)

  Matches the B2B credits used in Fibrechannel to provide a lossless service

- Ethernet jumbo frames (2180 Bytes)

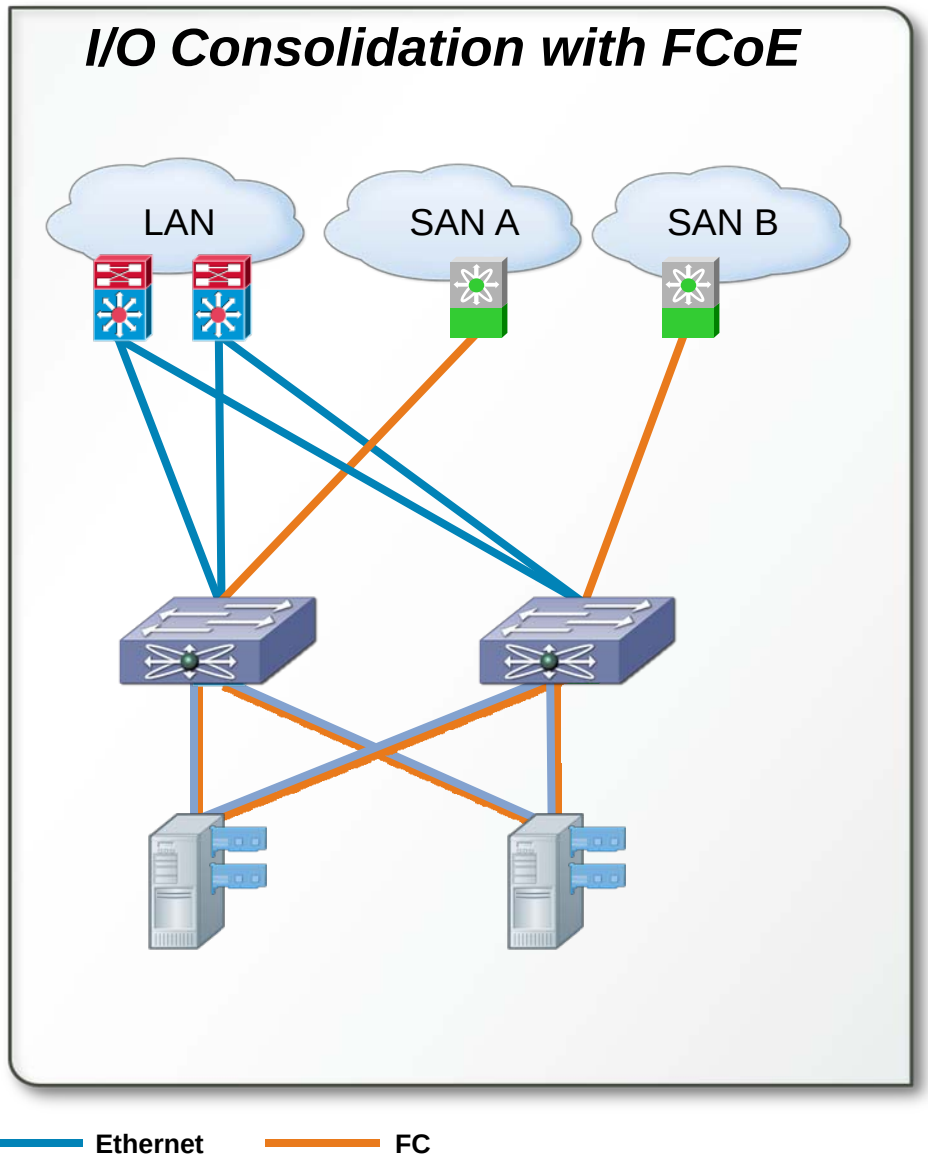
  Max FC frame payload = 2112 bytes



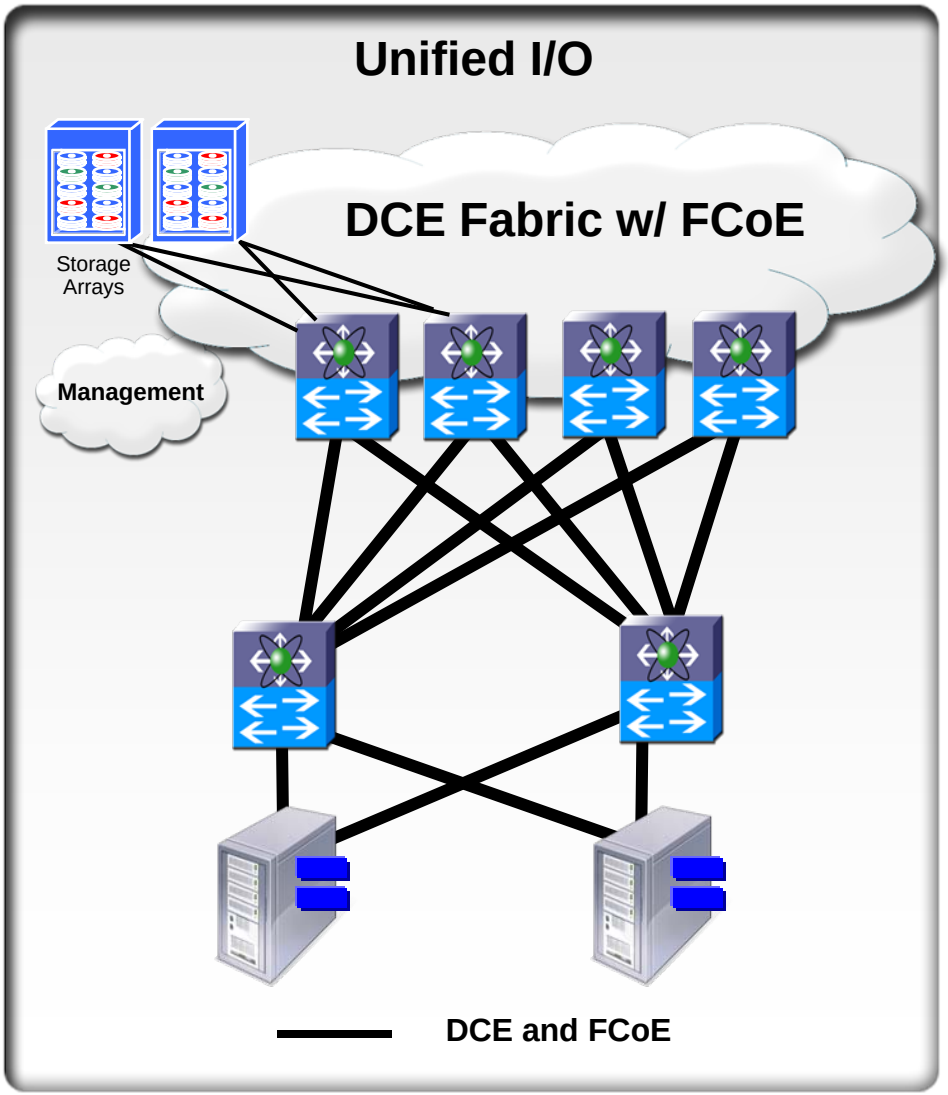
*) FIP (FCoE Initializing Protocol) uses Ethertype %x8914

