Data Center Bridging

IEEE 802 Tutorial
12th November 2007

Contributors and Supporters

- Hugh Barrass (Cisco)
- Jan Bialkowski (Infinera)
- Bob Brunner (Ericsson)
- Craig Carlson (Qlogic)
- Mukund Chavan (Emulex)
- Rao Cherukuri (Juniper Networks)
- Uri Cummings (Fulcrum Micro)
- Norman Finn (Cisco)
- Anoop Ghanwani (Brocade)
- Mitchell Gusat (IBM)
- Asif Hazarika (Fujitsu Microelectronics)
- Zhi Hern-Loh (Fulcrum Micro)
- Mike Ko (IBM)
- Menu Menuchehry (Marvell)
- Joe Pelissier (Cisco)
- Renato Recio (IBM)
- Guenter Roeck (Teak Technologies)
- Ravi Shenoy (Emulex)
- John Terry (Brocade)
- Pat Thaler (Broadcom)
- Manoj Wadekar (Intel)
- Fred Worley (HP)

Agenda

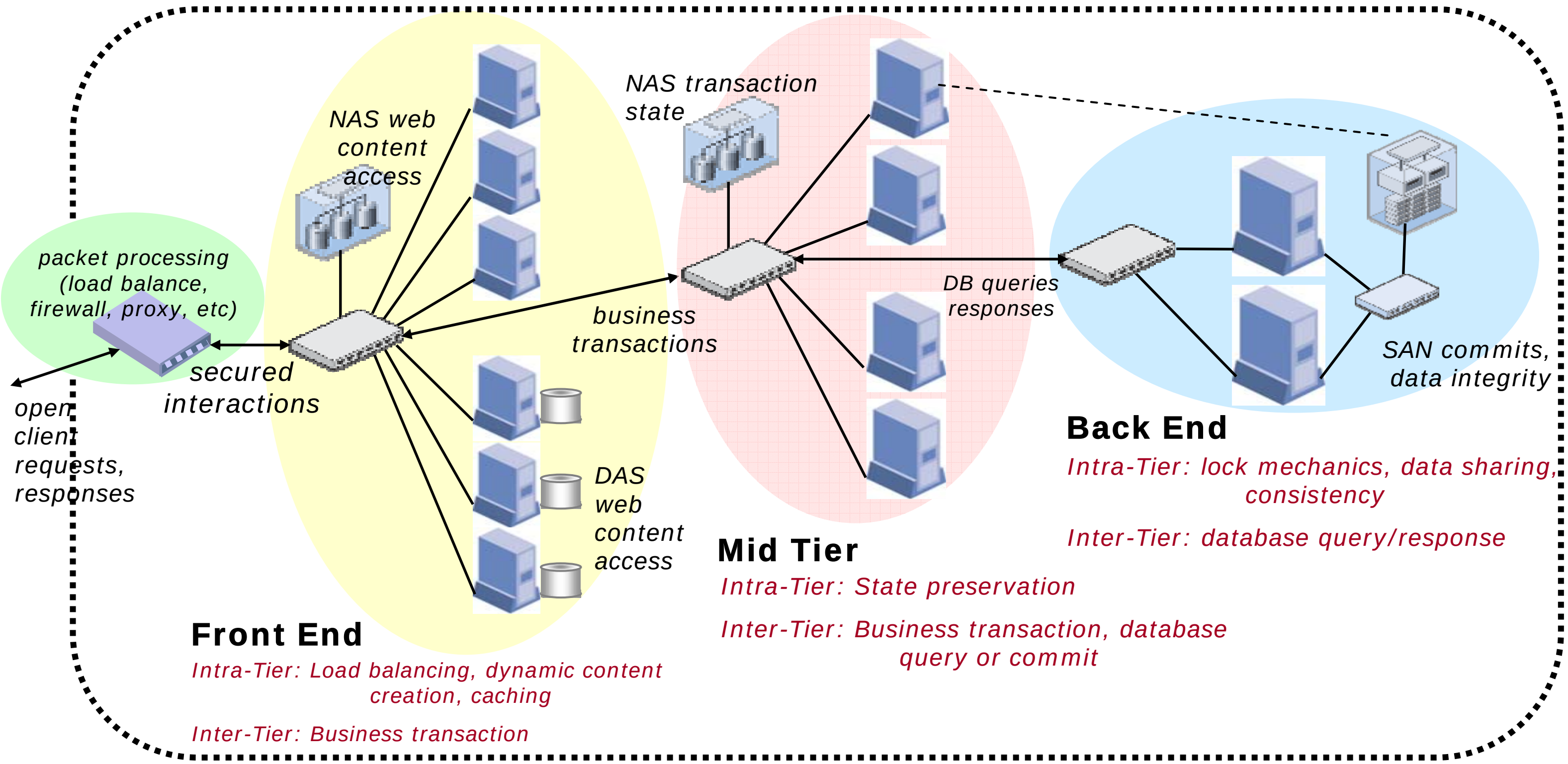
- **Introduction: Pat Thaler**
- **Background: Manoj Wadekar**
- **Gap Analysis: Anoop Ghanwani**
- **Solution Framework: Hugh Barrass**
- **Potential Challenges and Solutions: Joe Pelissier**
- **802.1 Architecture for DCB: Norm Finn**
- **Q&A**

- **802.1Qau Congestion Notification**
  - In draft development
- **802.1Qaz Enhanced Transmission Selection**
  - PAR submitted for IEEE 802 approval at this meeting
- **Priority-Based Flow Control**
  - Congestion Management task group is developing a PAR

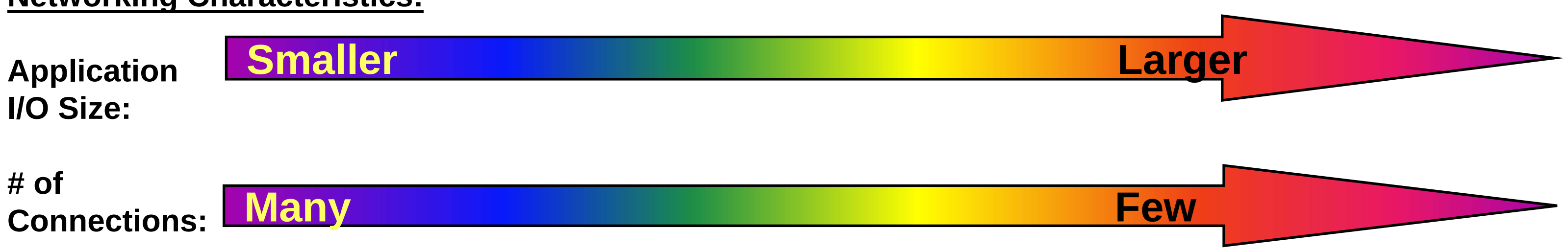
Background: Data Center I/O Consolidation

Manoj Wadekar

Data Center Topology



Networking Characteristics:



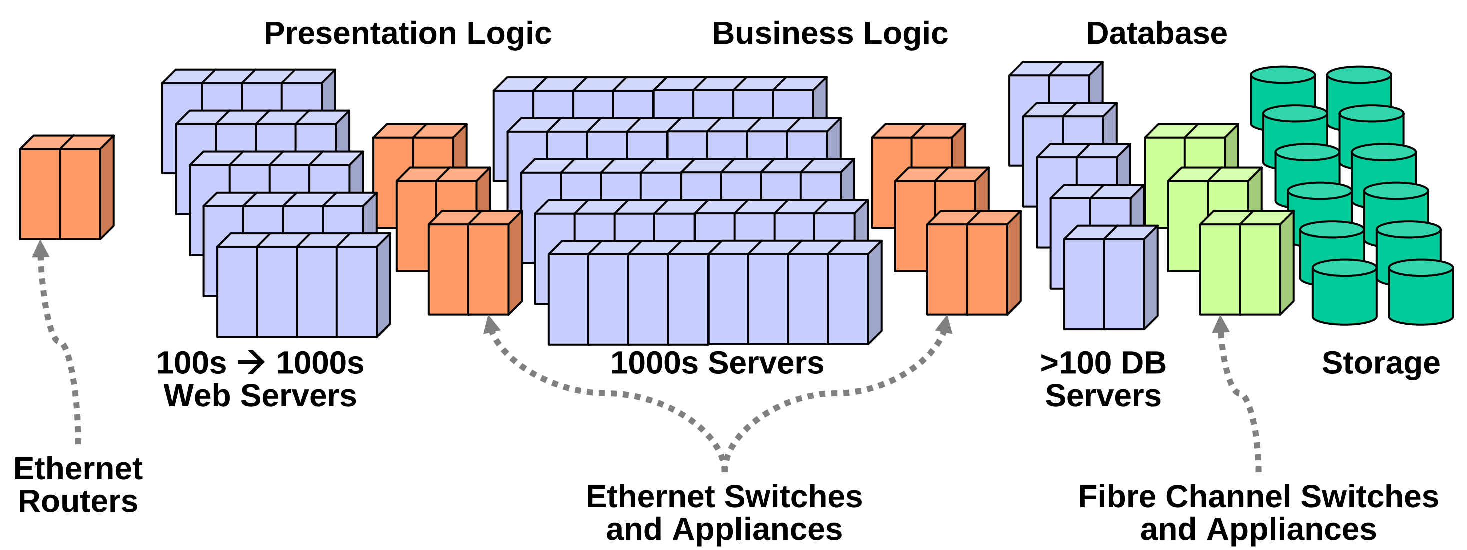
Characteristics of Data Center Market Segments

- **Internet Portal Data Centers**
  - Internet based services to consumers and businesses
  - Concentrated industry, with small number of high growth customers
- **Enterprise Servers Data Centers**
  - Business workflow, Database, Web 2.0/SOA
  - Large enterprise to medium Business
- **HPC and Analytics Data Centers**
  - Large 1000s node clusters for HPC (DOD, Seismic,…)
  - Medium 100s node clusters for Analytics (e.g. FSIs …)
  - Complex scientific and technical

