QLOGIC
The Ultimate in Performance

Industry Standards for the Exponential Growth of Data Center Bandwidth and Management

Craig W. Carlson

Or…

Finding the Fat Pipe through standards

Creative Commons, Flikr User davepaker

Overview

- **Part of managing a data center is having the bandwidth to get the job done, in this presentation:**
  - I will present the latest Standards that UP your bandwidth
  - I will then also give a look into what the Standards future holds for bandwidth

Creative Commons, Flikr User Grotuk

Skinny Pipes cause problems!

Agenda

- **Overview**
- **The Latest**
  - FC Standards
    - 16 GFC Completed
    - FCoE
  - IEEE 802
    - DCB – Lossless Ethernet
    - 40/100 G Completed
    - Energy Efficient Ethernet
  - Infiniband

Agenda Continued

- **Upcoming Standards**
  - FC Standards
    - 32 GFC
    - FCoE 2nd gen.
    - Energy Efficient Fibre Channel
  - IEEE 802
    - DCB – Virtual Bridging
- **And beyond…**
  - FCIA roadmap

The Latest

FC Standards

Creative Commons, Flikr User Daniel*1977

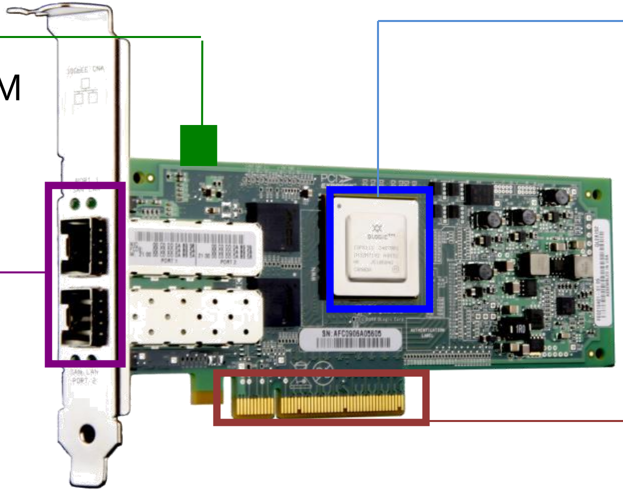
Fibre Channel Standards

- **16GFC Standard Finished**
  - This June NCITS T11 FC-PI-5 forward to ANSI for publication
  - October 2009 PI-5 was “technically stable” standard
    - Also finished joint T11.2/T11.3 work for auto-negotiation 8b/10b to 64b/66b
  - 16GFC provides for a good fit between 10G Ethernet and FC
- **16GFC Products now available**

Port Compatible

- **16GFC is port compatible**
  - Speed Negotiation from 2, 4, and 8 GFC on a single port

Server Connectivity



Optimized for Virtualization
Support for industry standards and OEM proprietary virtualization technologies

Simultaneous Data and Storage Connectivity
Ethernet LAN & NAS; FC, FCoE, and iSCSI storage

Lowest CPU Utilization
Protocol offload for FC, FCoE, TCP/IP, iSCSI and IPSec

PCIe Gen3 Compliant & SR-IOV Support
Greater performance & scalability in VM environments

The new look for adaptable, high performance fabrics . . .

16GFC

- **16GFC specifications**
  - 10-14 meters on copper cabling
  - 100 meters on multi-mode OM3
  - 10 km on single-mode
  - New encoding
    - Prior to 16GFC used 8B/10B encoding
    - 16GFC uses 64/66 Encoding
    - 64/66 allows improved signal integrity
    - 16GFC ports are fully backwards compatible with slower speed FC ports
      - Speed negotiation compatible with 2, 4, 8 GFC

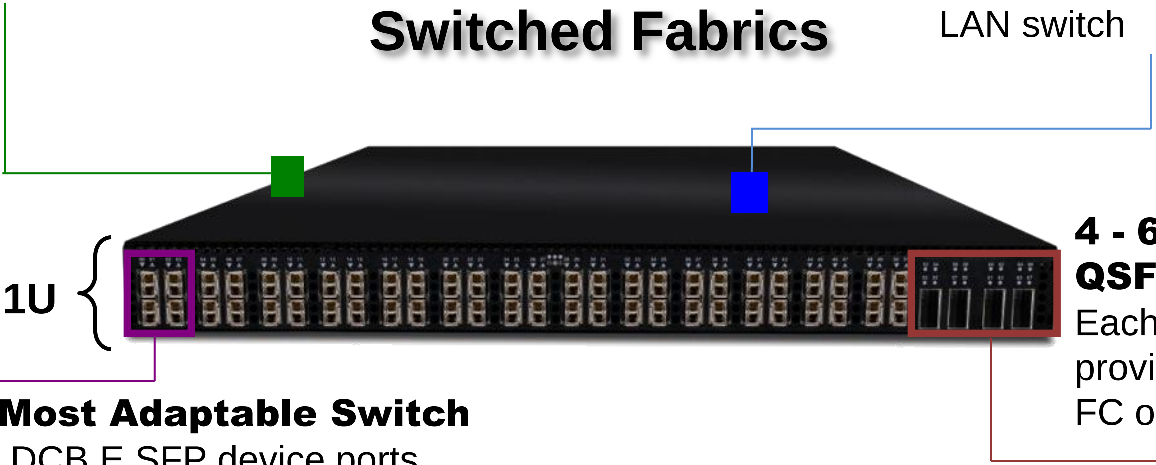
16GFC

- Multiple port speeds supported on a single switch port, 2, 4, 8, 16 GFC

Highest Port Count FCoE TOR Switch
No connectivity limitations

Switched Fabrics

Hybrid Fabrics + ETR, VEPA
Easily attach to any SAN or LAN switch



4 - 64Gb FC / 40GbE QSFP Uplink Ports
Each QSFP port can also provide four additional 16Gb FC or 10GbE SFP ports

Highest Density, Most Adaptable Switch
52 - 16Gb FC / 10Gb DCB E SFP device ports

The Latest

IEEE 802 Standards

Creative Commons, Flikr User Daniel*1977

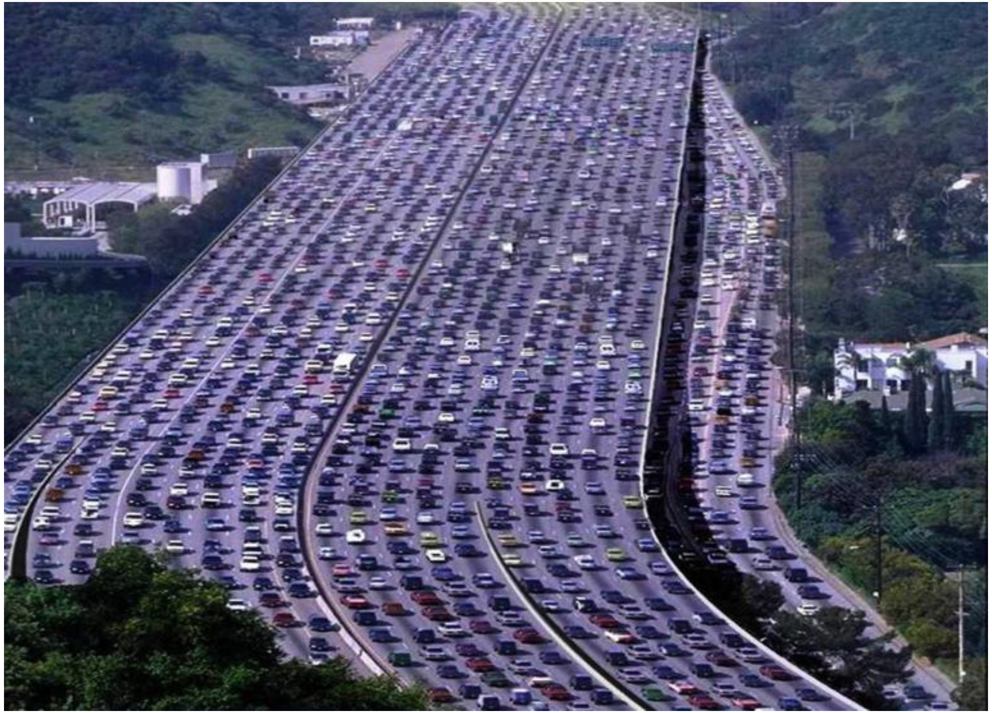
DCB

- **IEEE 802.1 – Data Center Bridging Work Group**
  - Last of Lossless Ethernet Standards completed in 2011
    - 802.1Qbb – Priority Flow Control
    - 802.1Qau – Congestion Management
    - 802.1Qaz – Enhanced Transmission selection