I/O Consolidation

- **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**



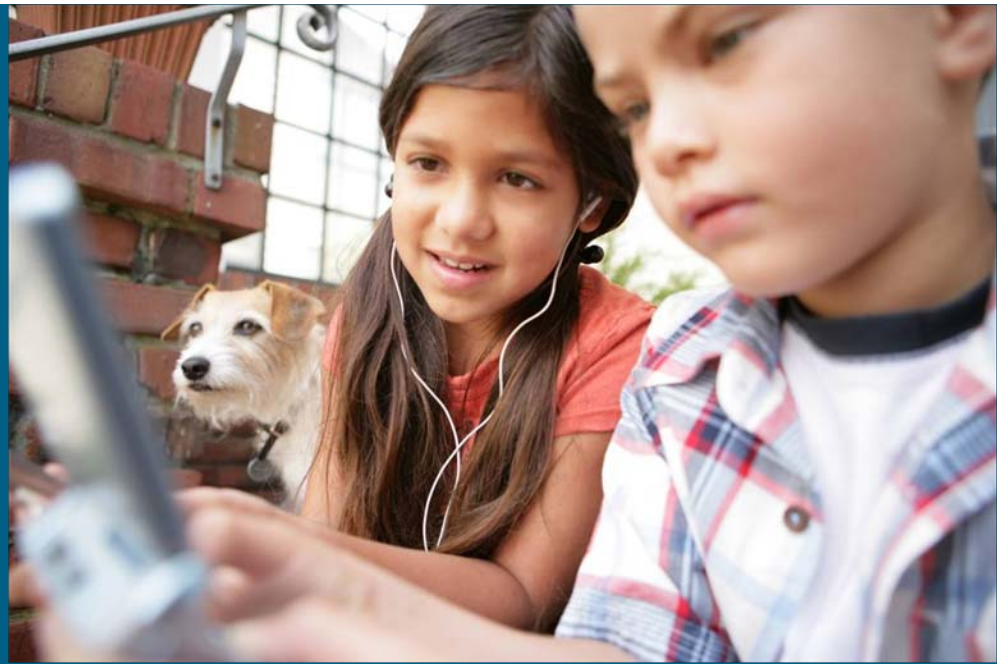
Unified Fabric



Unified I/O

- Datacenter wide Unified Fabric for LAN and SAN
- L2/L3 Multipathing end to end
- Consistent network policies across datacenter
- Lower TCO

Data Centre Ethernet



What is Data Center Ethernet (DCE) ?

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

Sometimes also called

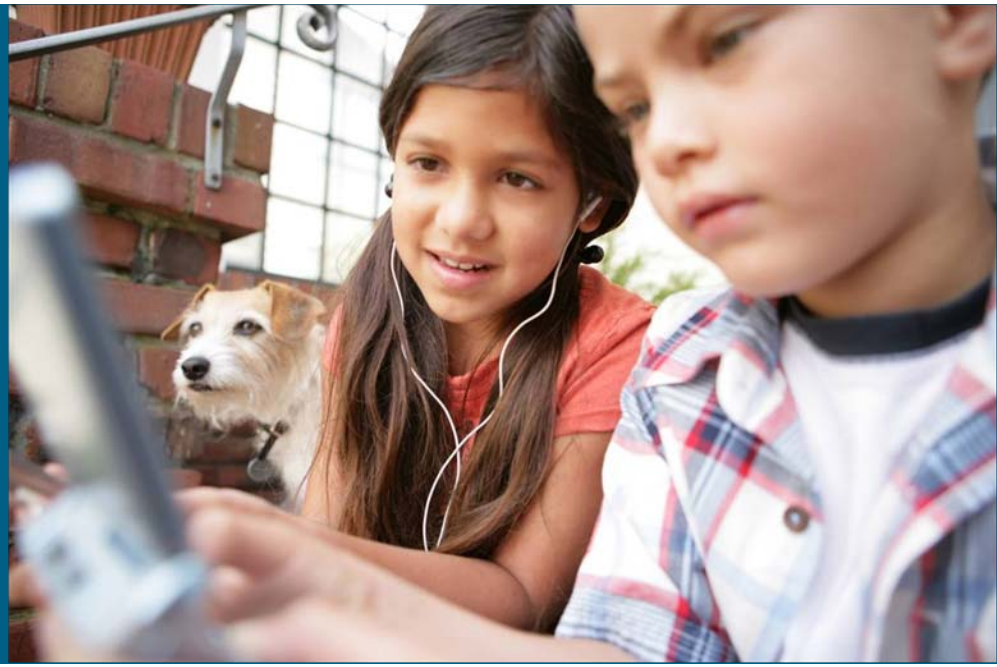
CEE = Converged Enhanced Ethernet

DCB = Data Center Bridging (IEEE)

Data Center Ethernet Features
Overview

Feature	Benefit
Priority-based Flow Control (PFC)	Provides class of service flow control. Ability to support storage traffic
CoS Based BW Management	Grouping classes of traffic into “Service Lanes” IEEE 802.1Qaz, CoS based Enhanced Transmission
Congestion Notification (BCN/QCN)	End to End Congestion Management for L2 network
Data Center Bridging Exchange	Auto-negotiation for Enhanced Ethernet capabilities DCBX (Switch to NIC)
L2 Multi-path for Unicast & Multicast	Eliminate Spanning Tree for L2 topologies Utilize full Bi-Sectional bandwidth with ECMP
Lossless Service	Provides ability to transport various traffic types (e.g. Storage, RDMA)

Priority Based Flow Control (PFC)



Overview Virtual Links

A lesson learned from Fibre Channel & InfiniBand:

- **Credited networks require smaller buffers than non-credited ones**
- **Link virtualization allows to run different types of traffic with no interference**
- **Ethernet Priority concept extended to support Virtual Links (VLs)**

  The .1Q Tag is used to identify the VL…

  … and therefore DCE supports up to 8 VLs
- **We have introduced**

  Per VL PAUSE frames for flow control

  Per VL behavior

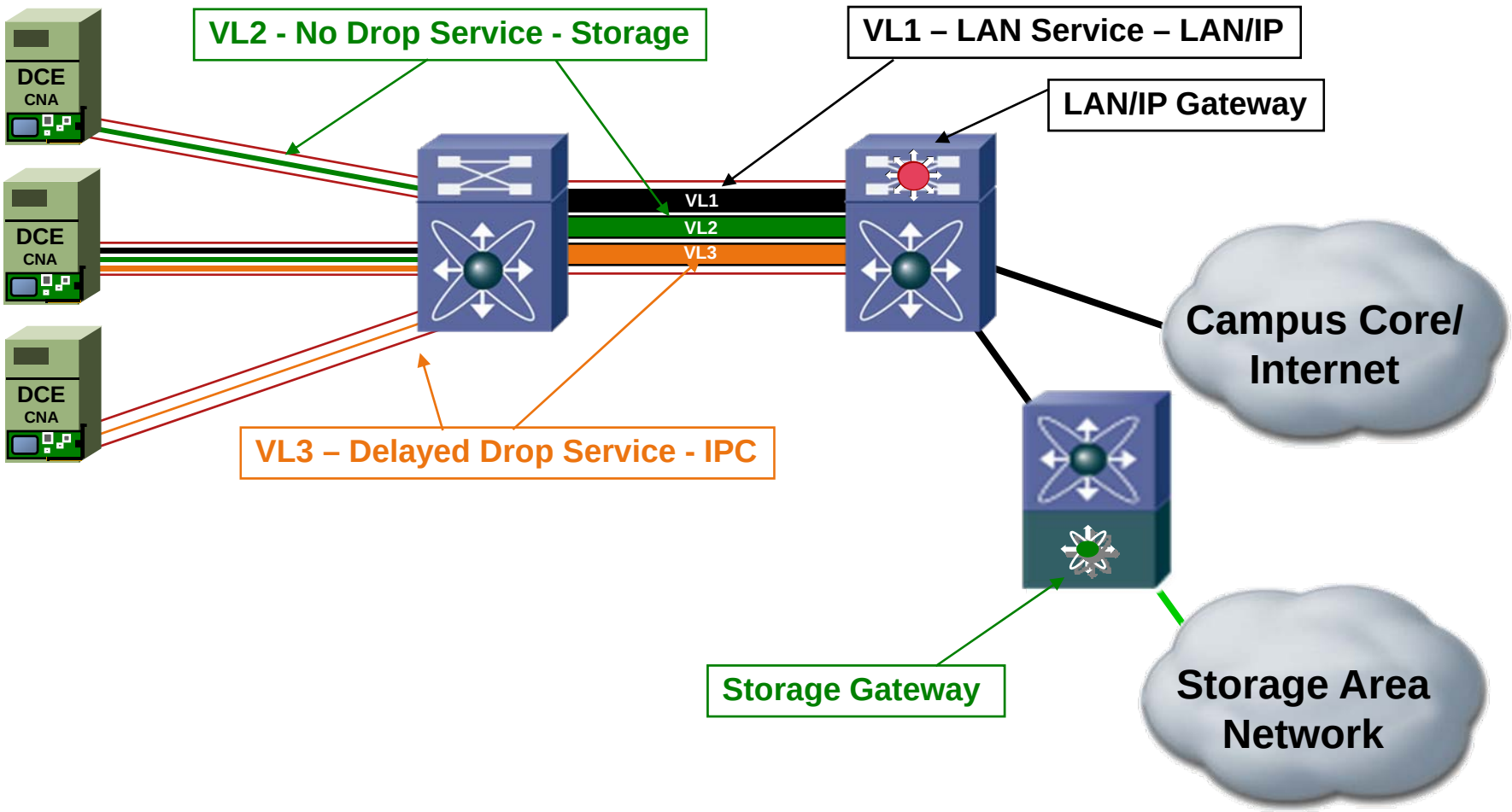
  - Buffer allocation

  - BW allocation

  - Drop type (Drop, No Drop, Delayed Drop)