Data Centers Today

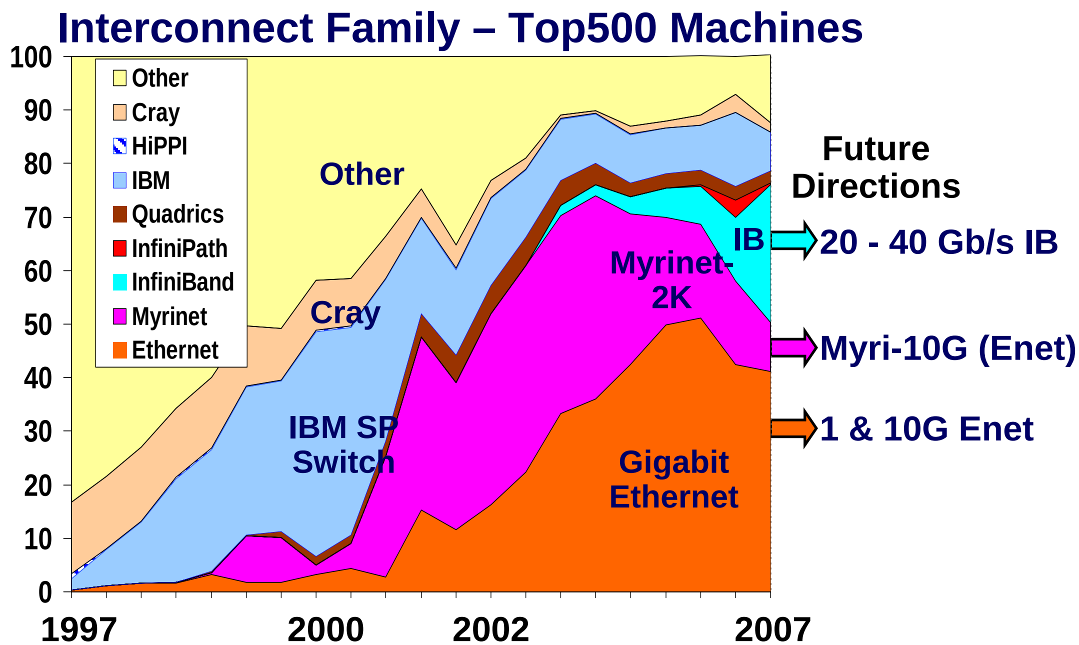
Type	Internet Portal Data Center	Enterprise Data Center	HPCC Data Center
Characteristics	➢SW based enterprises, Non mission critical HW base ➢**10K to 100K servers** ▪ Primary Needs: ➢Low capital cost ➢Reduced power and cooling ➢Configuration solution flexibility	➢Large desktop client base ➢ **100 to 10K servers** ▪ Primary Needs: ➢Robust RAS ➢Security and QOS control ➢Simplified management	➢ Non mission critical HW architecture ➢**100 to 10K servers** ▪ Primary Needs: ➢Low Latency ➢High throughput

Topology examples



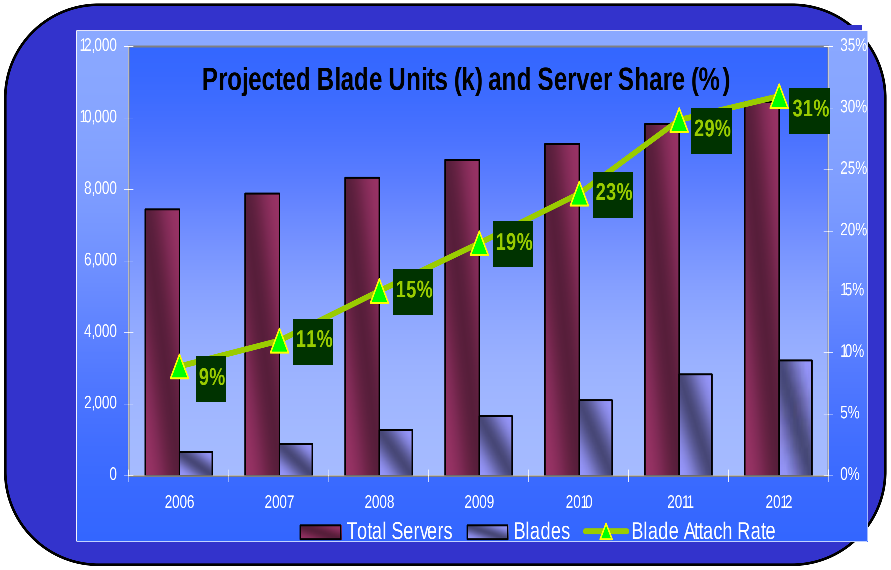
- **Fabric preference is**
  - Low Cost
  - Standard high volume
- **Looks nice on the surface, but…**
  - Lots of cables (cost and power) underneath

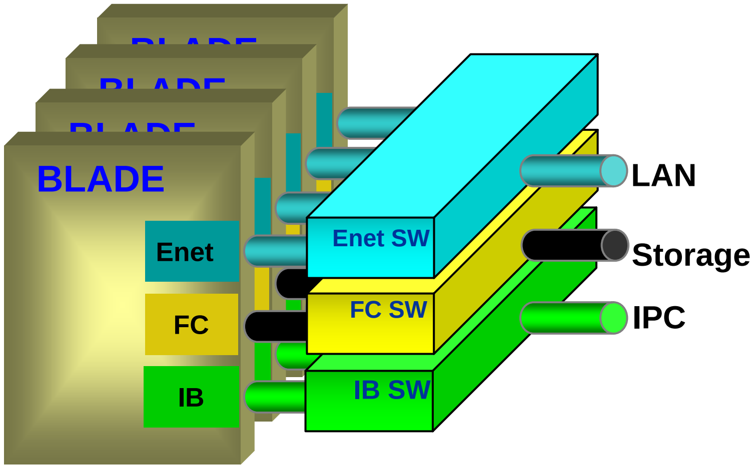
HPC Cluster Network Market Overview - Interconnects in Top 500



- **Standard networks Ethernet & Infiniband (IB) replacing proprietary networks:**
  - IB leading in aggregate performance
  - Ethernet dominates in volume
- **Adoption of 10 Gbps Ethernet in the HPC market will likely be fueled by:**
  - 10 Gbps NICs on the motherboard, 10 GBASE-T/10GBASE-KR
  - Storage convergence over Ethernet (iSCSI, FCoE)
- **There will always be need for special solution for high end (E.g. IB, Myrinet)**
  - But 10 GigE will also play a role in the HPC market in the future
- **Challenges: Need Ethernet enhancements - IB claims better technology**
  - Low Latency, traffic differentiation, "no-drop" fabric, multi-path, bi-sectional bandwidth

Data Center: Blade Servers





- **Blade Servers are gaining momentum in market**
- **Second Generation Blade Servers in market**
- **Provides server, cable consolidation**
- **Ethernet default fabric**
- **Optional fabrics for SAN and HPC**

- **Challenges: IO Consolidation is strong requirement for Blade Servers:**
  - Multiple fabrics, mezzanine cards,
  - Power/thermal envelope,
  - Management complexity
  - Backplane complexity

REF: Total Servers: Worldwide and Regional Server 2007-2012 Forecast, April, 2007, IDC
Blade Servers: Worldwide and US Blade Server 2006-2010 Forecast, Oct'06, IDC

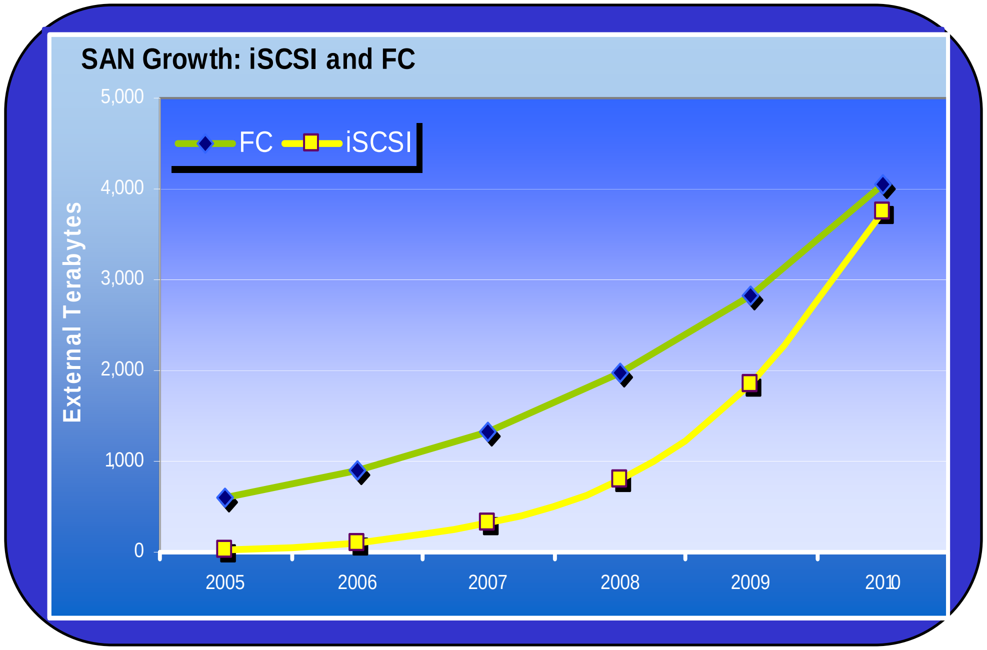
Data Center: Storage Trends

Lot of people using SAN:

- **Networked Storage growing 60% year**
- **FC is incumbent and growing: Entrenched enterprise customers**
- **iSCSI is taking off…: SMB and Greenfield deployment - choice driven by targets**

Storage convergence over Ethernet:

- **iSCSI for new SAN installations**
- **FC tunneled over Ethernet (FCoE) for expanding FC SAN installation**



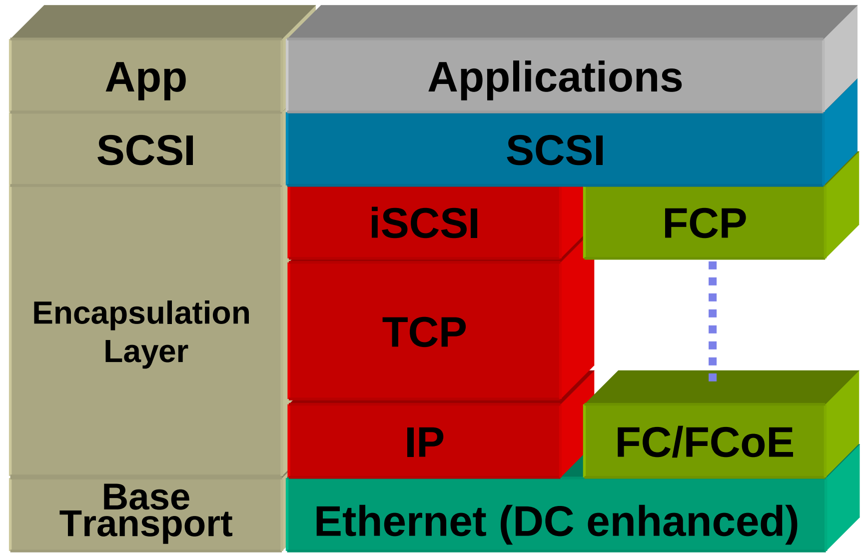
IDC Storage and FC HBA analyst reports, 2006