DCB – Lossless Ethernet

- **Problem Statement**
  - Challenges using storage protocols on Ethernet
  - Dropped Frames – Ethernet switches are traditionally designed to drop frames in some conditions such as congestion
  - Flow Control – Flow control does exist for Ethernet in the form of PAUSE, but this affects all traffic
  - Convergence – Convergence of Storage traffic with other types of traffic is possible on Ethernet networks
- **All of these items can be caused by, or cause congestion**

Congestion



- **Congestion in Ethernet switches…**

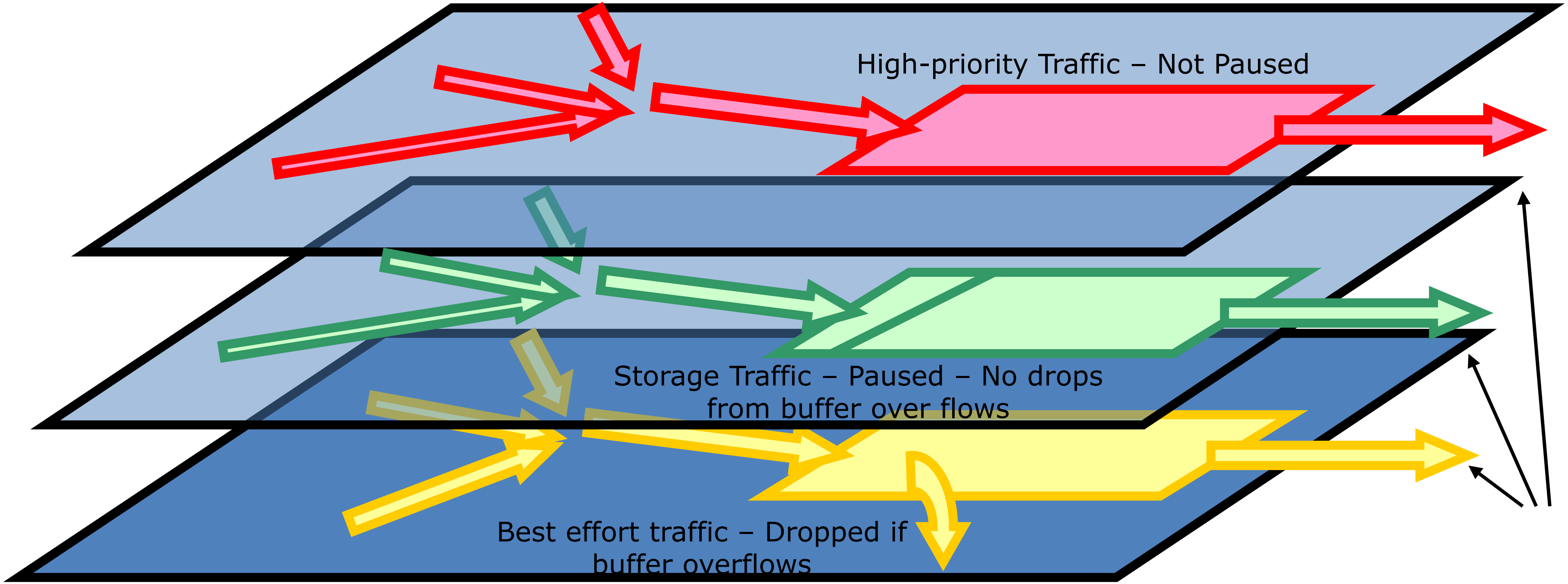
Drops

- **Can cause dropped packets…**

802.1Qbb – Priority-based Flow Control

- **PAUSE mechanism was defined for Ethernet in 802.3**
  - This PAUSE is a single PAUSE for ALL traffic on a link
  - Can cause congestion spreading if one type of traffic is hogging the link
  - Some types of traffic may not see a benefit from PAUSE
- **Priority-based Flow Control allows PAUSE on a specific Ethernet Priority**
  - PAUSE can be enabled only for traffic which needs it

802.1Qbb – Priority Based Flow Control

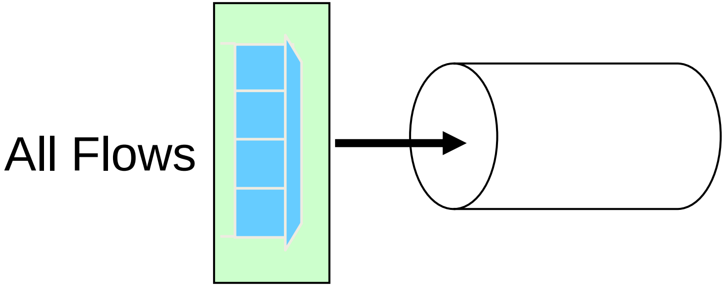

High-priority Traffic – Not Paused
Storage Traffic – Paused – No drops from buffer over flows
Best effort traffic – Dropped if buffer overflows

802.1Qaz – Enhanced Transmission Selection

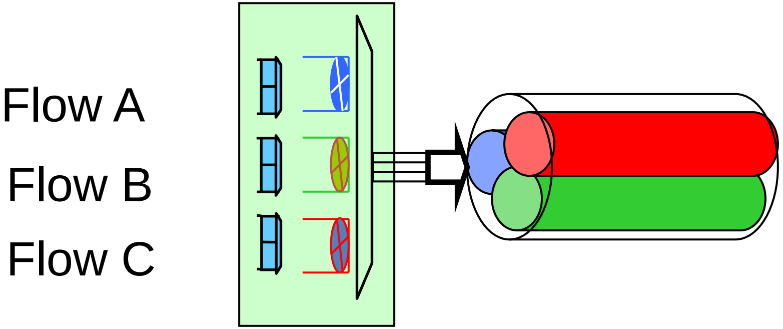
- **Allows assignment of “Best Effort” Quality of Service to Ethernet Traffic Classes**
  - Bandwidth is allocated by percentage of the available bandwidth
  - If one Traffic class does not consume all of it’s share, then others may use that unused portion of bandwidth
- **Designed for converging multiple traffic types (FCoE, TCP, HPC) on a single link**
  - Best effort nature means that bandwidth is not wasted on a traffic type which is not using it’s share

802.1Qaz – Enhanced Transmission Selection

- **Without traffic differentiation all traffic is sent together**



- **With ETS traffic may be divided**
- **For example:**
  - Flow A and B, managed by ETS, can be given 50% traffic each
  - If Flow C is not managed by ETS then Flow A and B receive 50% available after Flow C usage
  - If Flow B and C use nothing, then A can use full link bandwidth

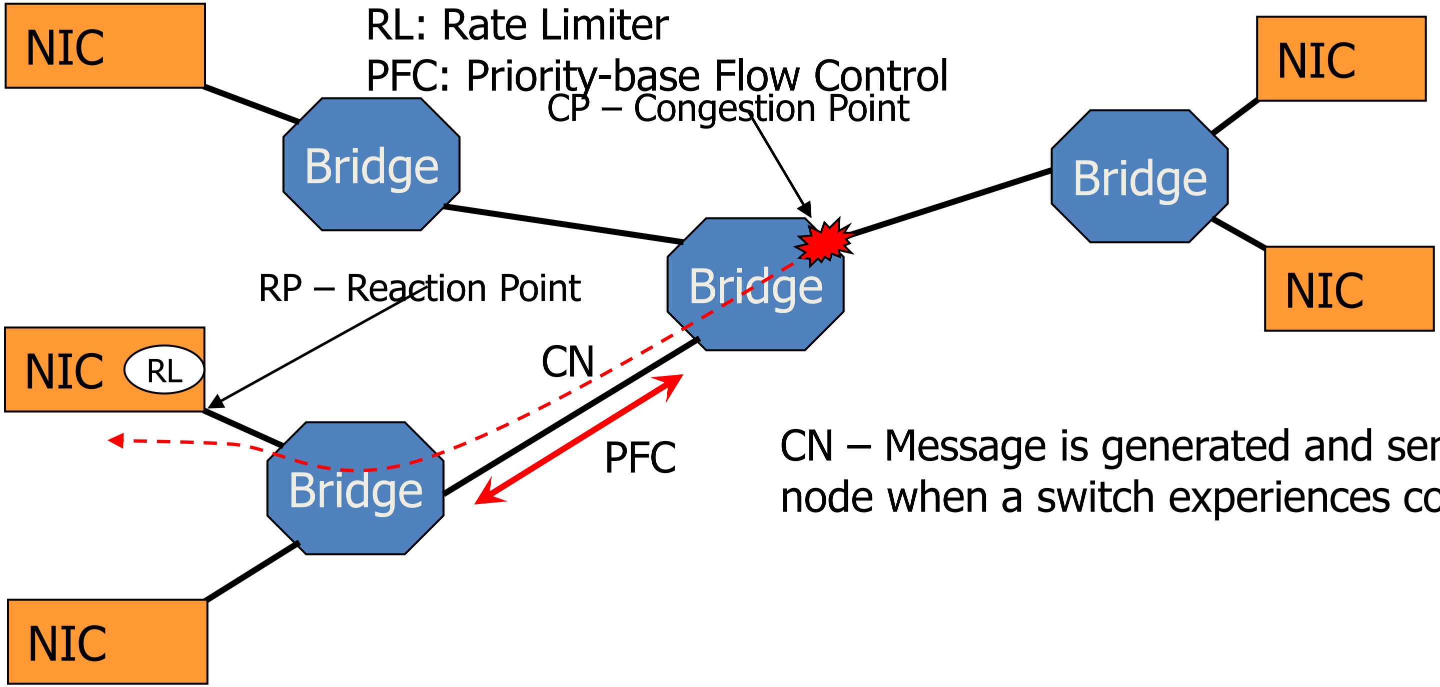


802.1Qau – Congestion Management

- **Provides a mechanism for managing congestion in an Ethernet network, provides the following:**
  - Detection
  - Notification
  - Rate Limiter
- **This is done using two types of entities within a network**
  - Congestion Point – an element of a switch or end station that can detect congestion and send notification messages
  - Reaction Point – an element of an end station which can limit the rate of transmission based on messages received