Virtual Lanes/Links

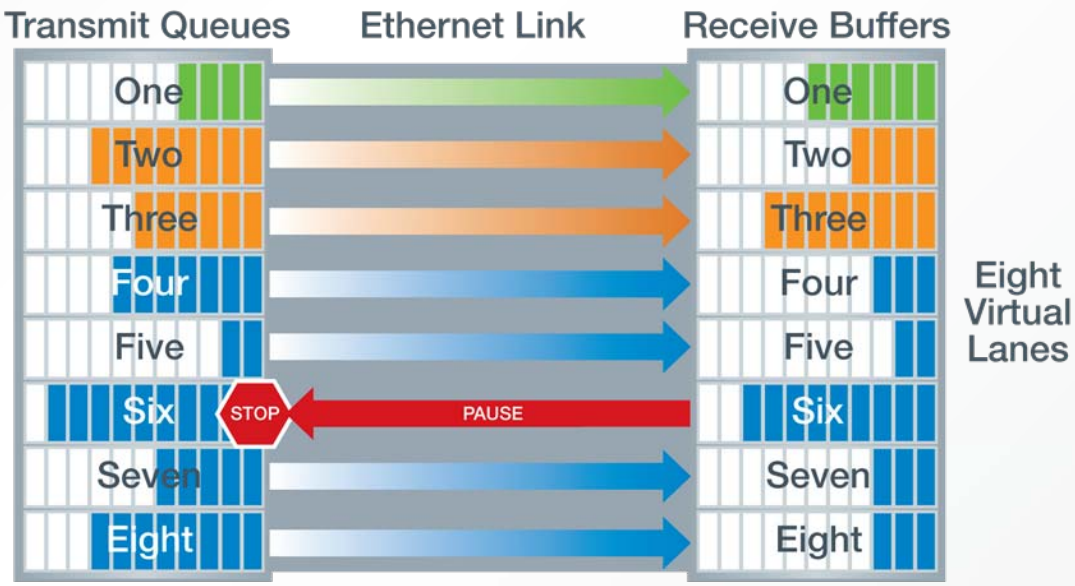
Up to 8 VL's per physical link
Ability to support QoS queues within the lanes



Drop/No Drop VL

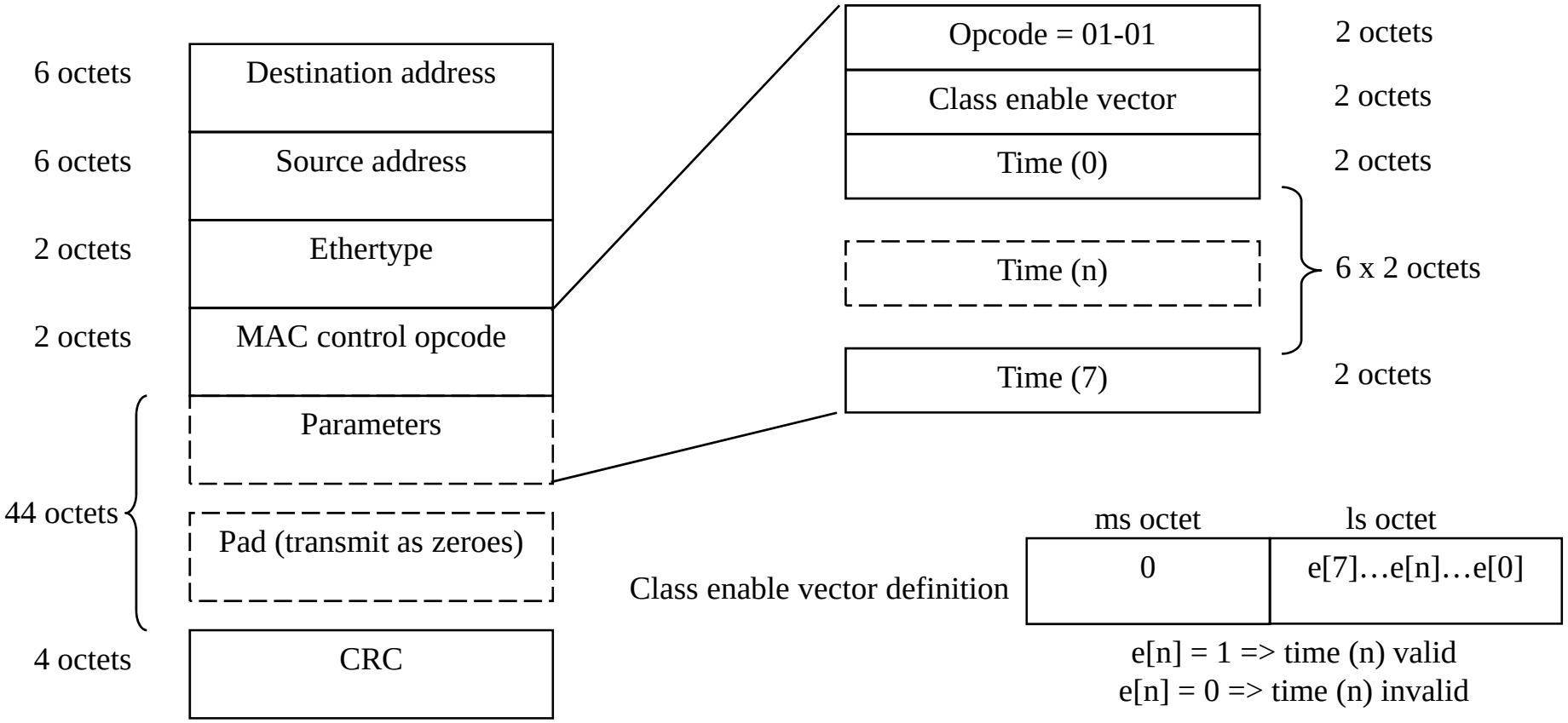
- FC provides a very close to no drop transfer
  - Drops are still possible due to topology changes or link errors
- No Drop VL provides similar characteristic as FC
- Per-priority PAUSE used to provide no-drop VL
  - Is a modification to Ethernet PAUSE (802.3x)
  - Each VL can be independently PAUSEd

Priority-Based Flow Control (PFC)



- Enables lossless Fabrics for each class of service
- PAUSE sent per virtual lane when buffers limit exceeded
- Network resources are partitioned between VL's (E.g. input buffer and output queue)
- The switch behavior is negotiable per VL