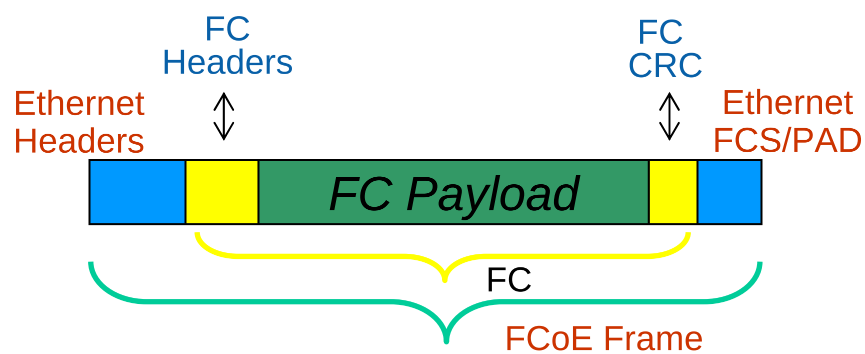
Challenges:

- **Too many ports, fabrics, cables, power/thermal**
- **Need to address FC as well as iSCSI**

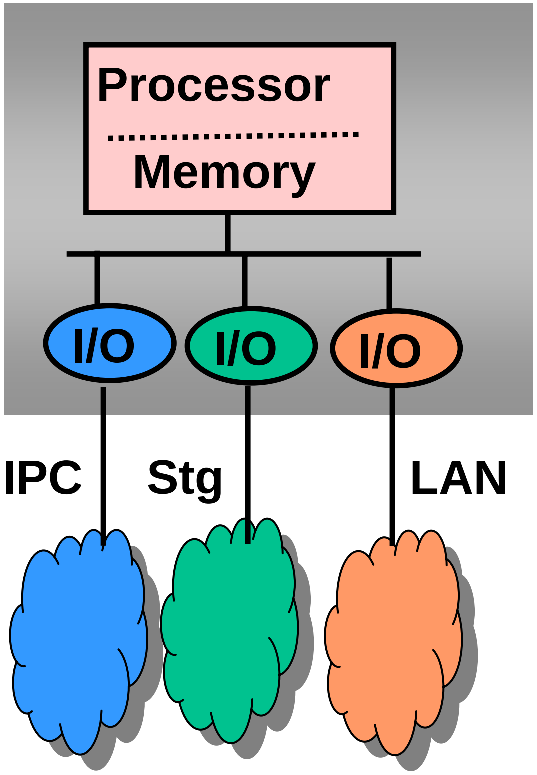
SAN Protocols

- <u>FCoE</u>:
  - FC Tunneled through Ethernet
  - Addresses customer investment in legacy FC storage
  - Expects FC equivalent no-drop behavior in underlying Ethernet interconnect
  - Needs Ethernet enhancements for link convergence and "no-drop" performance
- <u>iSCSI</u>:
  - SCSI over TCP/IP that provides reliability
  - High speed protocol acceleration solutions benefit from reduced packet drops

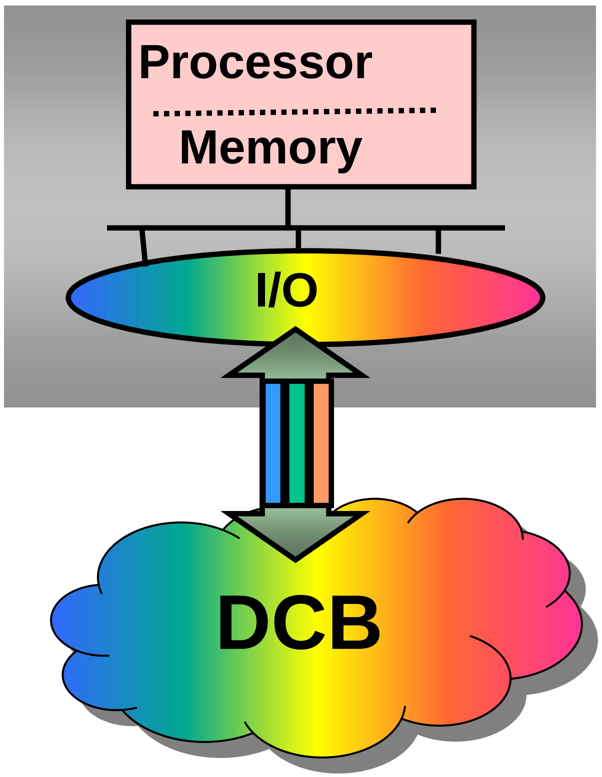




Fabric convergence needs:



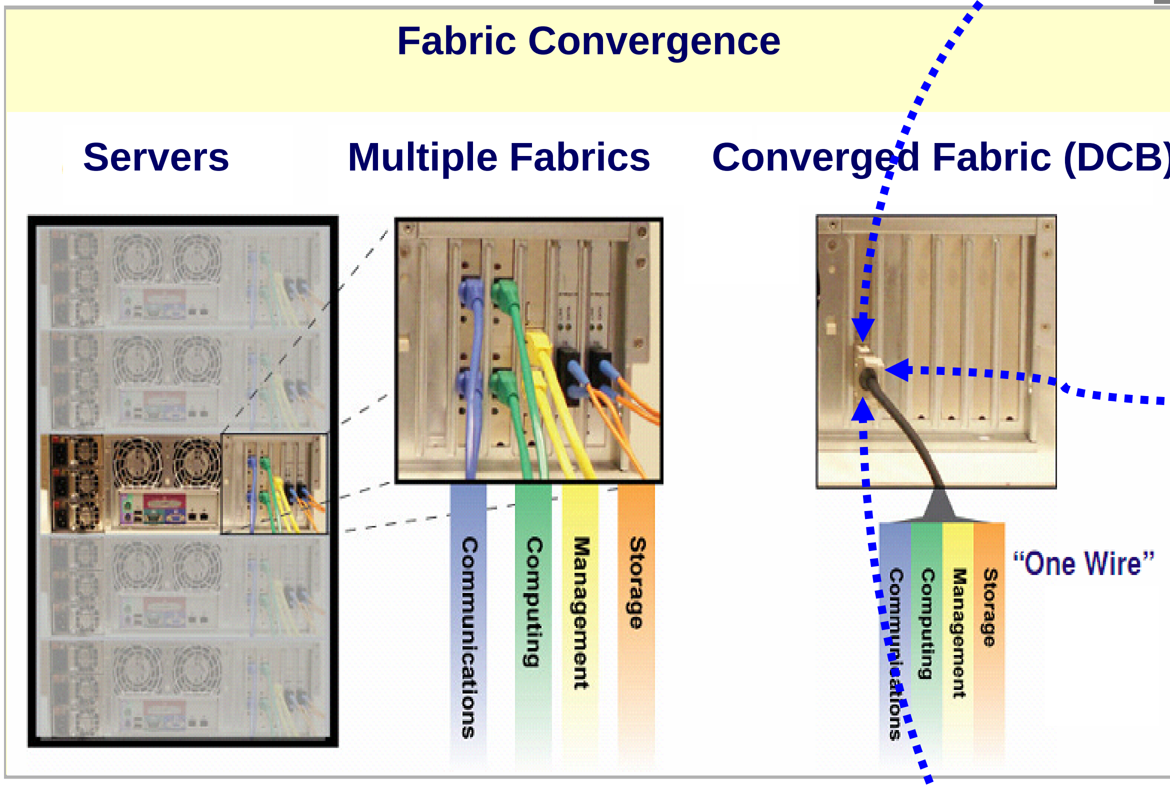
- **Traffic Differentiation:**
  - LAN, SAN, IPC traffic needs differentiation in converged fabric
- **Lossless Fabric:**
  - FC does not have transport layer – retransmissions are at SCSI!
  - iSCSI acceleration may benefit from lossless fabric too
- **Seamless deployment:**
  - Backward compatibility
  - Plug and play
- **Ethernet needs these enhancements to be true converged fabric for Data Center**



- **What is DCB?**
  - Data Center Bridging provides Ethernet enhancements for Data Center needs (Storage, IPC, Blade Servers etc.)
  - Enhancements apply to bridges as well as end stations
- **Why DCB?**
  - DC market demands converged fabric
  - Ethernet needs enhancements to be successful converged fabric of choice
- **Scope of DCB:**
  - Should provide convergence capabilities for Data Center – short range networks

Data Center Bridging → Value Proposition

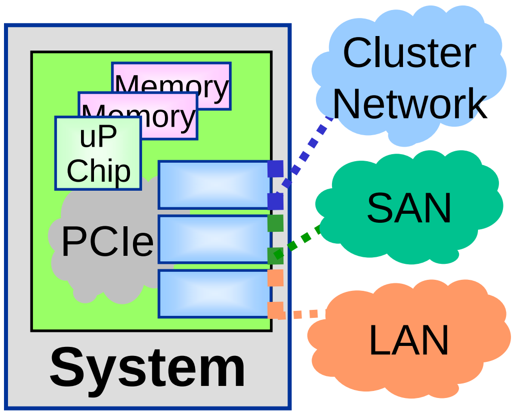
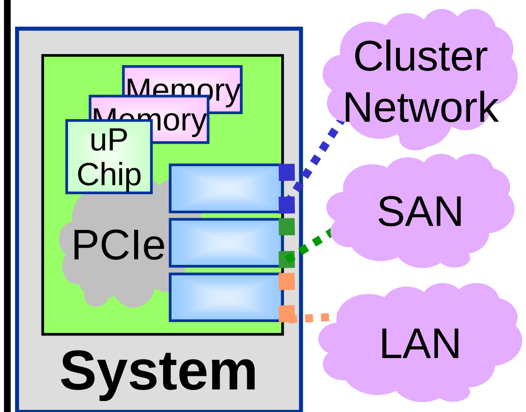
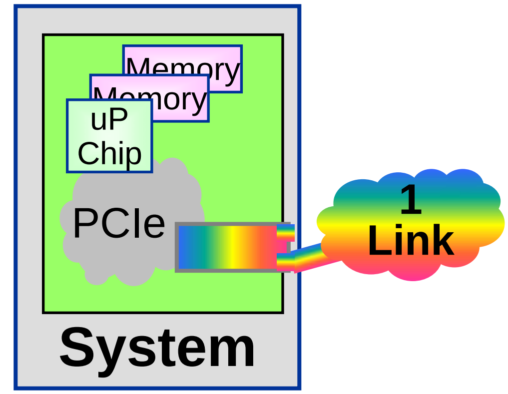
Simpler Management
Single physical fabric to manage. Simpler to deploy, upgrade and maintain.