802.1Qau – Congestion Notification

CN: Congestion Notification
RL: Rate Limiter
PFC: Priority-base Flow Control

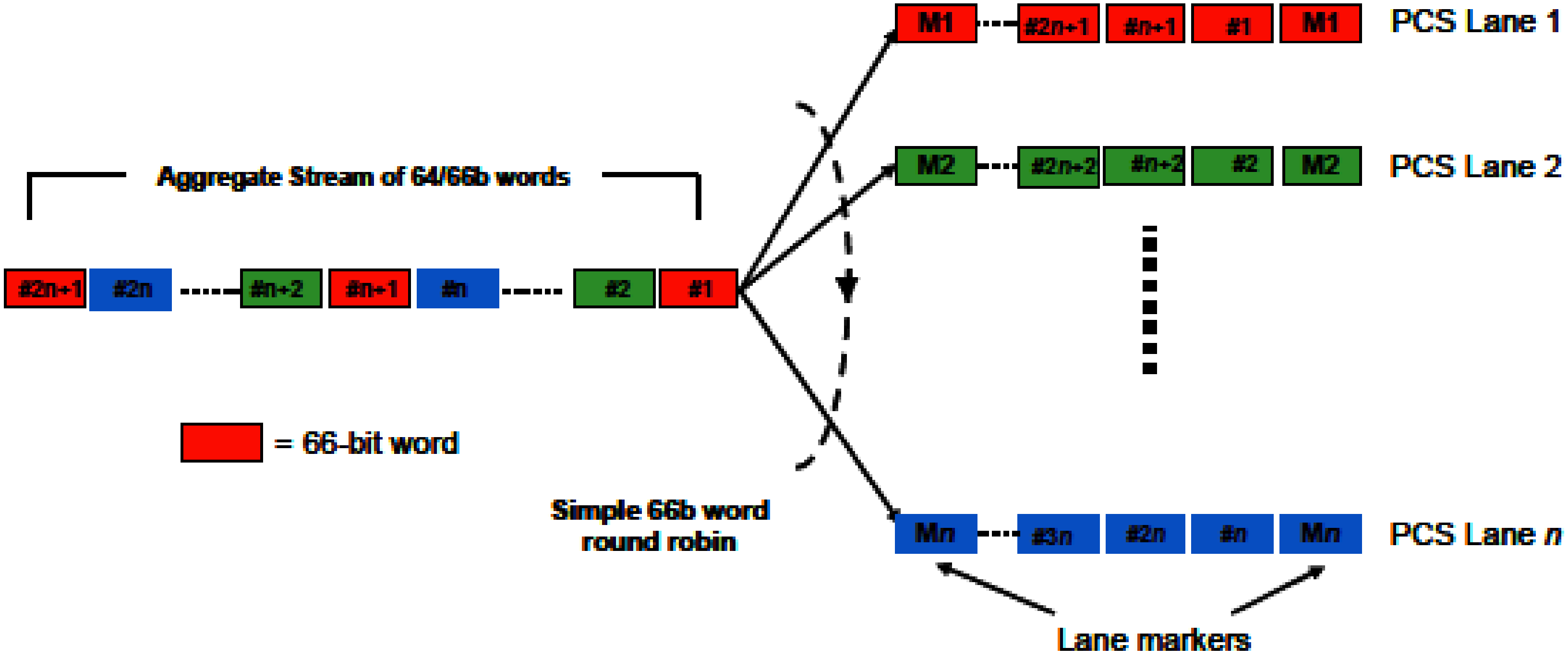


CN – Message is generated and sent to ingress node when a switch experiences congestion

40/100 Gb Ethernet

- **IEEE 802.3ba**
  - Completed 2010
  - Defined as multi-lane 64/66 encoding
- **Specs**
  - 40 Gb/s Ethernet
    - 1 m backplane
    - 10 m copper
    - 100 m OM3 Multi-mode fiber
    - 10 km single-mode fiber
  - 100 Gb/s Ethernet
    - 10 m Copper
    - 100 m OM3 Multi-mode
    - 10 Km single-mode fibre

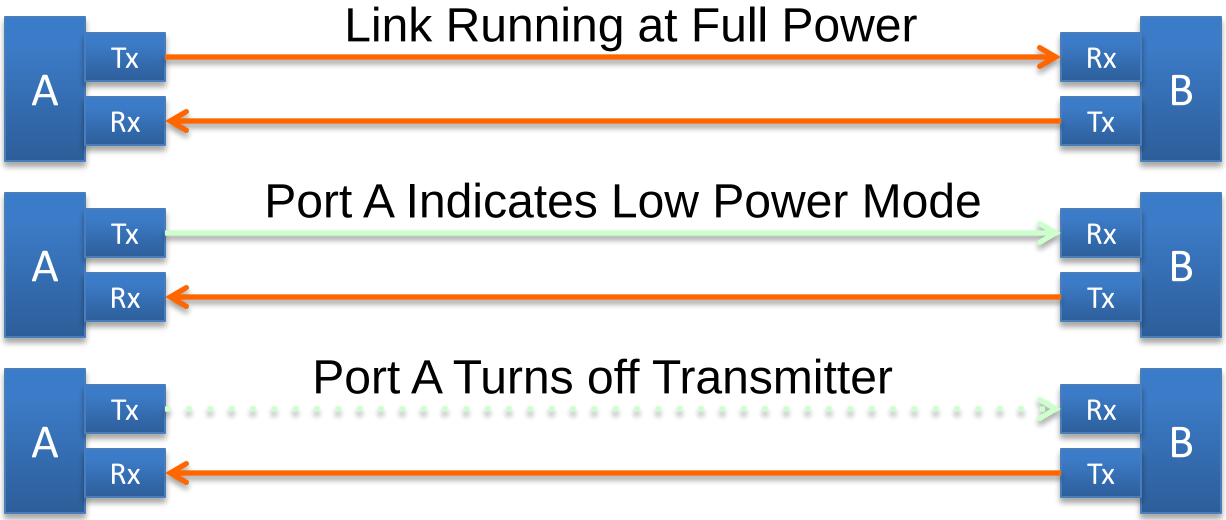
Multi-lane example



Energy Efficient Ethernet

- **802.3az defines an Energy Efficient Ethernet mode of operation**
  - During periods of low link utilization (i.e., no data to transmit)
    - Link can enter periods of low power usage by shutting of the transmitter/receiver for periods of time
    - Designed to have minimum impact on data traffic – Quick wake up times
  - Optional behavior, links decide when to go into low power mode