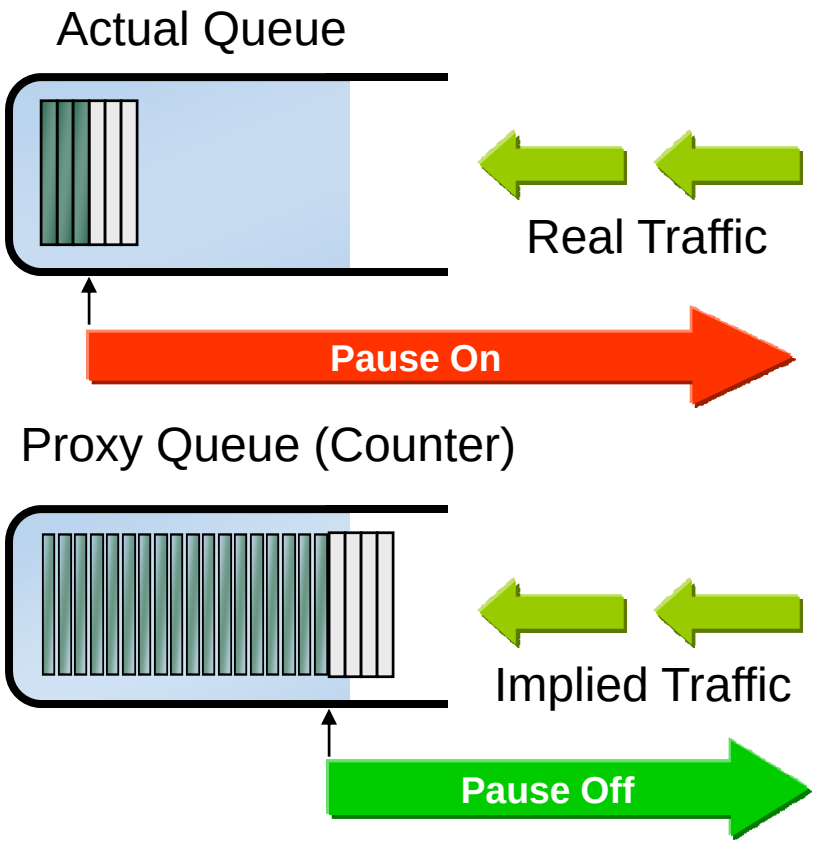
Per Priority PAUSE (PFC) *)



Time (n) is defined as the pause timer for class n, defined in the same manner as 31B.2

*) Based on a public proposal by Cisco that has high level of Industry support, now standardized in IEEE 802.1

Delayed Drop



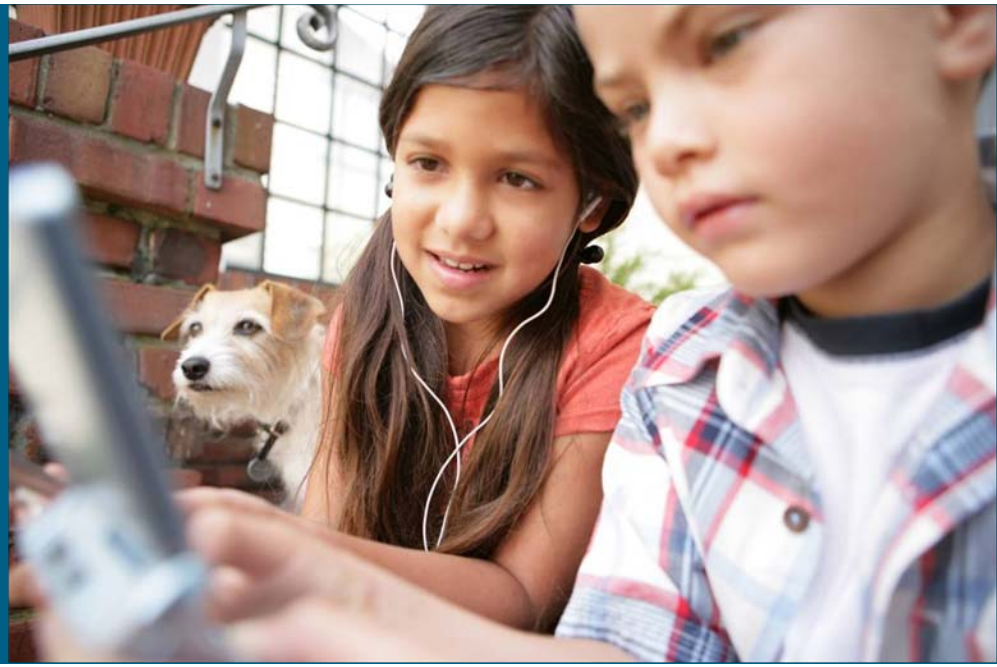
Actual Queue	*Proxy Queue*
Adds a frame	**Adds a frame**
Issues a PAUSE	**Adds frames at line rate**
Drains a frame	**Drains a frame**
Is Empty	**Drains frames at line rate**

The Proxy Queue measures the duration of the traffic flow

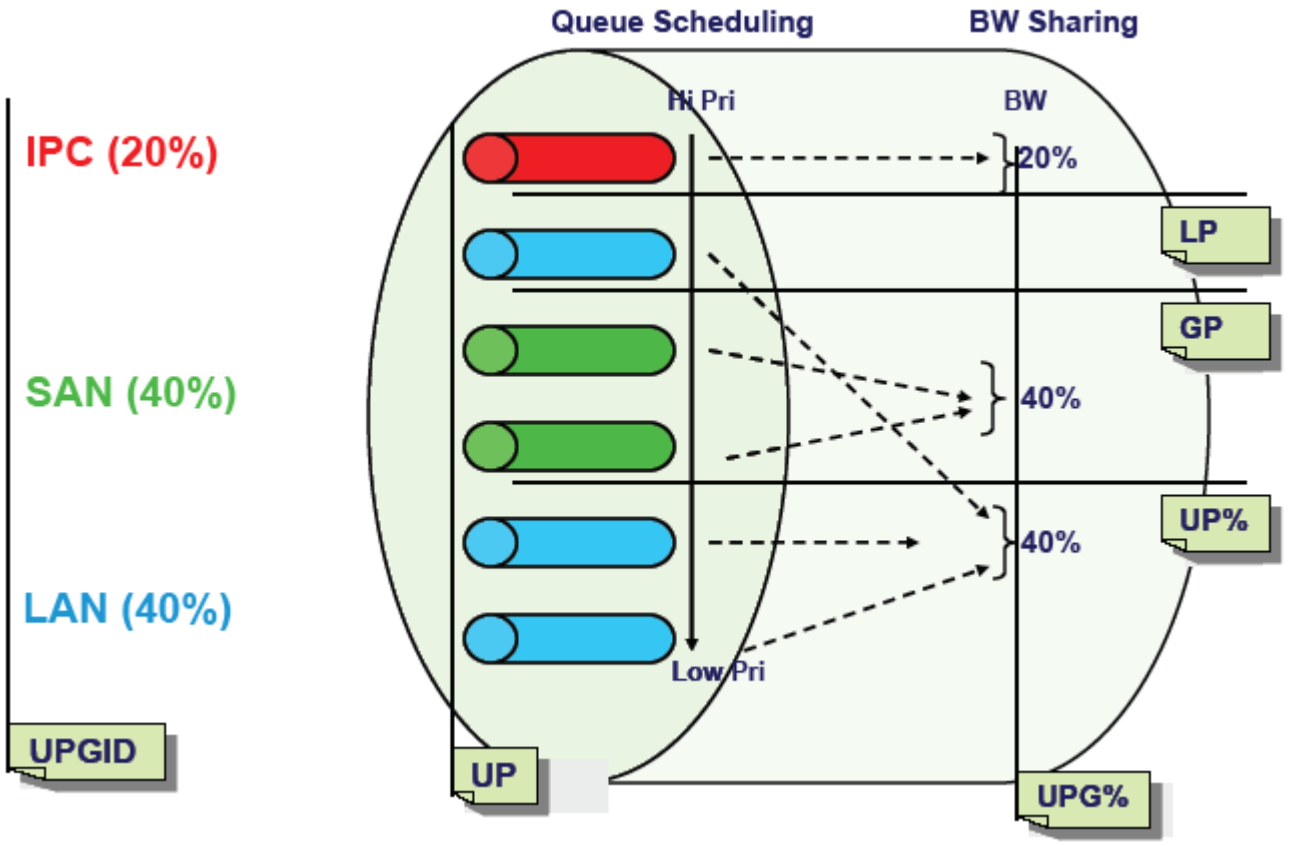
Delayed drop is a means of using PAUSE to mitigate the effects of short-term traffic bursts while maintaining Packet drop for long-term congestion