Lower Cost
Less adapters, cables & switches
Lower power/thermals

Improved RAS
Reduced failure points, time, misconnections, bumping.

Comparison of convergence levels

	No convergence (dedicated)	Converged Fabric Management	DCB
			
Number of Fabric Types	**2-3**	**1**	**1**
IO interference	**No**	**No**	**Yes**
Technologies Managed	**3**	**1 to 2**	**1 to 2**
HW cost	**3x adapters**	**3x adapters**	**1x adapter**
RAS	**More HW**	**More HW**	**Least HW**
Cable mess	**3-4x**	**3-4x**	**1x**

DCB provides IO Consolidation:

- **Lower CapEx, Lower OpEx, Unified management**

Needs standards-based Ethernet enhancements:

- **Need to support multiple traffic types and provide traffic differentiation**
- **"No-drop" option for DC applications**
- **Deterministic network behavior for IPC**
- **Does not disrupt existing infrastructure**
  - Should allow "plug-and-play" for enhanced devices
  - Maintain backward compatibility for legacy devices

Gap Analysis: Data Center Current Infrastructure

Anoop Ghanwani

- **Data Center Bridging (DCB) requirements**
- **What do 802.1 and 802.3 already provide?**
- **What can be done to make bridges more suited for the data center?**

- **Single physical infrastructure for different traffic types**

- **Traffic in existing bridged LANs**
  - Tolerant to loss – apps that care about loss use TCP
  - QoS achieved by using traffic classes; e.g. voice vs data traffic

- **Data center traffic has different needs**
  - Some apps expect lossless transport; e.g. FCoE
    - This requirement cannot be satisfactorily met by existing standards

- **Building a converged network**
  - Multiple apps with different requirements; e.g.
    - Voice (loss- & delay-sensitive)
    - Storage (lossless, delay-sensitive)
    - Email (loss-tolerant)
    - …
  - Assign each app type to a traffic class
  - Satisfy the loss/delay/BW requirements for each traffic class

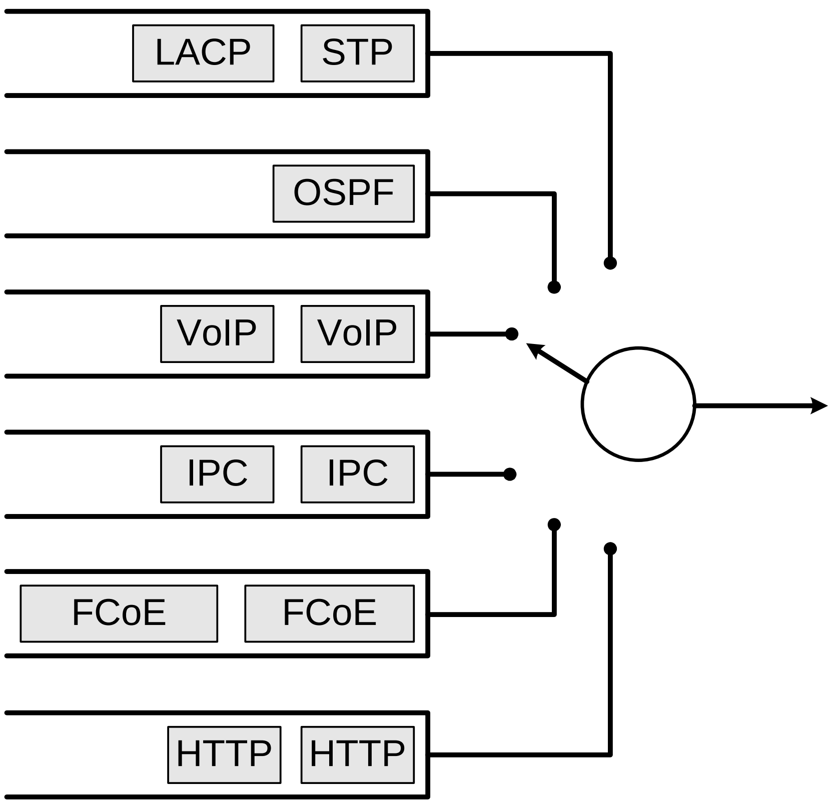
Relevant 802.1/802.3 Standards

- **For traffic isolation and bandwidth sharing**
  - Expedited traffic forwarding - 802.1p/Q
    - 8 traffic classes
    - Default transmission selection mechanism is strict priority
      - Others are permitted but not specified
    - Works well in for traffic in existing LANs
      - Control > Voice > Data

- **For achieving lossless behavior**
  - Congestion Notification - P802.1Qau [in progress]
    - Goal is to reduce loss due to congestion in the data center
    - Works end to end – both ends must be within the L2 network
    - Needed for apps that run directly over L2 with no native congestion control
  - PAUSE - 802.3x
    - On/off flow control
    - Operates at the link level

- **Can we build a converged data center network using existing standards?**

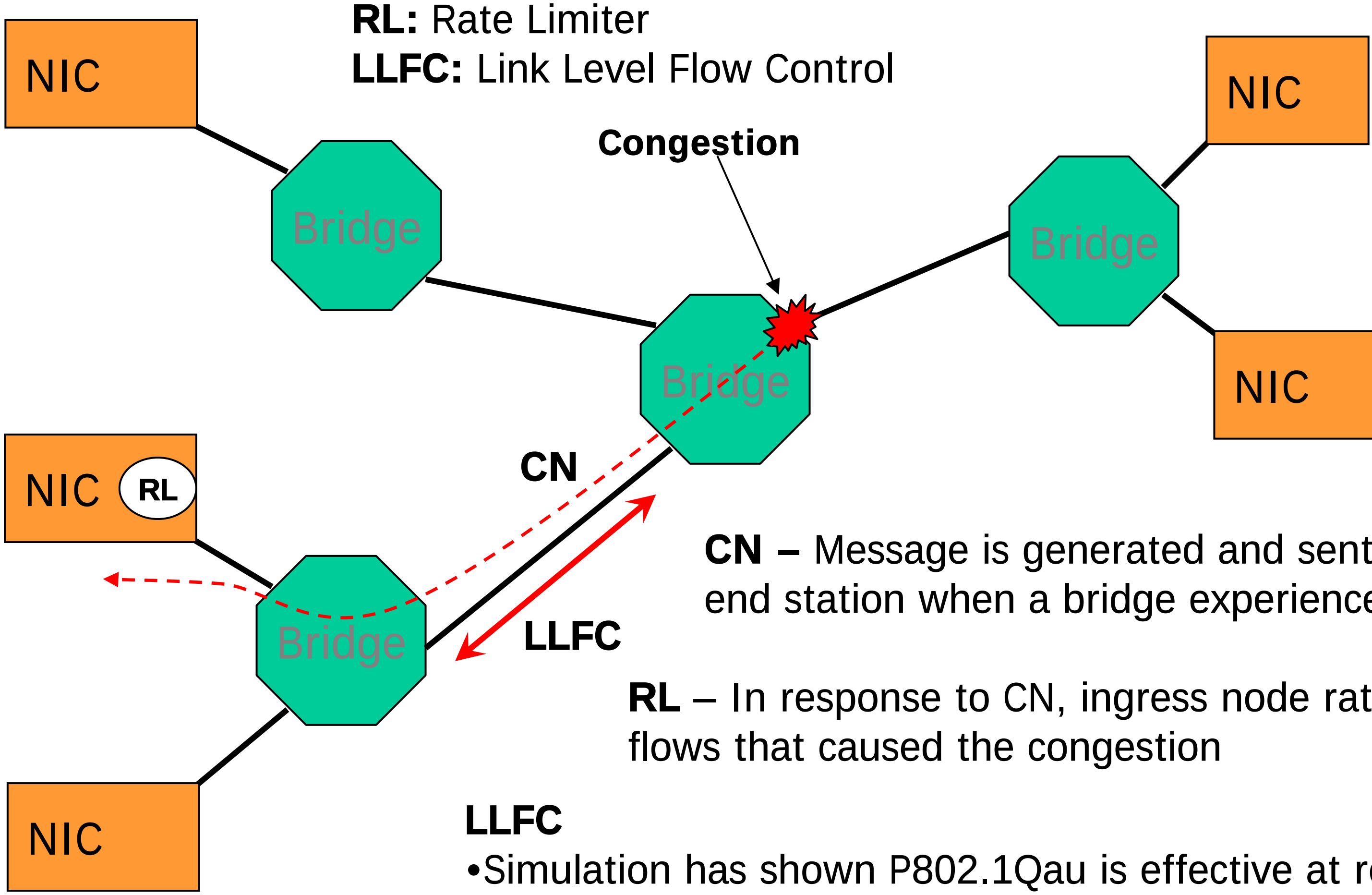
Transmission Selection for DCB



- **802.1p/Q's strict priority is inadequate**
  - Potential starvation of lower priorities
  - No min BW guarantee or max BW limit
  - Operator cannot manage bandwidth
    - 1/3 for storage, 1/10 for voice, etc.
- **Transmission selection requirements**
  - Allow for minimal interference between traffic classes
    - Congestion-managed traffic will back off during congestion
    - Should not result in non-congestion-managed traffic grabbing all the BW
  - Ideally, a "virtual pipe" for each class
- **Benefits regular LAN traffic as well**
  - Many proprietary implementations exist
- **Need a standardized behavior with a common management framework**