Low Power Example



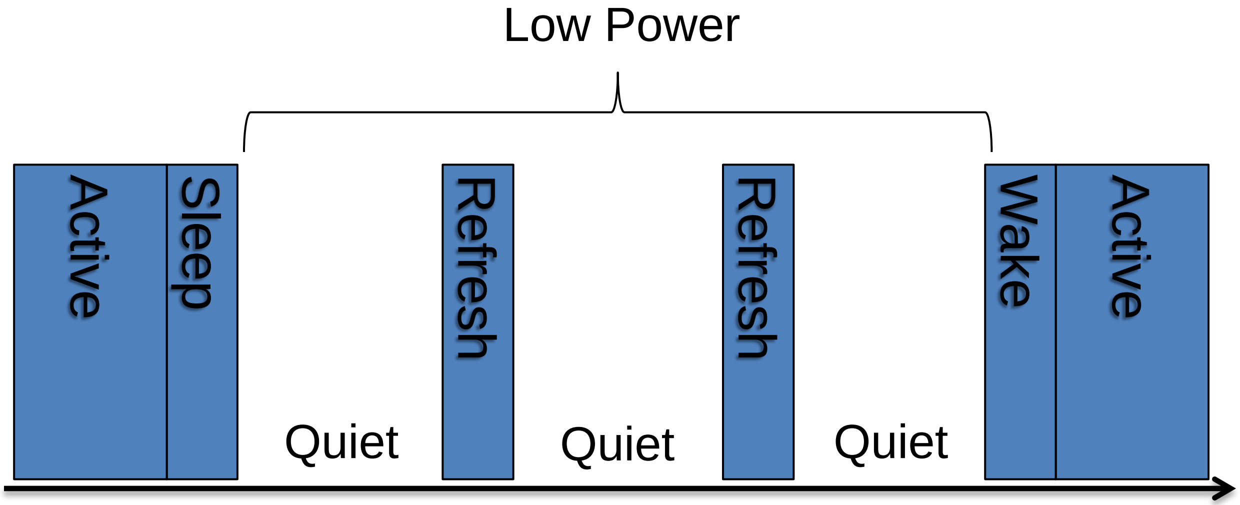
Link Running at Full Power

Port A Indicates Low Power Mode

Port A Turns off Transmitter

Port B may enter Low Power or
Remain in Full Power

Energy Efficient Ethernet Operation



- **Sleep/Refresh/Quiet times are short intervals**
  - Sleep/Refresh is in the microsecond range
  - Quite is in the millisecond range
  - Exact timings depend on interface type

The Latest

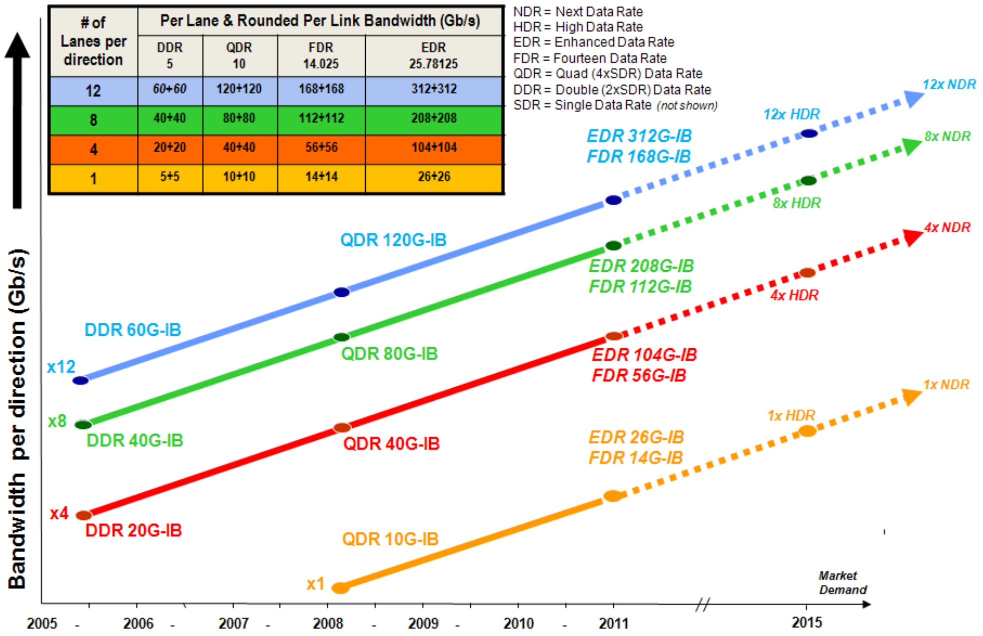
Infiniband

Creative Commons, Flikr User Daniel*1977

InfiniBand Performance Trends: Bandwidth

- **QDR → FDR Bandwidth**
  - 10G → 14.065G signaling
  - 8b/10b → 64b/66b encoding
  - 2.0x host: 27 Gbps → 54 Gbp
  - 1.7x ISL: 32 Gbps → 54 Gbp

- **QDR → EDR Bandwidth**
  - 10G → 25.78125G signaling
  - 8b/10b → 64b/66b encoding
  - 3.7x host: 27 Gbps → 100 Gb
    - Requires PCIe Gen3 x16
  - 3.1x ISL: 32 Gbps → 100 Gbps



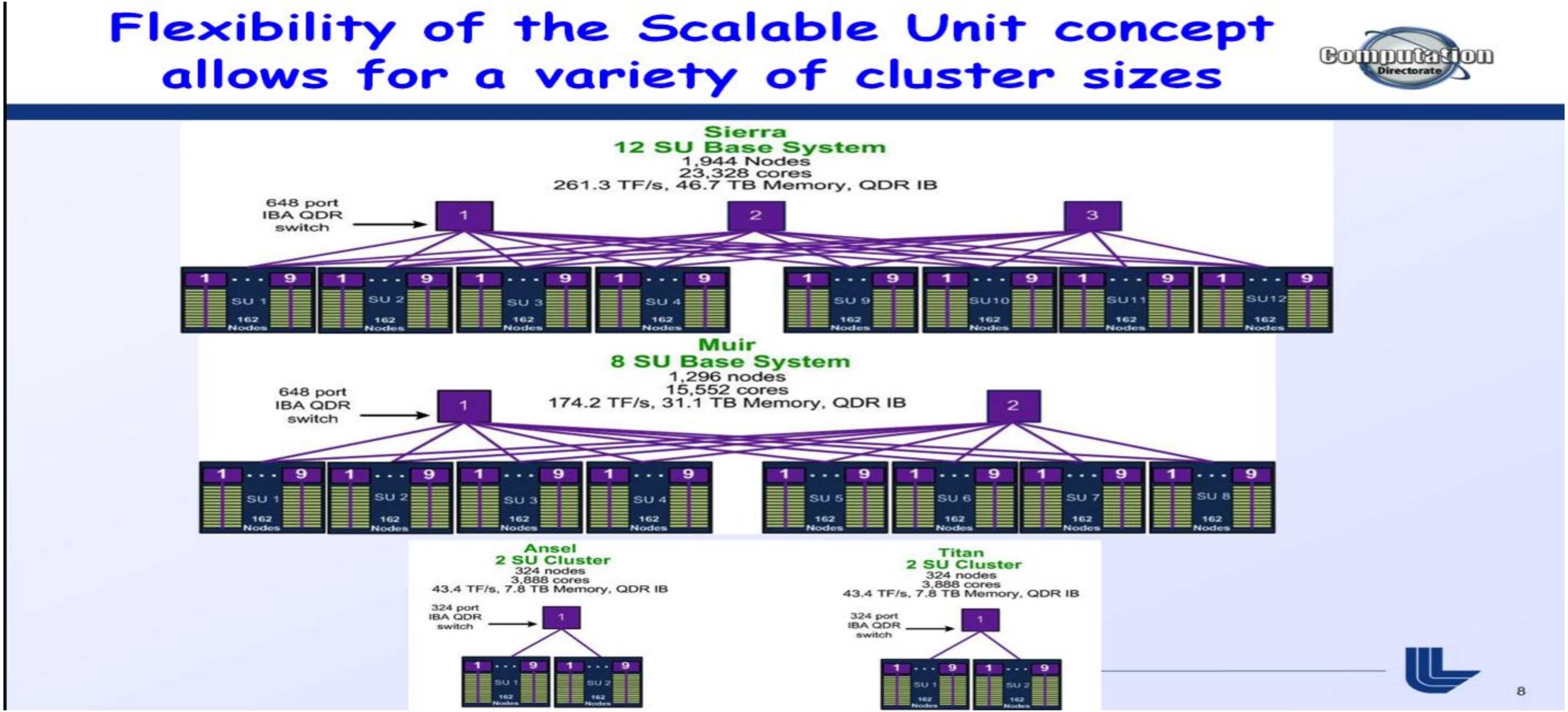
# of Lanes per direction	Per Lane & Rounded Per Link Bandwidth (Gb/s)			
	DDR 5	QDR 10	FDR 14.025	EDR 25.78125
12	60+60	120+120	168+168	312+312
8	40+40	80+80	112+112	208+208
4	20+20	40+40	56+56	104+104
1	5+5	10+10	14+14	26+26

NDR = Next Data Rate
HDR = High Data Rate
EDR = Enhanced Data Rate
FDR = Fourteen Data Rate
QDR = Quad (4xSDR) Data Rate
DDR = Double (2xSDR) Data Rate
SDR = Single Data Rate *(not shown)*

Lawrence Livermore National Laboratory

All systems implement Fat Tree topology using QLogic switches

QLOGIC
The Ultimate in Performance



Flexibility of the Scalable Unit concept allows for a variety of cluster sizes

Computation Directorate

Sierra
12 SU Base System
1,944 Nodes
23,328 cores
261.3 TF/s, 46.7 TB Memory, QDR IB

Muir
8 SU Base System
1,296 nodes
15,552 cores
174.2 TF/s, 31.1 TB Memory, QDR IB