Bandwidth Management

ETS (Enhanced Transmission Selection) Priority Groups IEEE 802.1Qaz



Priority Groups

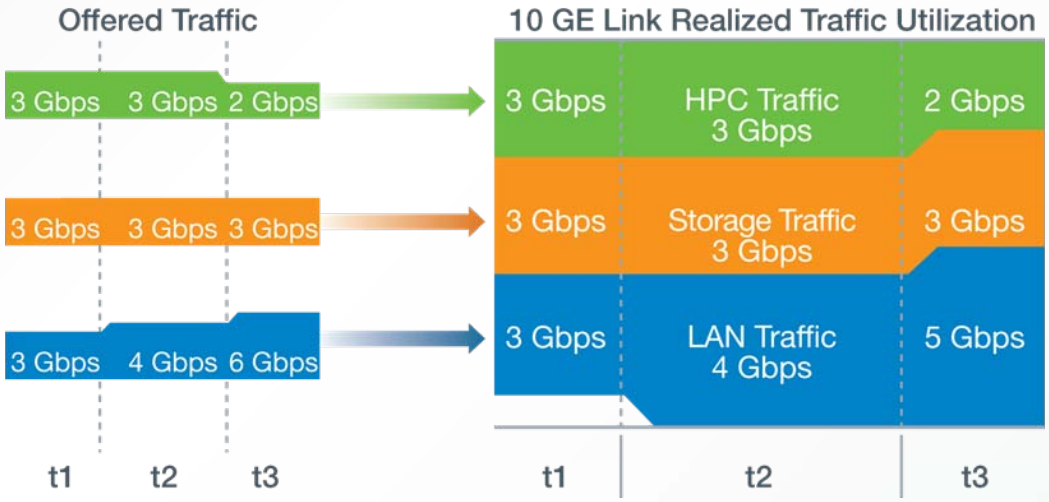


UP: User Priority	**This is actual marking of traffic on the wire (802.1p bits)**
User Priority Group (UPG) UPGID	**E.g. LAN, SAN, IPC, Management etc.**
UPG%	**% of Link Bandwidth allocated for a particular UPGID**
UP%	**% of Group Bandwidth allocated for a particular UP within UPGID**

Priority Groups (cont)

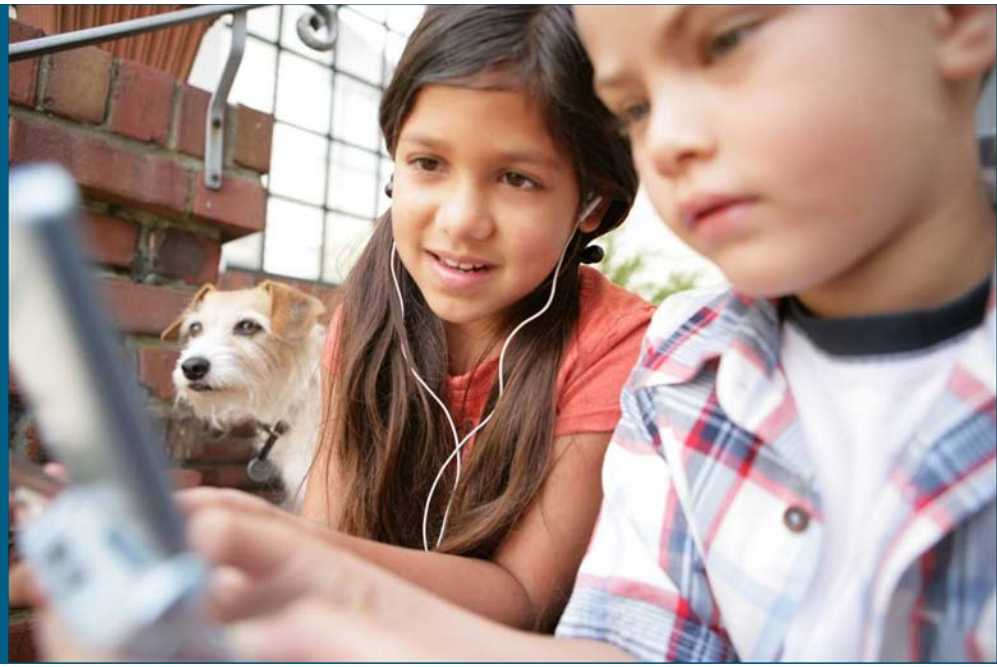
- ETS (Enhanced Transmission Selection) (IEEE 802.1Qaz)
  - Hardware efficient two-level DWRR with strict priority support
  - First level scheduling within each Priority Group
  - Second level scheduling between Priority Groups

Enhanced Transmission Selection (ETS)

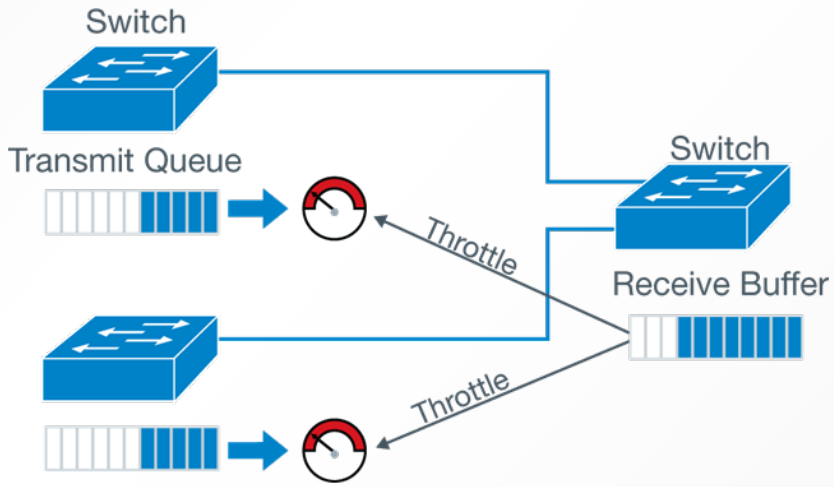


HPC Traffic: Priority Class High – 20% guaranteed bandwidth
LAN Traffic: Priority Class Medium – 50% guaranteed bandwidth
Storage Traffic: Priority Class Medium-High- 30% default bandwidth