Achieving Lossless Transport Using P802.1Qau and 802.3x

CN: Congestion Notification
RL: Rate Limiter
LLFC: Link Level Flow Control



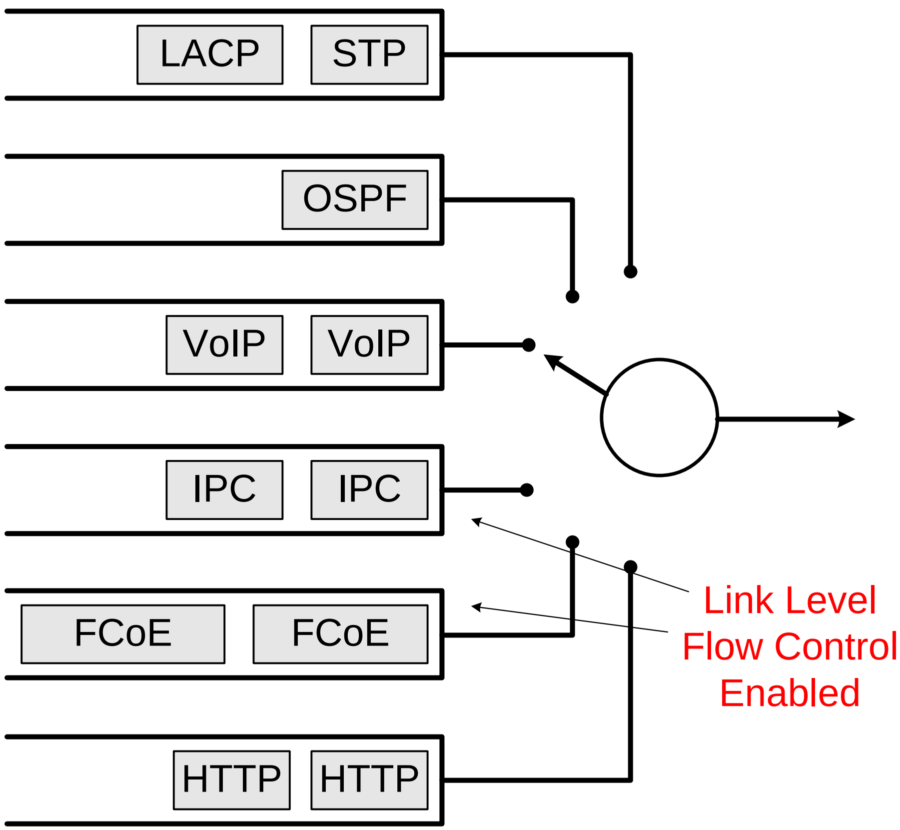
CN – Message is generated and sent to ingress end station when a bridge experiences congestion

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

LLFC

- Simulation has shown P802.1Qau is effective at reducing loss
- Packet drop still possible if multiple high-BW source go active simultaneously
- <u>LLFC is the only way to guarantee zero loss</u>

More on Link Level Flow Control Using 802.3x



- **802.3x is an on/off mechanism**
  - All traffic stops during the "off" phase
- **802.3x does not benefit some traffic**
  - Tolerant to loss; e.g. data over TCP
  - Low BW, high priority ensure loss is relatively rare; e.g. voice
- **802.3x may be detrimental in some cases**
  - Control traffic; e.g. LACP & STP BPDUs
  - Increases latency for interactive traffic
- **As a result most folks turn 802.3x off**
- **Need priority-based link level flow control**
  - Should only affect traffic that needs it
  - Ability to enable it per priority
  - **Not simply 8 x 802.3x PAUSE!**
  - Provides a complete solution when used together with P802.1Qau

Summary

- **The goal of DCB is to facilitate convergence in the data center**
  - Apps with differing requirements on a single infrastructure
- **Need improvements to existing standards to be successful**
  - Flexible, standards-based transmission selection
  - End to end congestion management
  - Enhanced link level flow control
- **Networks may contain DCB- and non-DCB-capable devices**
  - Discovery/management framework for DCB-specific features
- **The technology exists in many implementations today**
  - Standardization will promote interoperability and lower costs

Solution Framework: Data Center Bridging

Hugh Barrass

Data Center Bridging Framework

- **Solution must be bounded**
  - **- No "leakage" to legacy or non-DCB systems**
  - **- Can optimize behavior at "edges"**

- **Aim for "no drop" ideal**
  - **- Using reasonable architectures**
  - **- No congestion spreading, or latency inflation**

(Covered in 802.1Qau) ***(in discussion)***

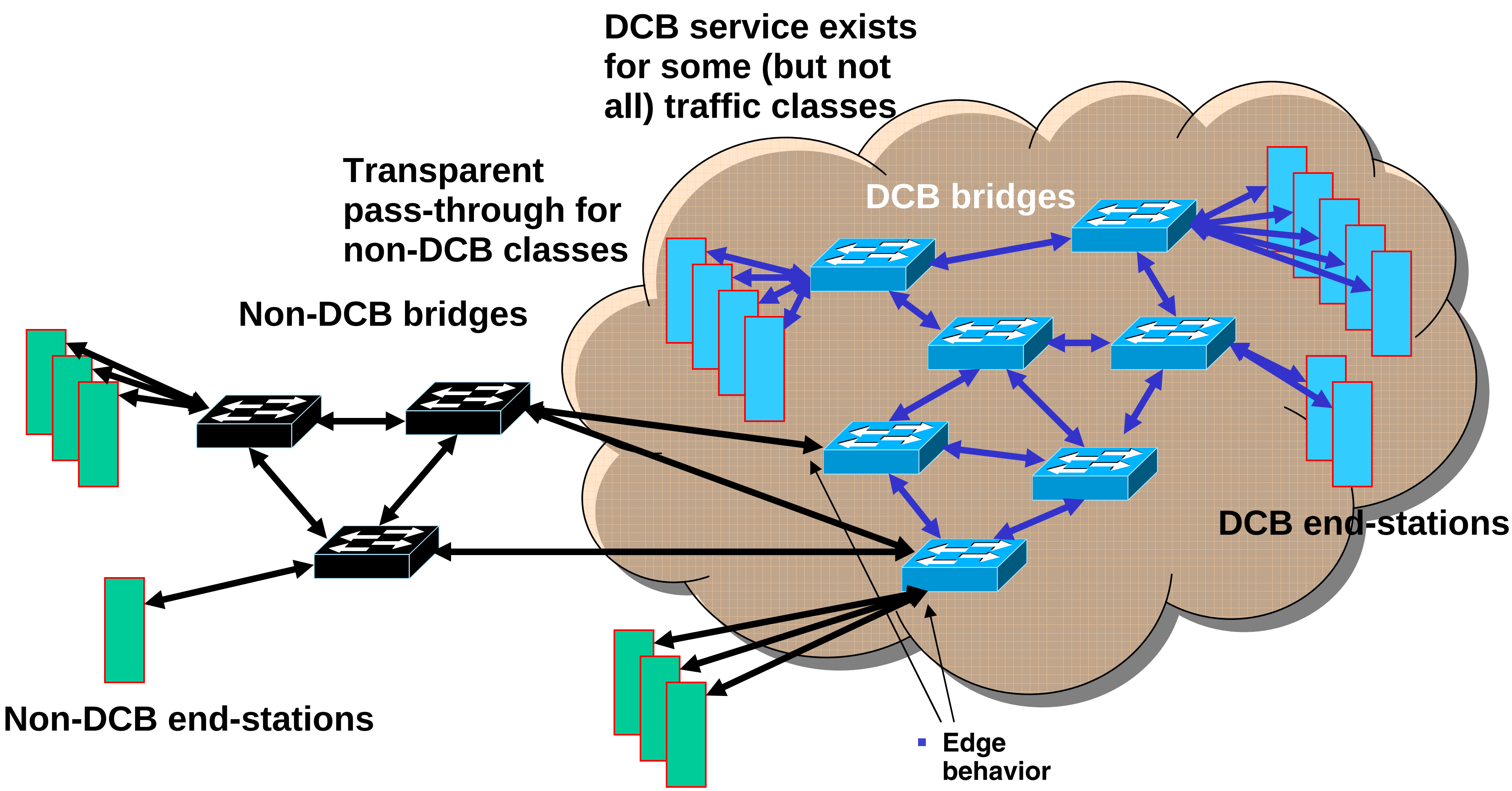
- **Relies on congestion notification with flow control backup**
  - **- Two pronged attack required for corner cases**

- **Support for multiple service architectures** ***(PAR under consideration)***
  - **- Priority queuing with transmission selection**

(Covered in 802.1Qau and extensions)

- **Discovery & capability exchange**
  - **- Forms & defines the cloud; services supported**
  - **- Management views & control**

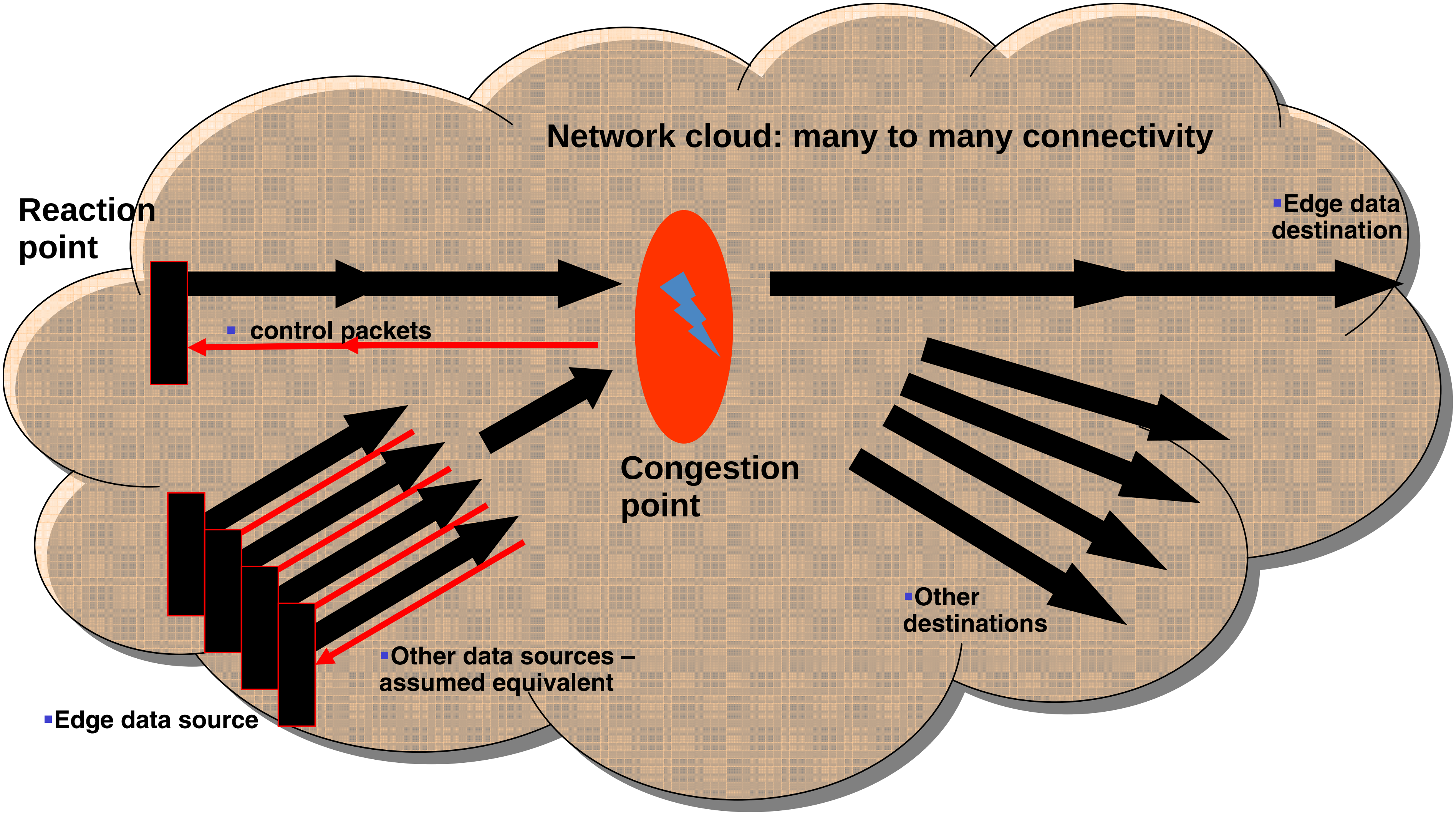
DCB devices in cloud, edge behavior



Introduction of DCB devices in key parts of network offers significant advantages