Ansel
2 SU Cluster
324 nodes
3,888 cores
43.4 TF/s, 7.8 TB Memory, QDR IB

Titan
2 SU Cluster
324 nodes
3,888 cores
43.4 TF/s, 7.8 TB Memory, QDR IB

Upcoming Standards

Creative Commons, Flikr User bcmacsac1

Agenda check

- **Upcoming Standards**
  - FC Standards
    - 32 GFC
    - FCoE 2nd gen.
    - Energy Efficient Fibre Channel
  - IEEE 802
    - DCB – Virtual Bridging
- **And beyond…**
  - FCIA roadmap

Upcoming Standards

Fibre Channel

Creative Commons, Flikr User Daniel*1977

Upcoming Standards – 32 GFC

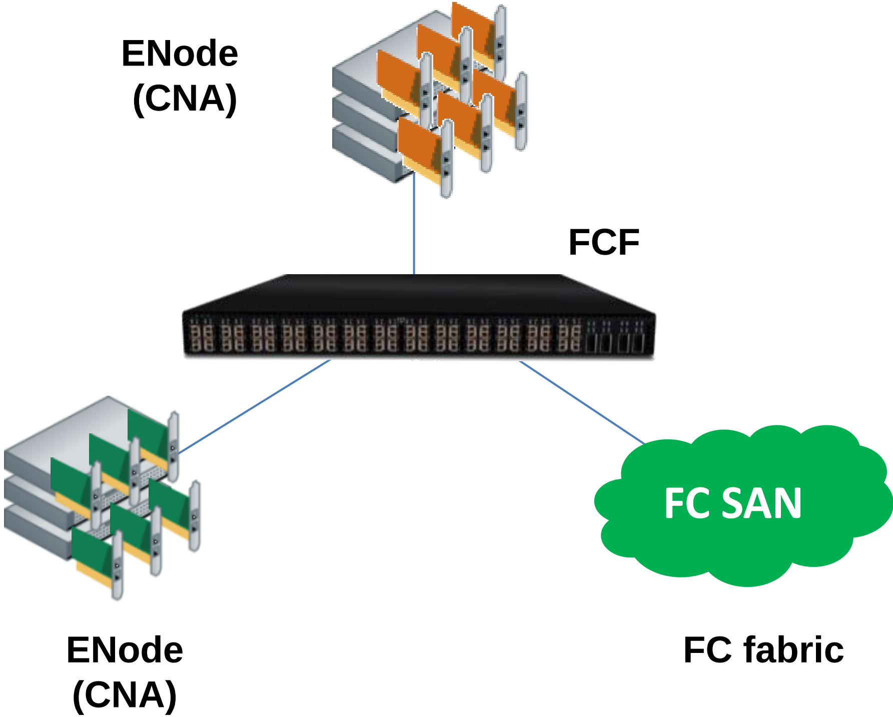
- **Next speed**
  - 32 GFC!
  - Definition taking place now
    - Technical completion 2012
    - Products 2014
    - Based on 64/66 encoding, same as 16 GFC
    - Port speed-negotiation backward compatible with 4,8,16 GFC

Upcoming Standards – FCoE update

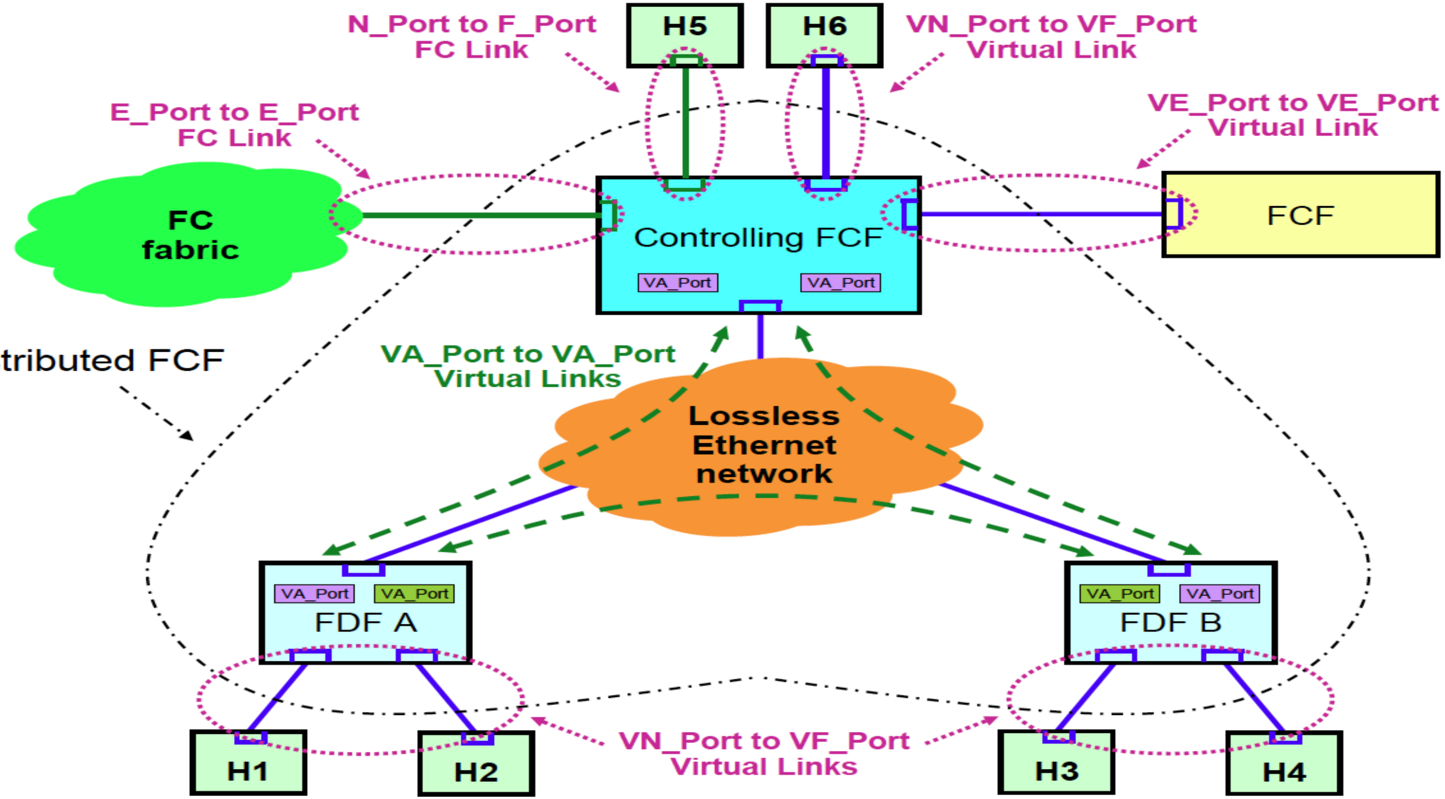
- **FC-BB-6 is under way defining the next generation of FCoE**
- **Three new topologies**
    - Distributed FCF – Including a redundancy protocol
    - Point-to-multipoint
    - Point-to-point

FCoE Term Refresher

- FCF – And FCoE aware Ethernet switch
- ENode – An FCoE interface adapater (also termed Converged Network Adapter – CAN)