Congestion Management



Congestion Management

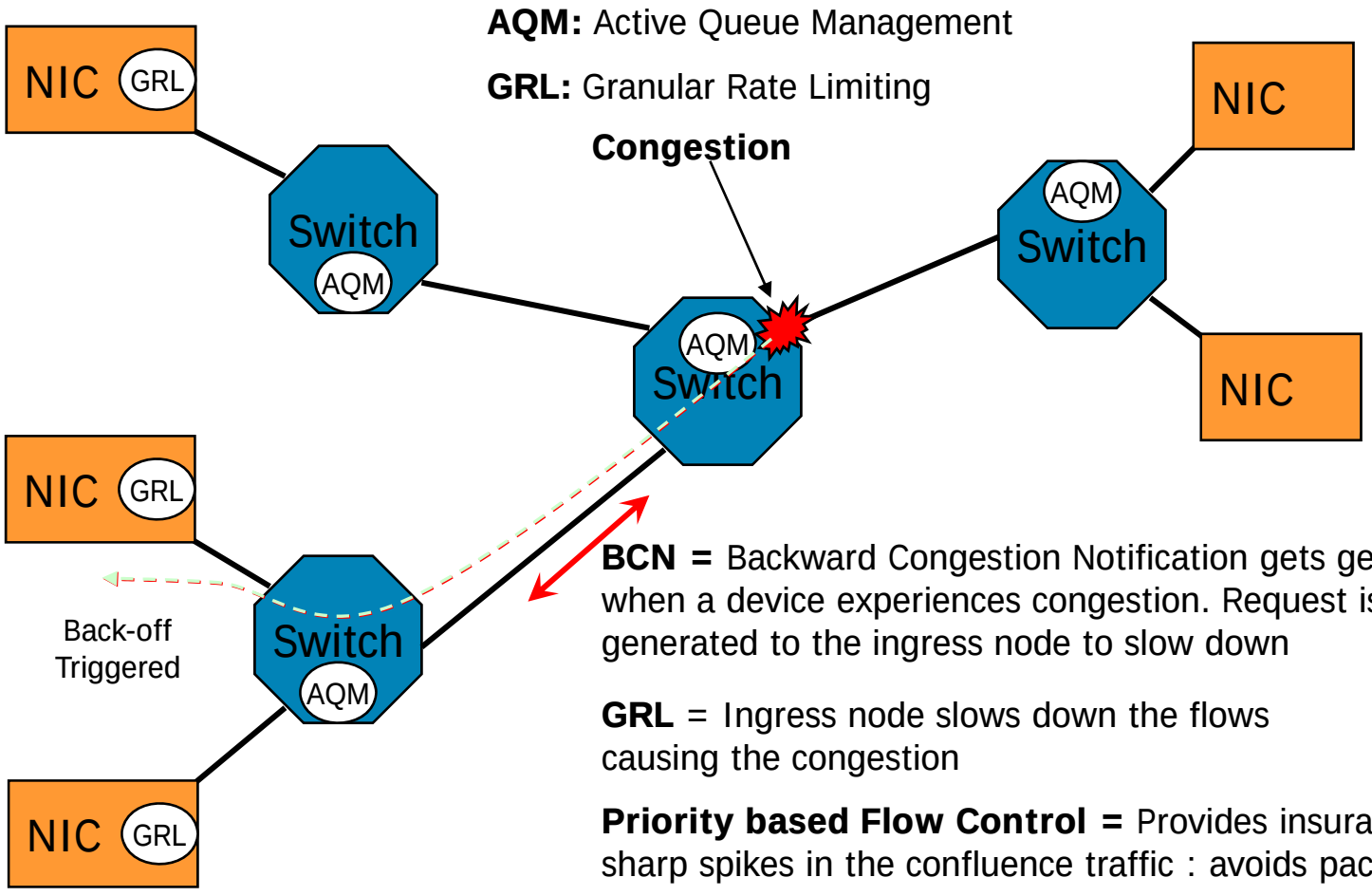


- **Lossless Ethernet: Congestion Spreading, HOL, Deadlock**
- **Moves congestion out of the core to avoid congestion spreading**
- **Allows End-to-End congestion management (Pause is hop by hop)**
- **Transient congestion - Priority Based Flow Control (PFC)**
- **Persistent congestion - Backward Congestion Notification**
- **Standards track in IEEE 802.1Qau**

Backward Congestion Notification (BCN)

- 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
  - AIMD (Additive Increase, Multiplicative Decrease) rate control
    - TCP window increases linearly in absence of congestion
    - Decreases exponentially (gets halved) at every congestion indication (either implicit or explicit)
  - Self-Clocking Control loop (acknowledgements)
- Derived from FCC (Fibre Channel Congestion Control)
- Works at L2 (applies to all traffic types, not only TCP)

Congestion Notification (IEEE 802.1Qau)



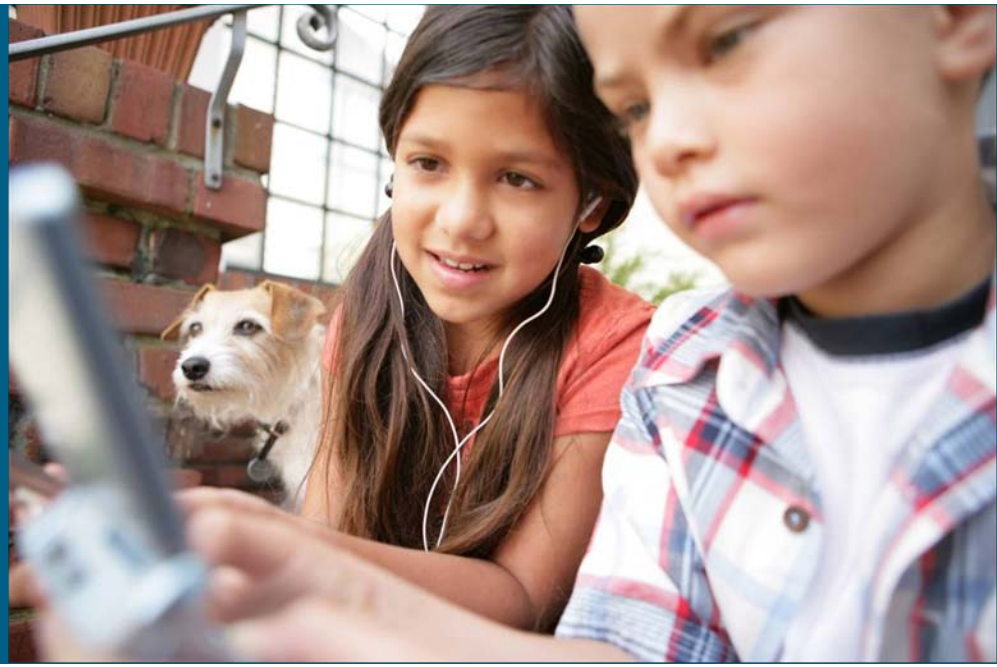
BCN = Backward Congestion Notification gets generated when a device experiences congestion. Request is generated to the ingress node to slow down

GRL = Ingress node slows down the flows causing the congestion

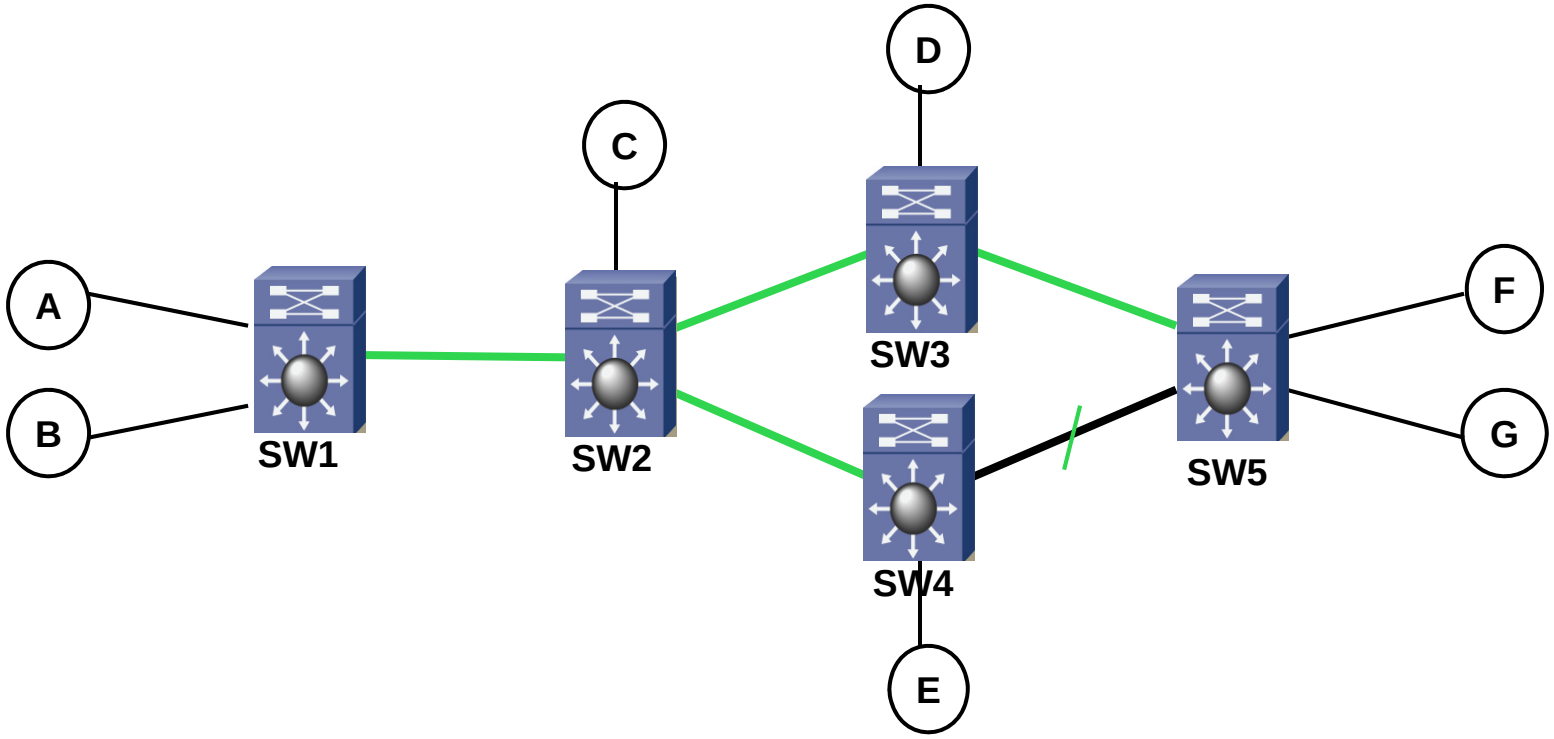
Priority based Flow Control = Provides insurance against sharp spikes in the confluence traffic : avoids packet drops