- DCB cloud is formed, only DCB devices allowed inside
  - **Function using LLDP defines cloud and operating parameters**
- If source, destination & path all use DCB then optimal behavior
- At edge of cloud, edge devices restrict access to CM classes
  - **DCB specific packets do not "leak out"**
- Optimal performance for small, high b/w networks
  - **E.g. datacenter core**

1 congestion point, 1 reaction point considered



Operation of congestion management

- CM is an end-to-end function
  - **Data from multiple sources hits a choke point**
  - **Congestion is detected, ingress rate is slowed**
  - **Control loop matches net ingress rate to choke point**

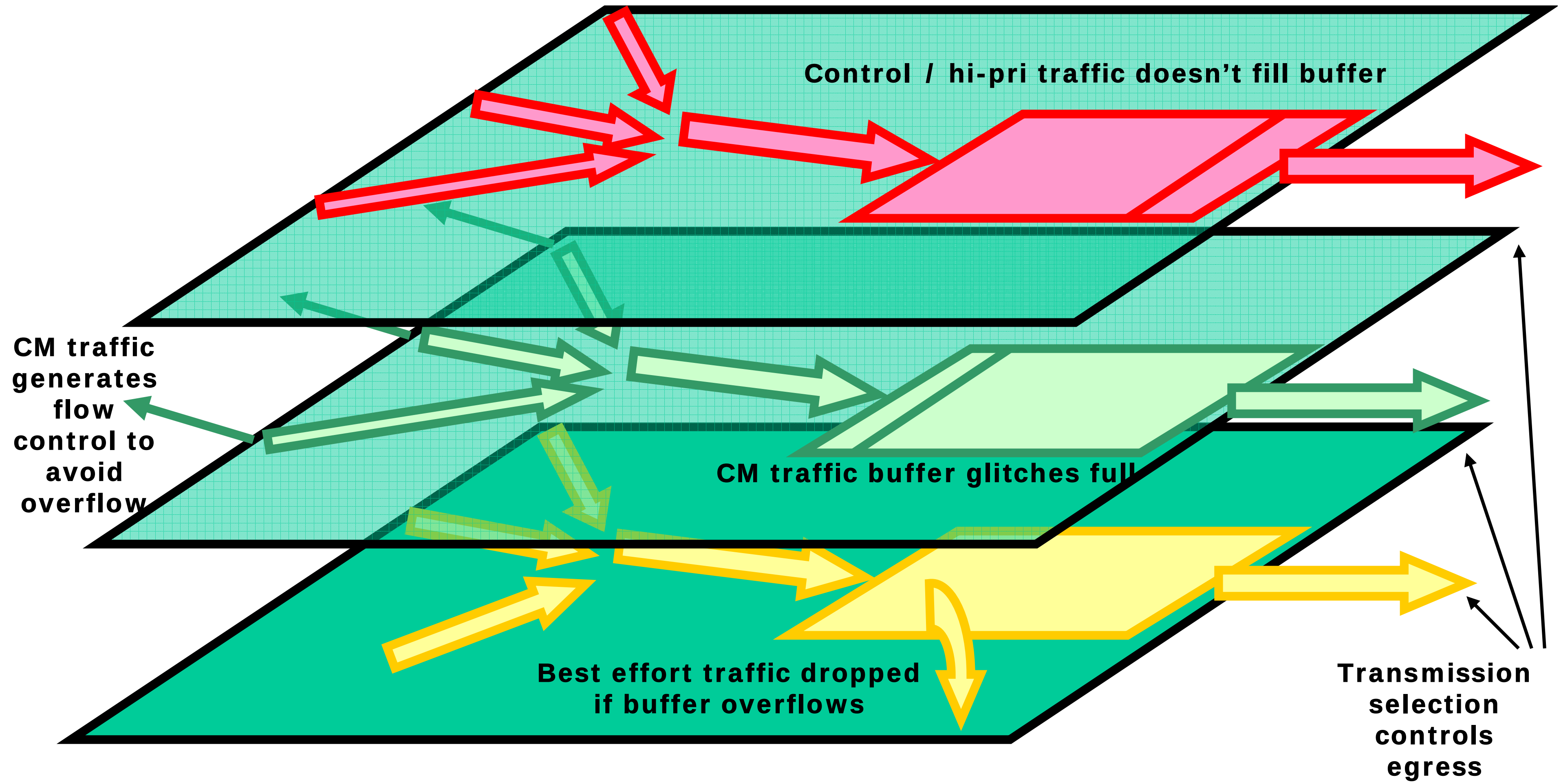
- Simple cases – 2 or more similar ingress devices; 1 choke point
  - **Each end device gets 1/N b/w**
  - **Minimal buffer use at choke point**

- More complex cases with asymmetric sources, multi-choke etc.
  - **CM is always better than no CM!**

- Corner cases still cause problems – particularly for buffering
  - **b/w spikes overfill buffers before CM can take control**
  - **Requires secondary mechanism for safety net**

Priority based flow control

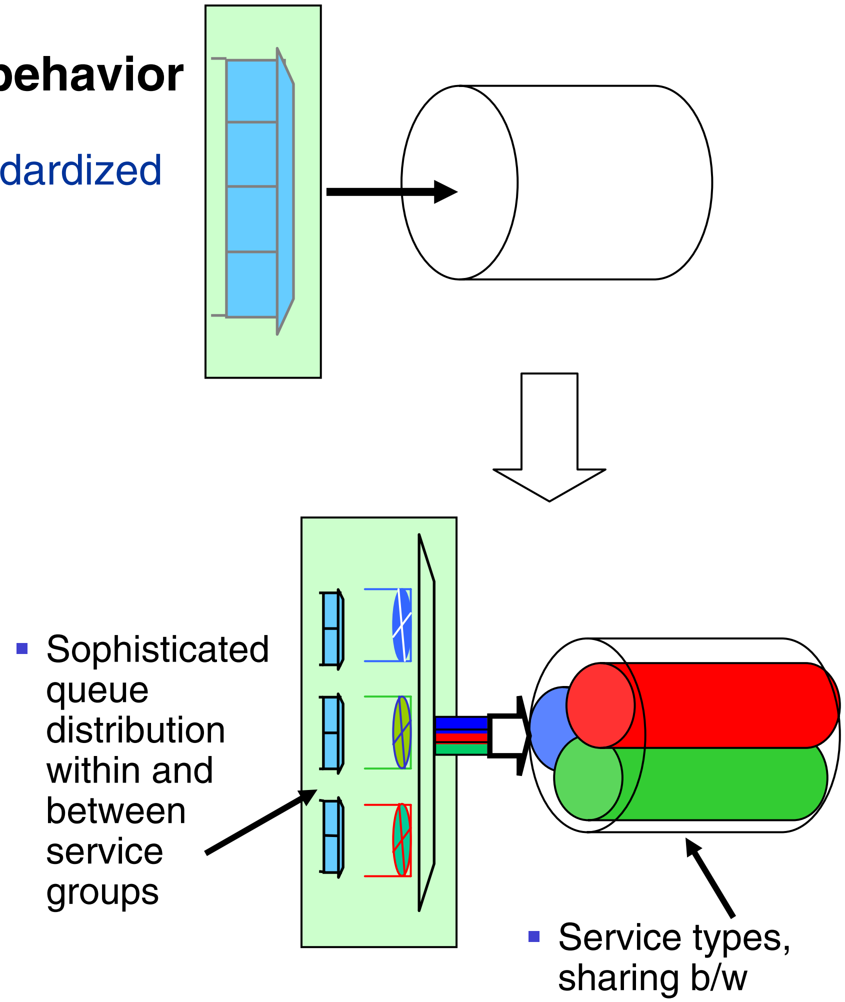
- **In corner cases, and edge conditions, CM cannot react quickly enough to prevent queue overflow. For certain traffic types the packet loss is unacceptable.**



Multiple service architectures using 802.1p

- **Current standard defines only simplistic behavior**
  - Strict priority defined
  - More complex behavior ubiquitous, but not standardized
- **Enhancements needed to support more sophisticated, converged, networks**

- **Within the defined 8 code points**
  - Define grouping & b/w sharing
  - Queue draining algorithms defined to allow minimum & maximum b/w share; weighted priorities etc.
- **Common definition & management interface allow stable interoperability**



DCB Framework

- **Congestion Management**
  - Persistent Congestion: 802.1Qau (approved TF)
  - Transient Congestion: Priority based Flow Control (under discussion)

- **Traffic Differentiation**
  - Enhanced Transmission Selection: 802.1Qaz (proposed PAR)

- **Discovery and Capability Exchange**
  - Covered for 802.1Qau
  - Additional enhancements may be needed for other DCB projects

In summary

- Complete datacenter solution needs whole framework
  - Components rely on each other
  - Utility is maximized only when all are available

- Fully functional Data Center Bridging solves the problem
  - Allows convergence of datacenter network
  - Currently discrete networks per function
  - Eliminates niche application networks

- High bandwidth, low latency, “no drop” network…
- … alongside scalability & simplicity of 802.1 bridging
  - Supports rapid growth to meet datacenter demands
  - Net benefits for users and producers

Challenges and Solutions: DCB

Joe Pelissier

- **Congestion Spreading:**
  - Priority based Flow Control causes congestion spreading, throughput melt-down
- **Deadlocks:**
  - Link level Flow Control can lead to deadlocks
- **DCB not required for all applications/products**
  - DCB functionality applicable to DC networks, but..
  - May become unnecessary burden for some products
- **Compatibility with existing devices**
  - Data Centers contain applications that tolerate “drop”
  - Legacy devices and DCB devices interoperability challenges