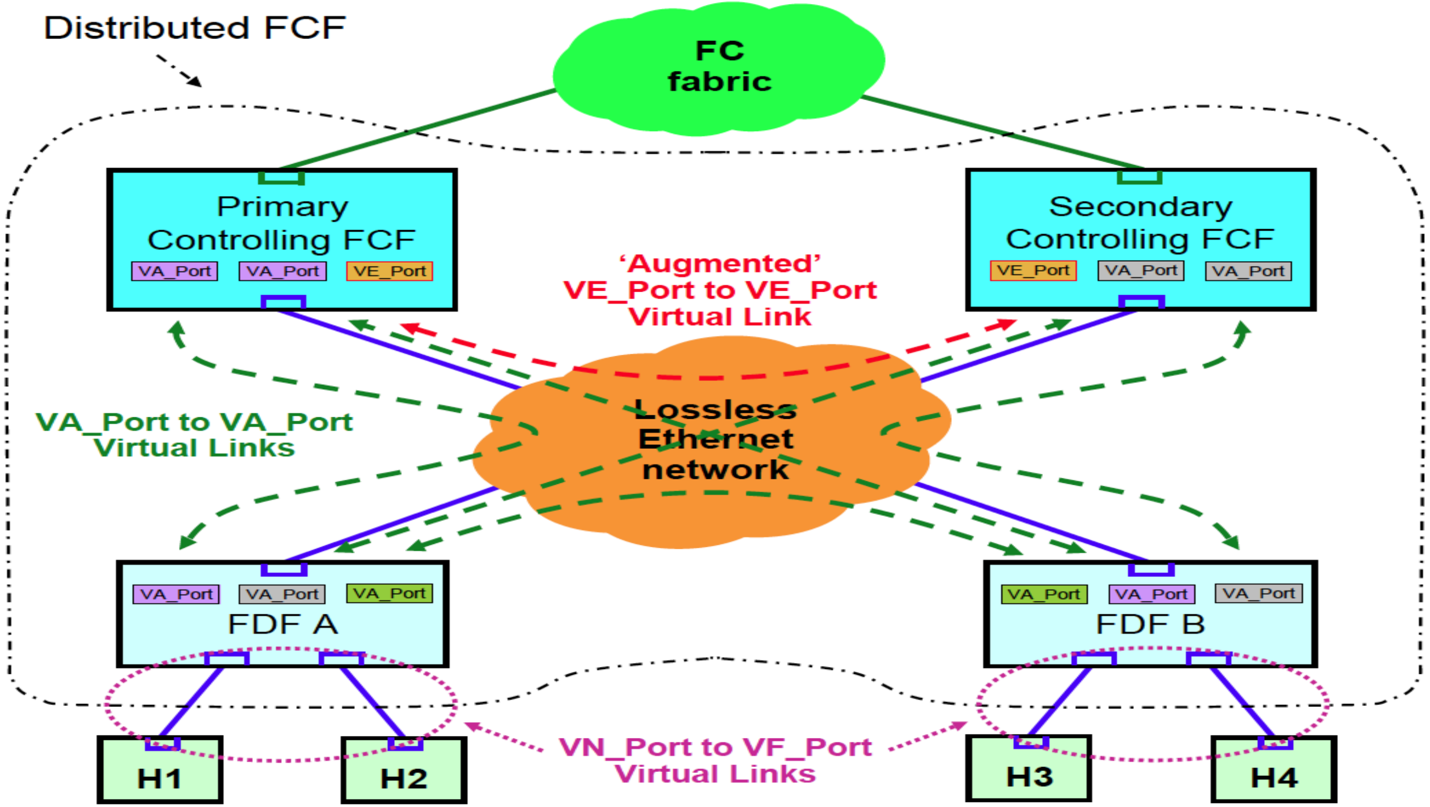
Distributed FCF



Distributed FCF Redundancy

- **Goal is to eliminate single point of Failure for a Distributed FCF**
- **Within a Distributed FCF**
  - A Primary and Secondary Controlling FCF is elected
  - Primary Controlling FCF operations Virtual Domain until failure – Then Secondary takes over
- **Only covers a single point of failure, not a double failure**
  - Protocol allows for a Primary and Secondary, but no further redundancy

Distributed FCF Redundancy

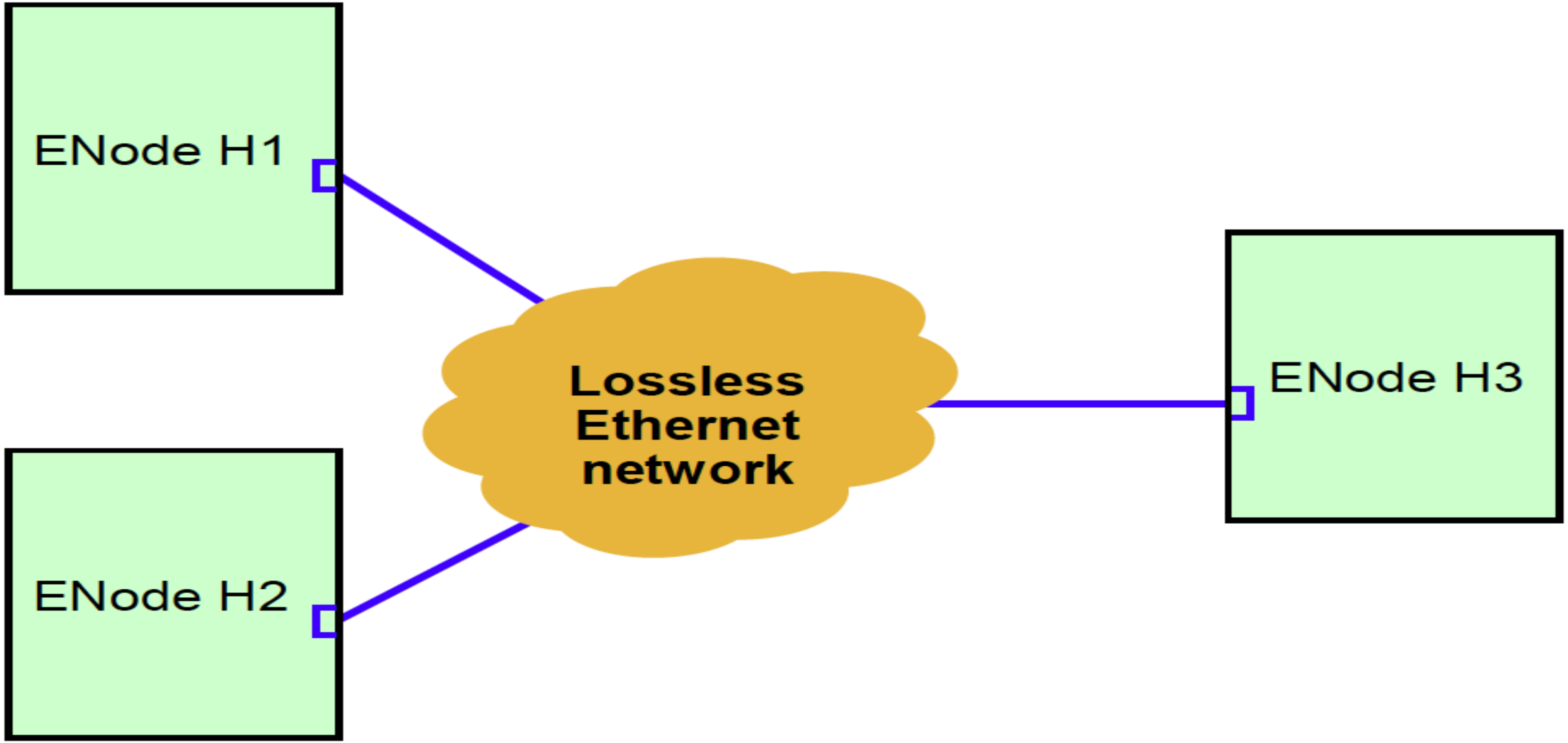


Distributed FCF
FC fabric
Primary Controlling FCF
VA_Port
VA_Port
VE_Port
Secondary Controlling FCF
VE_Port
VA_Port
VA_Port
'Augmented' VE_Port to VE_Port Virtual Link
VA_Port to VA_Port Virtual Links
Lossless Ethernet network
VA_Port
VA_Port
VA_Port
FDF A
VA_Port
VA_Port
VA_Port
FDF B
VN_Port to VF_Port Virtual Links
H1
H2
H3
H4

Point to multi-point (VN2VN)

- **Similar to Arbitrated Loop in Fibre Channel**
  - But, not a loop – works through Ethernet switches without an FCF
  - Requires new FIP commands to establish addresses
    - Selected by a random process
    - Claimed via a FIP message
    - After an address is claimed, an advertisement is sent