BCN = When congestion disappears, positive notification is generated to the ingress device allowing to grow the rate

L2 Multipathing



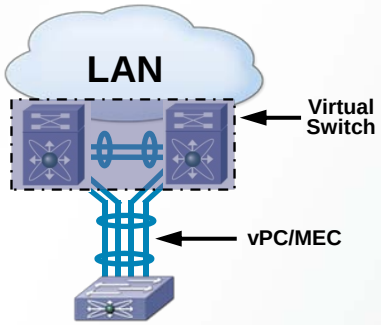
Classical Ethernet – Spanning Tree



- Link between SW4 & SW5 is not used

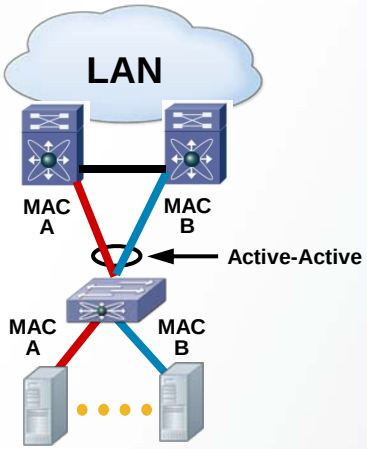
Layer 2 Multi-Pathing

Phase 1



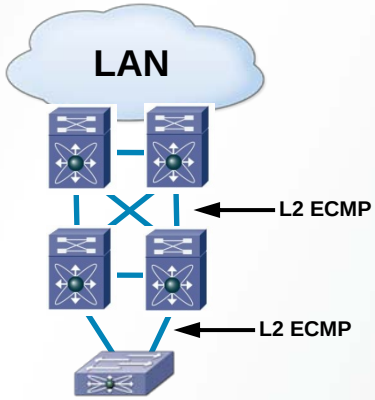
- Virtual Switch (VSS on C6K, vPC on Nexus 7K)
- Virtual port channel mechanism is transparent to hosts or switches connected to the virtual switch
- STP as fail-safe mechanism to prevent loops even in the case of control plane failure

Phase 2



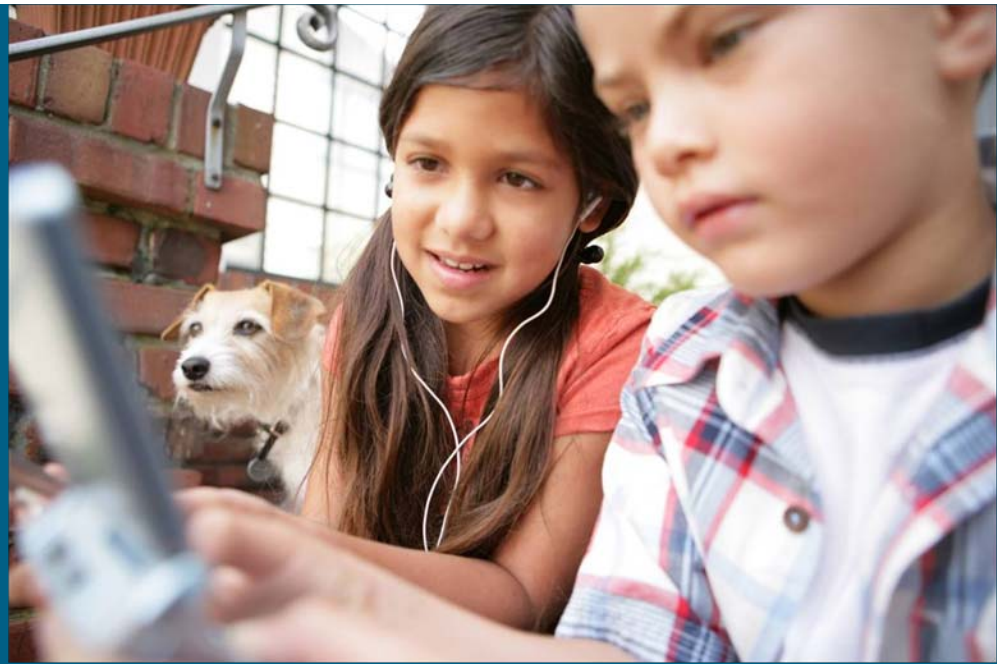
- Host Mode
- Eliminates STP on Uplink Bridge Ports
- Allows Multiple Active Uplinks Switch to Network
- Prevents Loops by Pinning a MAC Address to Only One Port
- Completely Transparent to Next Hop Switch

Phase 3

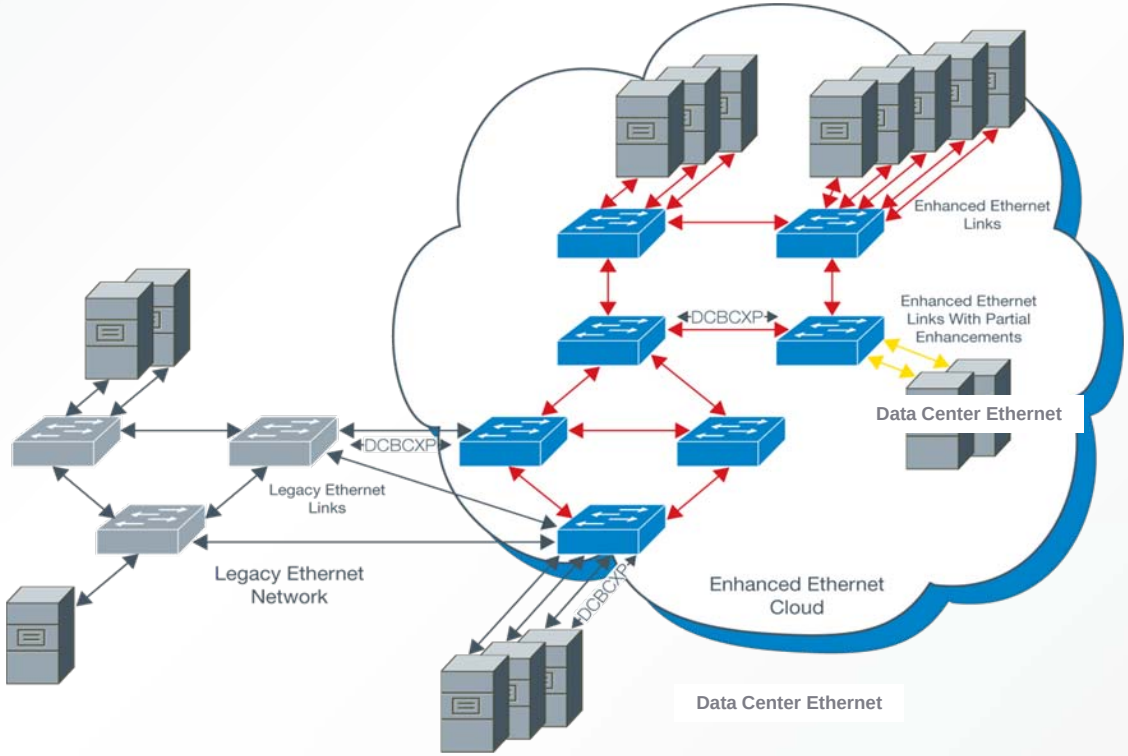


- Uses ISIS based topology
- Up to 16 way ECMP
- Eliminates STP from L2 domain
- Preferred path selection
- IETF Trill
- IEEE 802.1Qaq (Shortest-Path Bridging)