Requirements of a “Lossless” Fabric

- **Many IPC and storage protocols do not provide for low–level recovery of lost frames**
  - Done by higher level protocol (e.g. class driver or application)
  - Recovery in certain cases requires 100’s to 1000’s of ms

- **Excessive loss (e.g. loss due to congestion vs. bit errors) may result in link resets, redundant fabric failovers, and severe application disruption**

- **These protocols therefore require (and currently operate over) a flow control method to be enabled**
  - With 802.3x PAUSE, this implies a separate fabric for these protocols
    - Since traditional LAN/WAN traffic is best served without PAUSE
  - These protocols are not “broken”
    - They are optimized for layer 2 data center environments
      - Such as those envisioned for DCB
    - Maintaining this optimization is critical for broad market adoption

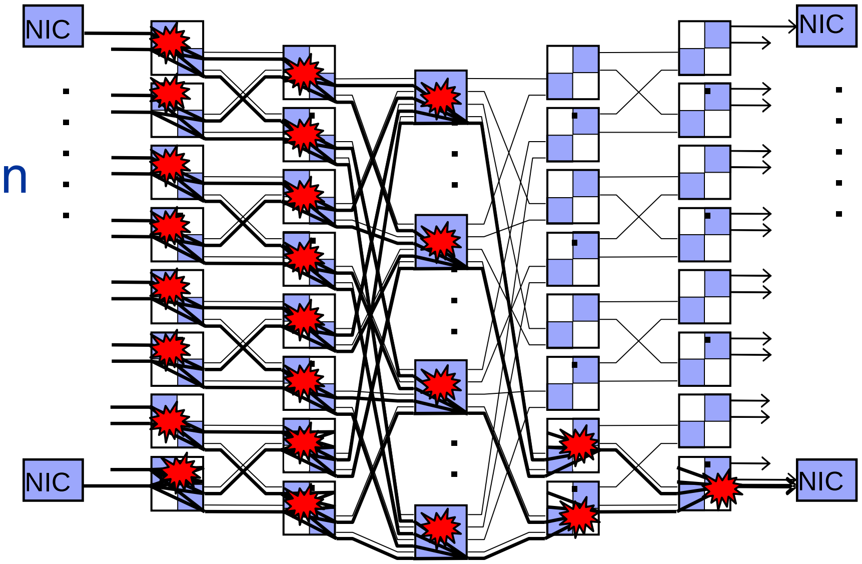
- **Many IPC and storage protocols are not "congestion aware"**
  - Do not expect frame loss due to congestion
    - Potential for congestion caused frame loss highly application sensitive
    - Traffic may be very bursty – very real potential for fame loss
  - Do not respond appropriately
    - Huge retransmission attempts
    - Congestion collapse
- **Storage and IPC applications can tolerate low frame loss rates**
  - Bit errors do occur
- **Frame loss due to congestion requires different behavior compared to frame loss due to bit errors**
  - Back-off, slow restart, etc. to avoid congestion collapse

Congestion Notification and Frame Loss

- **Tremendous effort has been expended in developing a congestion notification scheme for Bridged LANs**
  - Simulation efforts indicate that these schemes are likely to dramatically reduce frame loss
  - However, frame loss not sufficiently eliminated
    - Especially under transitory congestion events and in topologies that one would reasonably expect for storage and IPC traffic
    - Congestion Notification does reduce the congestion spreading side effect of flow control
  - Therefore a supplemental flow control mechanism that prevents frame loss is viewed as a requirement for successful deployment of storage and IPC protocols over Bridged LANs
    - These protocols operate over a small portion of the network (i.e. the portion to be supported by DCB).
    - A simple method is sufficient

Congestion Spreading

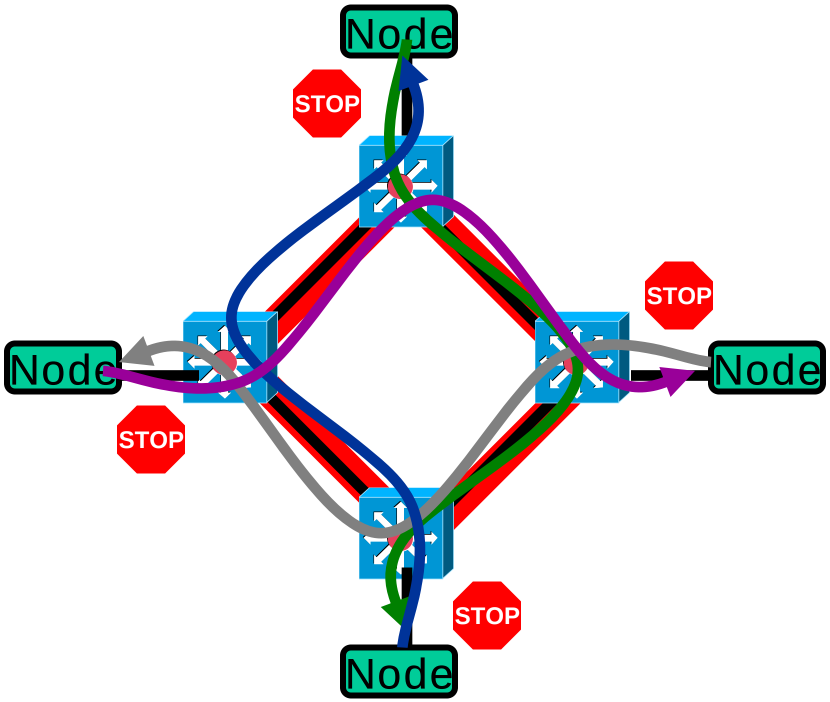
- **Multiple hops in a flow controlled network can cause congestion that spreads throughout the network**
  - Major issue with 802.3x PAUSE



- **Effects mitigated by:**
  - Limited to flow controlled DCB region of network
  - Limited to traffic that traditionally operates over flow controlled networks
    - E.g. IPC and storage
  - Isolated from and independent of traffic that is prone to negative impacts of congestion spreading
    - Priority Based Flow Control
      (and selective transmission) create "virtual pipes" on the link
      - Flow control is enabled (or not) on each "pipe" as appropriate for the traffic type

Deadlock (1)

- **Flow control in conjunction with MSTP or SPB can cause deadlock**
  - See "Requirements Discussion of Link Level-Flow Control for Next Generation Ethernet" by Gusat et al, January '07 (au-ZRL-Ethernet-LL-FC-requirements-r03)
- **To create deadlock, all of the following conditions must occur:**
  - A cyclic flow control dependency exists
  - Traffic flows across all corners of the cycle
  - Sufficient congestion occurs on *all* links in the cycle *simultaneously* such that each bridge is unable to permit more frames to flow
- **At this point, all traffic on the affected links halts until frames age out**
  - Generally after one second
- **Feeder links also experience severe congestion and probable halt to traffic flow**
  - A form of congestion spreading



Deadlock (2)

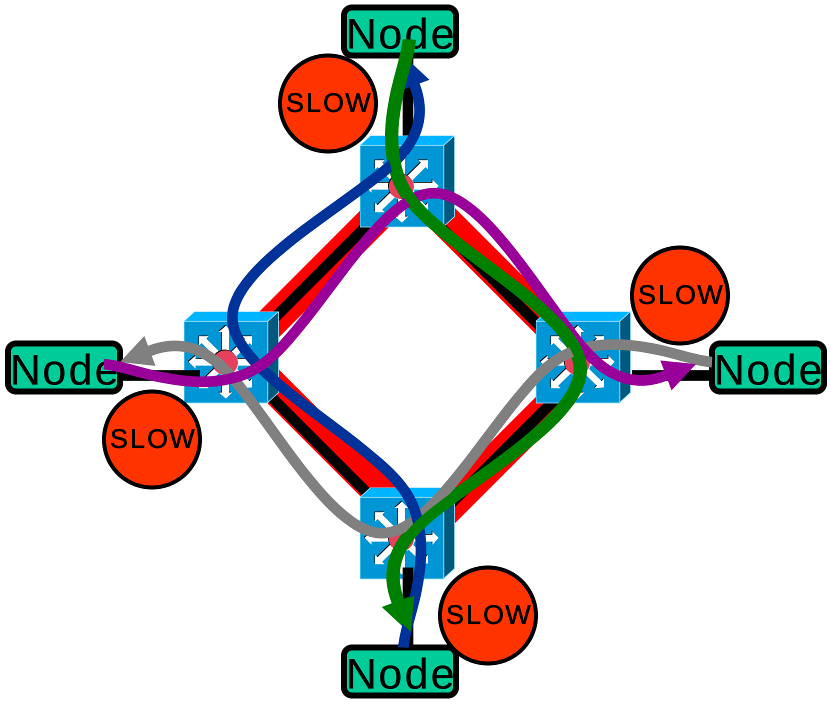
- **However:**
  - The probability of an actual deadlock is very small
  - Deadlock recovers due to frame discard mandated by existing Maximum Bridge Transit Delay (clause 7.7.3 in IEEE Std 802.1D™-2004)

- **Low Deadlock Probability:**
  - Congestion Notification renders sufficient congestion at all necessary points in the network highly unlikely
  - Many Data Center topologies (e.g. Core / Edge or fat tree) do not contain cyclic flow control dependencies
  - MSTP or SPB routing may be configured to eliminate deadlock
    - Edges would not be routed as an intermediate hop between core bridges
  - Traffic flow frequently is such that deadlock cannot occur
    - E.g. storage traffic generally travels between storage arrays and servers
    - Insufficient traffic passes through certain “corners” of the loop to create deadlock

Deadlock (3)

- **The previous assumptions are not without demonstration**
  - IPC and storage networks are widely deployed in mission critical environments:
    - Most of these networks are flow controlled

- **What happens if a deadlock does occur?**
  - Traffic not associated with the class (i.e. different priority level) is unaffected
  - Normal bridge fame lifetime enforcement frees the deadlock
    - Congestion Notification kicks in to prevent reoccurrence



Summary

- **Congestion Spreading:**
  - Two-pronged solution: End-to-end congestion management using 802.1Qau and Priority based Flow Control
    - 802.1Qau reduces both congestion spreading and packet drop
    - Priority based Flow Control provides “no-drop” where required
- **Deadlocks:**
  - Deadlock had been shown to be a rare event in existing flow controlled networks
  - Addition of Congestion Notification further reduces occurrence
  - Normal Ethernet mechanisms resolve the deadlock
  - Non flow controlled traffic classes unaffected
- **CM not required for all applications/products**
  - Work related to Data Center should be identified as Data Center Bridging (following “Provider Bridging”, “AV Bridging”)
  - Provides clear message about usage model
- **Compatibility with existing devices**
  - Capability Exchange protocol and MIBs will be provided for backward compatibility