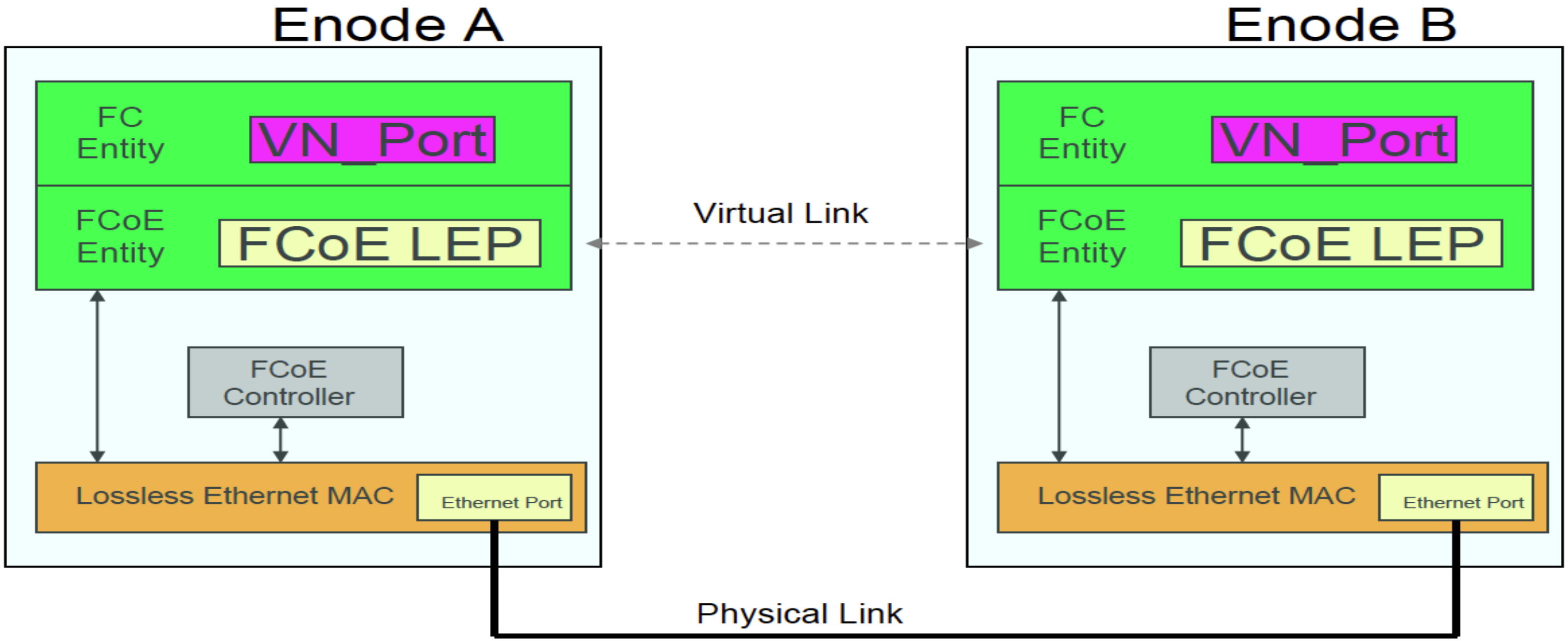
Point-to-multipoint Topology



Point-to-Point

- **Point-to-point configuration as in Fibre Channel point-to-point**
  - Works in a true point-to-point configuration - not intended to be used with a switch
  - If more than one neighbor is detected, drops out of point-to-point mode
  - Initialization speed is the advantage

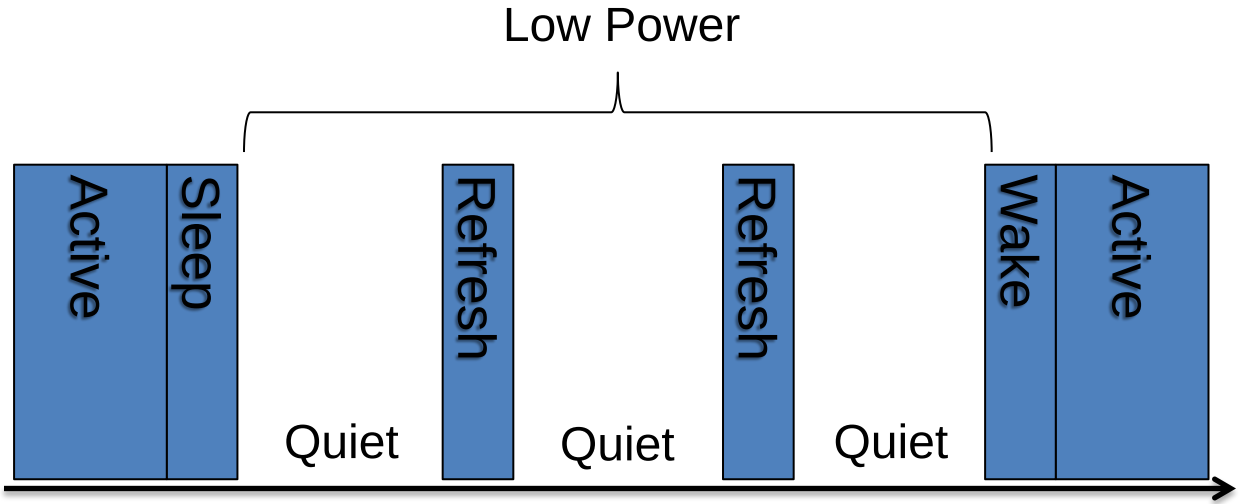
Point-to-point



Upcoming Standards – Energy Efficient Fibre Channel

- **ANSI INCITS T11 Fibre Channel committee is now working on Energy Efficient Fibre Channel – FC-EE**
  - Copper Fibre Channel Interfaces to be based on Energy Efficient Ethernet
  - Developing concepts for Energy Efficient Optical Interfaces

Energy Efficient Fibre Channel Operation



- **Sleep/Refresh/Quiet cycle similar to Energy Efficient Ethernet**
- **Being developed for 32GFC timeframe**