Management



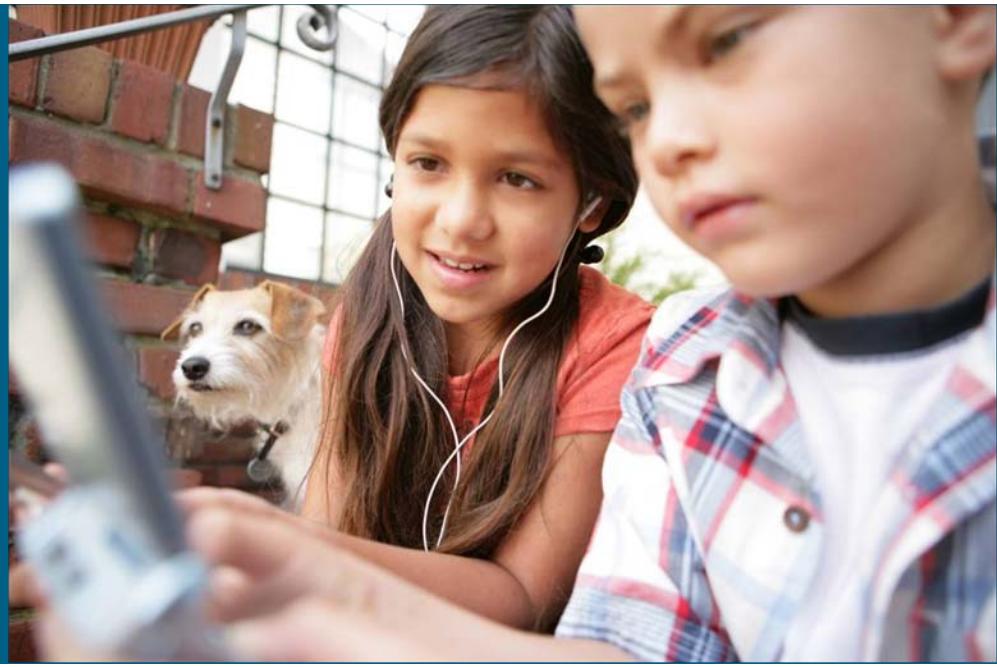
Data Center Bridging eXchange



Handshaking Negotiation for:

- **CoS BW Management**
- **Class Based Flow Control**
- **Congestion Management (BCN/QCN)**
- **Application (user_priority usage)**
- **Logical Link Down**

Summary



IEEE Enhancements for Data Center

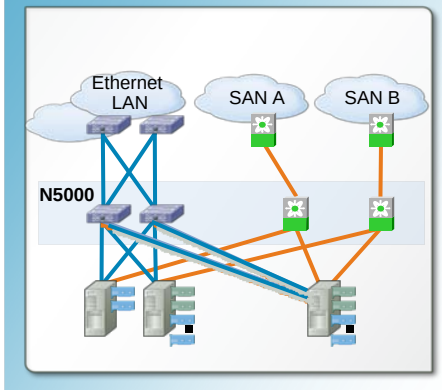
- IEEE projects necessary for IO Consolidation in Data Center
  - Congestion Notification: Approved project IEEE 802.1Qau
  - Shortest Path Bridging: Approved project IEEE 802.1Qaq
  - ETS (Priority Groups): Approved project IEEE 802.1Qaz
  - Priority based Flow Control (PFC): IEEE 802.1Qbb (Congestion Management task group is developing a PAR)
  - DCB Capability Exchange Protocol: Part of various projects above
- DCB Standards trending for ratification in ~2009

Cisco Unified Fabric: An Innovative Architecture To Simplify Data Center Transformation

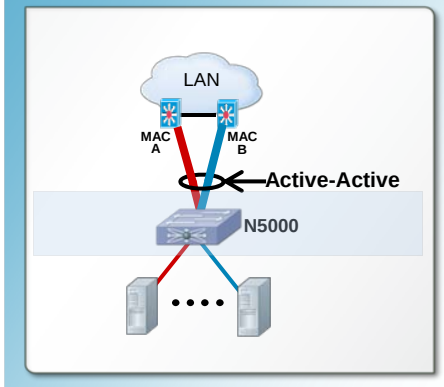
Standards

IETF IEEE ANSI

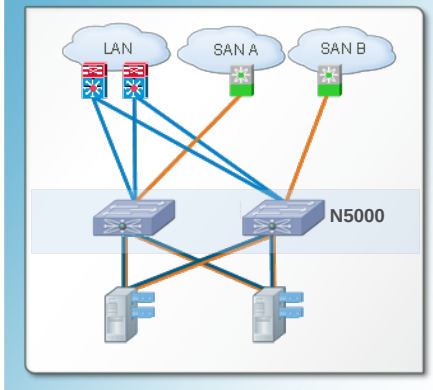
Wire Speed 10GbE Switching
Capacity



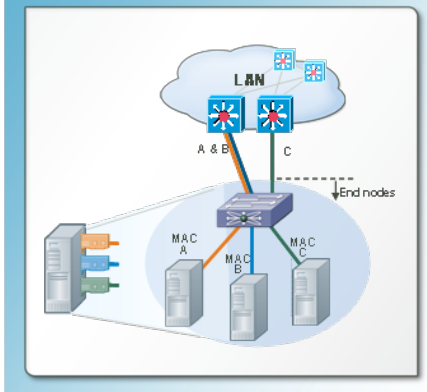
Data Center Ethernet (DCE)
Scalability



Fibre Channel over Ethernet (FCoE)
Consolidation



VM Aware Network Services
Virtualization



Eco-System

EMULEX NETXEN APC by Schneider Electric

QLOGIC intel PANDUIT vmware

Evolution of Ethernet Physical Media

Role of Transport in Enabling these Technologies!

Mid 1980's

10Mb

UTP Cat 3

Late 2000's

10Gb

X2

SFP+ Cu (BER better than 10^{-18})

SFP+ Fiber
Cat 6/7

Technology	Cable	Distance	Power (each side)	Transceiver Latency (link)
SFP+ CU Copper	Twinax	10m	~0.1W	~0.1µs
SFP+ USR ultra short reach	MM OM2 MM OM3	10m 100m	1W	~0
SFP+ SR short reach	MM 62.5µm MM 50µm	82m 300m	1W	~0
10GBASE-T	Cat6 Cat6a/7 Cat6a/7	55m 100m 30m	~8W ~8W ~4W	2.5µs 2.5µs 1.5µs

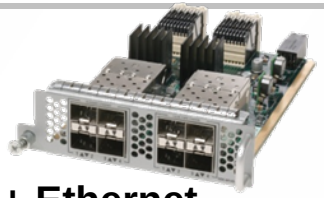
Cisco Nexus 5020

Industry's First I/O Consolidation Virtualization Fabric for Enterprise Data Center

Nexus 5020

56-Port L2 Switch
- 40 Ports 10GE/FCoE/DCE, fixed
- 2 Expansion Modules

Expansion Modules



FC + Ethernet
- 4 Ports 10GbE/FCoE/DCE
- 4 Ports 1/2/4G FC

Ethernet
- 6 Ports 10GE/FCoE/DCE

Eco System Partners QLOGIC EMULEX intel NETXEN PANDUIT APC vmware

OS	**Cisco NX-OS**
Mgmt	**Cisco Fabric Manager and Cisco Data Center Network Manager**

References

- http://ieee802.org/802_tutorials/nov07/Data-Center-Bridging-Tutorial-Nov-2007-v2.pdf
- http://ieee802.org/1/files/public/docs2007/new-cm-barrass-pause-proposal.pdf
- http://ieee802.org/1/files/public/docs2007/new-wadekar-priority-groups-1107-v1.pdf
- http://www.ieee802.org/1/files/public/docs2007/au-bergamasco-ethernet-congestion-manager-070313.pdf
- http://download.intel.com/technology/eedc/dcb_cep_spec.pdf
- http://fcoe.com/
- http://www.open-fcoe.org/