Upcoming Standards

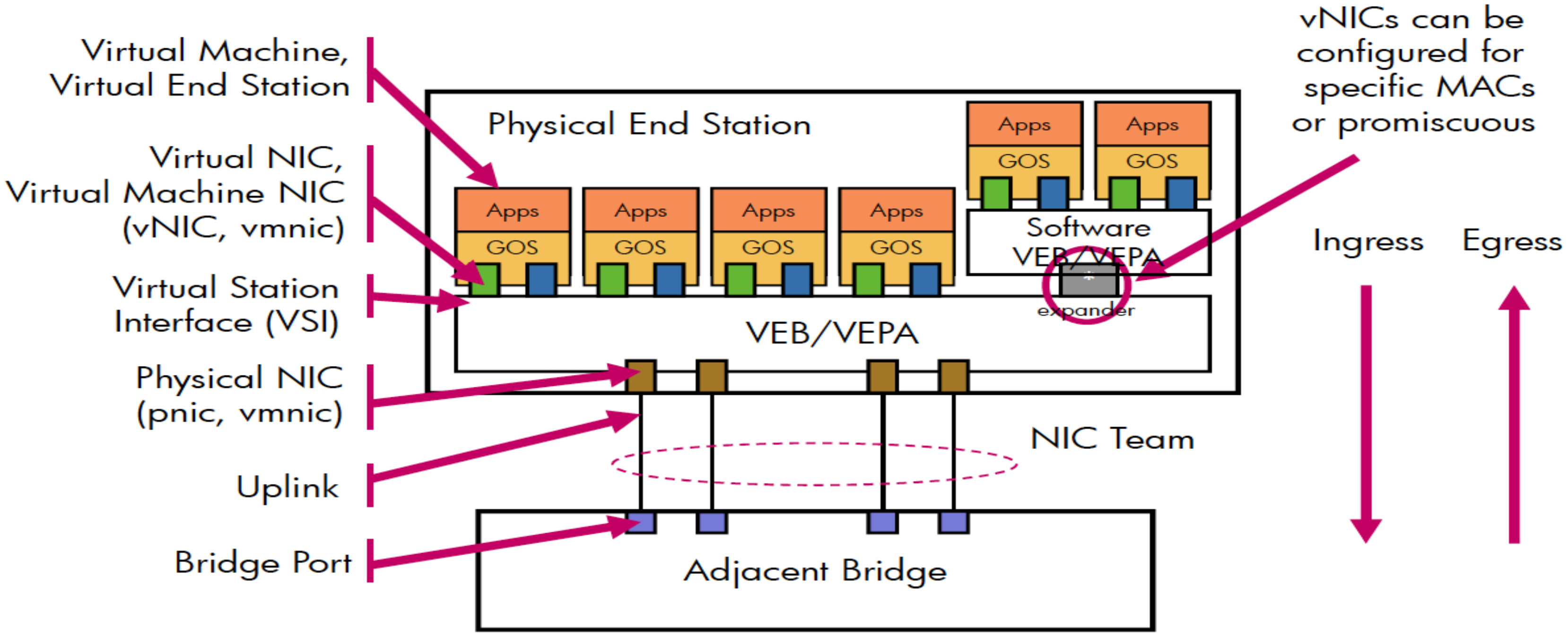
IEEE 802

Creative Commons, Flikr User Daniel*1977

DCB VEB projects

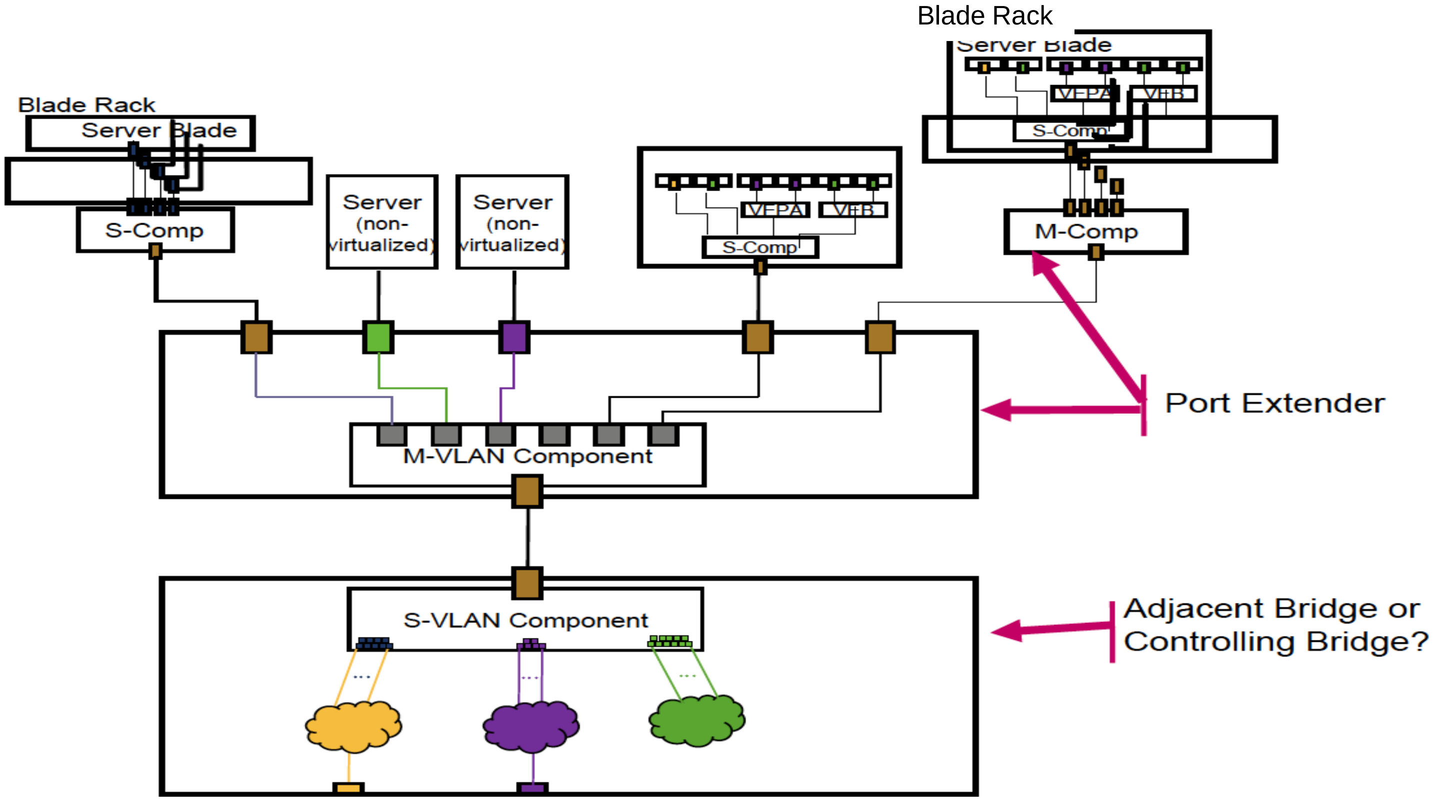
- **IEEE 802.1 DCB has two Virtual Ethernet Bridge projects**
  - 802.1Qbh – Edge Virtual Bridging
  - 802.1BR – Port Bridge Extension
- **VEB is designed for use in a Virtual Machine environment**
  - Provides for performance improvement for local VM traffic
  - Provides a framework which makes for easier VM portability

VEB



Virtual Machine, Virtual End Station
Virtual NIC, Virtual Machine NIC (vNIC, vmnic)
Virtual Station Interface (VSI)
Physical NIC (pnic, vmnic)
Uplink
Bridge Port
Physical End Station
Apps
GOS
Software VEB/VEPA
expander
VEB/VEPA
NIC Team
Adjacent Bridge
vNICs can be configured for specific MACs or promiscuous
Ingress
Egress

Port Extenders



Port Extender

Adjacent Bridge or Controlling Bridge?

And beyond…

Creative Commons, Flikr User bcmacsac1

And beyond…

- **FCIA Roadmap**
  - Defining Fibre Channel speeds beyond 32GFC

QLOGIC
The Ultimate in Performance

FCIA Official Roadmap v13

2010

Page 1 of 4

Fibre Channel Speed Roadmap - v13 (Page 2 of 4)

FC

Product Naming	Throughput (MBps)	Line Rate (GBaud)	T11 Spec Technically Completed (Year)‡	Market Availability (Year)‡
1GFC	200	1.0625	1996	1997
2GFC	400	2.125	2000	2001
4GFC	800	4.25	2003	2005
8GFC	1600	8.5	2006	2008
16GFC	3200	14.025	2009	2011
32GFC	6400	28.05	2012	2014
64GFC	12800	TBD	2015	Market Demand
128GFC	25600	TBD	2018	Market Demand
256GFC	12800	TBD	2021	Market Demand
512GFC	25600	TBD	2024	Market Demand

“FC” used throughout all applications for Fibre Channel infrastructure and devices, including edge and ISL interconnects. Each speed maintains backward compatibility at least two previous generations (I.e., 8GFC backward compatible to 4GFC and 2GFC)

†Line Rate: All “FC” speeds are single-lane serial stream

‡Dates: Future dates estimated

Fibre Channel Speed Roadmap - v13 (Page 3 of 4)

ISL
(Inter-Switch Link)

Product Naming	Throughput (MBps)	Equivalent Line Rate (GBaud)†	Spec Technically Completed (Year) ‡	Market Availability (Year)
10GFC	2400	10.52	2003	2004
20GFC	4800	21.04	TBD	2008‡
40GFC/FCoE	9600	41.225	2010	Market Demand‡
100GFC/FCoE	24000	103.125	2010	Market Demand
400GFC/FCoE	96000	TBD	TBD	Market Demand
1TFC/FCoE	240000	TBD	TBD	Market Demand

ISLs are used for non-edge, core connections, and other high speed applications demanding maximum bandwidth. Except for 100GFC (which follow Ethernet),

†Equivalent Line Rate: Rates listed are equivalent data rates for serial stream methodologies.

‡ Some solutions are Pre-Standard Solutions: There are several methods used in the industry to aggregate and/or "trunk" 2 or more ports and/or data stream lines to achieve the core bandwidth necessary for the application. Some solutions follow Ethernet standards and compatibility guidelines. Refer to the FCoE page 4 for 40GFCoE and 100GFC0E.

Fibre Channel Speed Roadmap - v13 (Page 4 of 4)

FCoE

Product Naming	Throughput (MBps)	Equivalent Line Rate (GBaud)†	Spec Technically Completed (Year)‡	Market Availability (Year)‡
10GFCoE	2400	10.3125	2008	2009
40GFCoE	9600	41.225	2010*	Market Demand
100GFCoE	24000	103.125	2010*	Market Demand

Fibre Channel over Ethernet tunnels FC through Ethernet. For compatibility all 10GFCoE FCFs and CNAs are expected to use SFP+ devices, allowing the use of all standard and non standard optical technologies and additionally allowing the use of direct connect cables using the SFP+ electrical interface. FCoE ports otherwise follow Ethernet standards and compatibility guidelines.

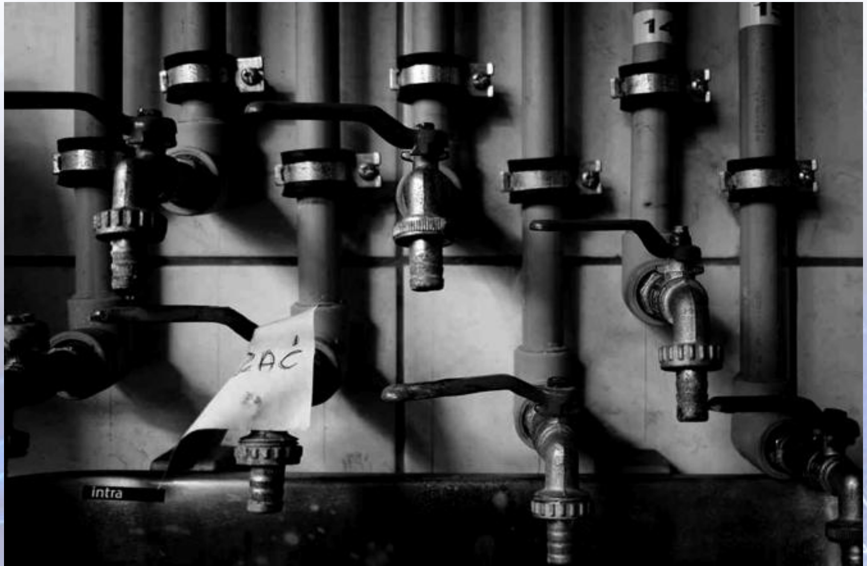
‡Dates: Future dates estimated

* It is expected that 40GFCoE and 100GFCoE based on 2010 standards will be used exclusively for Inter-Switch Link cores, thereby maintaining 10GFCoE as the predominant FCoE edge connection

QLOGIC®

The Ultimate in Performance

Questions?



Creative Commons, Flikr User Daniel*1977