802.1 Architecture for DCB

Norm Finn

Subclause 8.6 The Forwarding Process

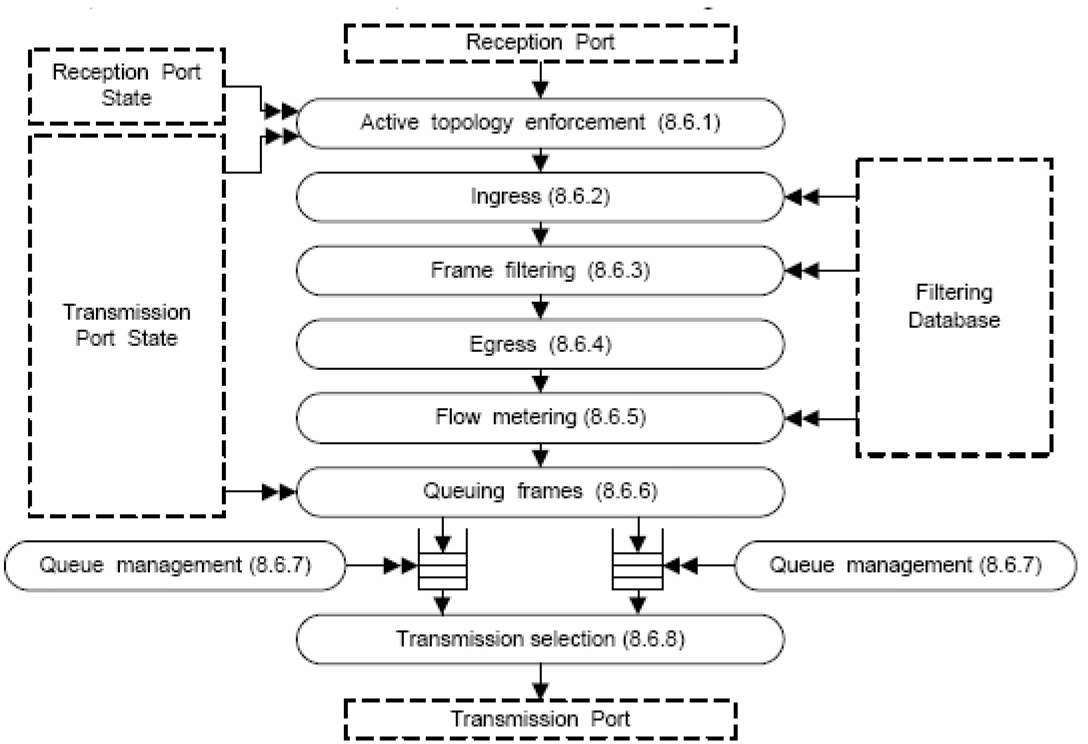


Figure 8-9—Forwarding Process functions

Subclause 22 CFM in systems

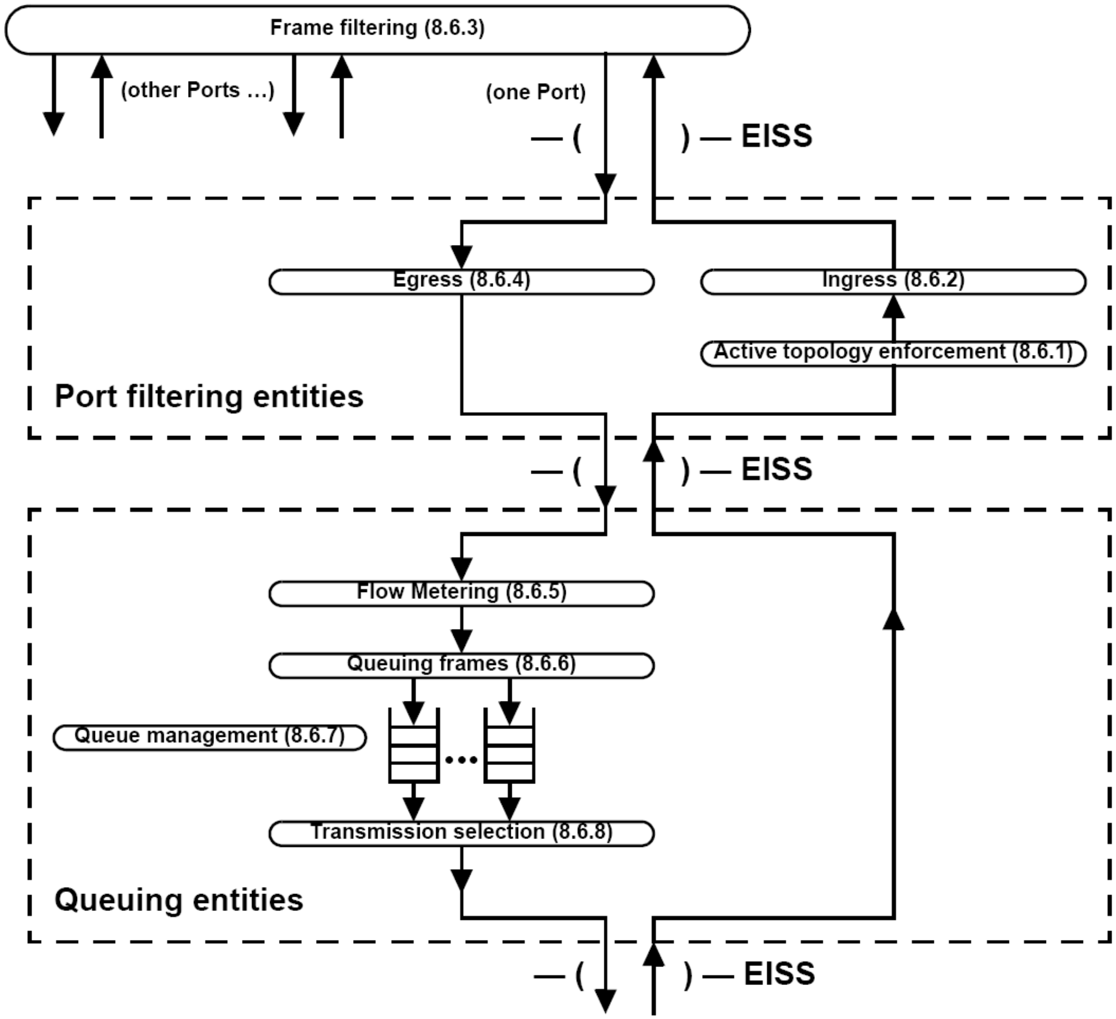


Figure 22-2—Alternate view of Forwarding process

Subclause 22
CFM in systems

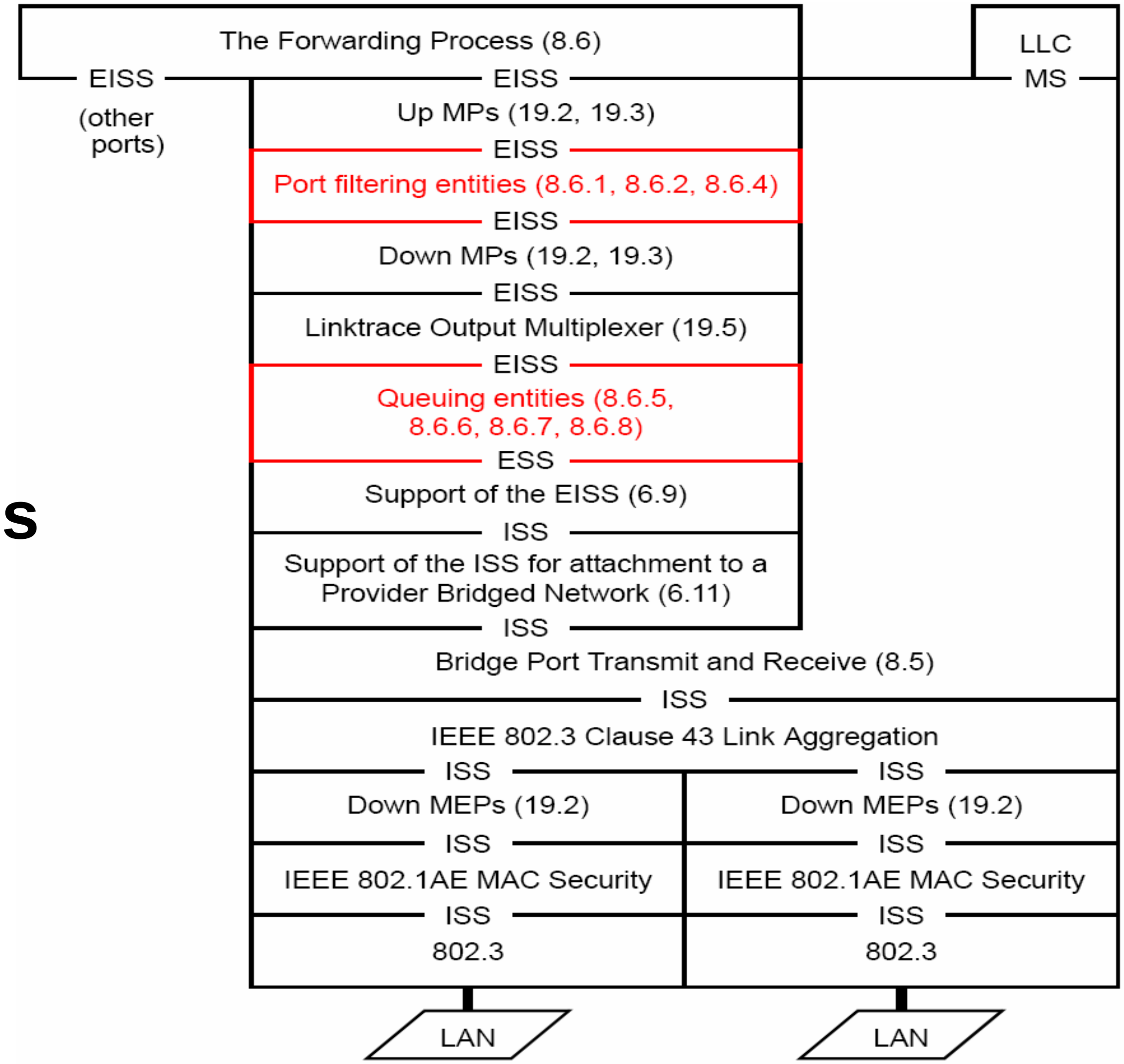


Figure 22-8—Maintenance Point placement relative to other standards

Subclause 31 Congestion notification entity operation

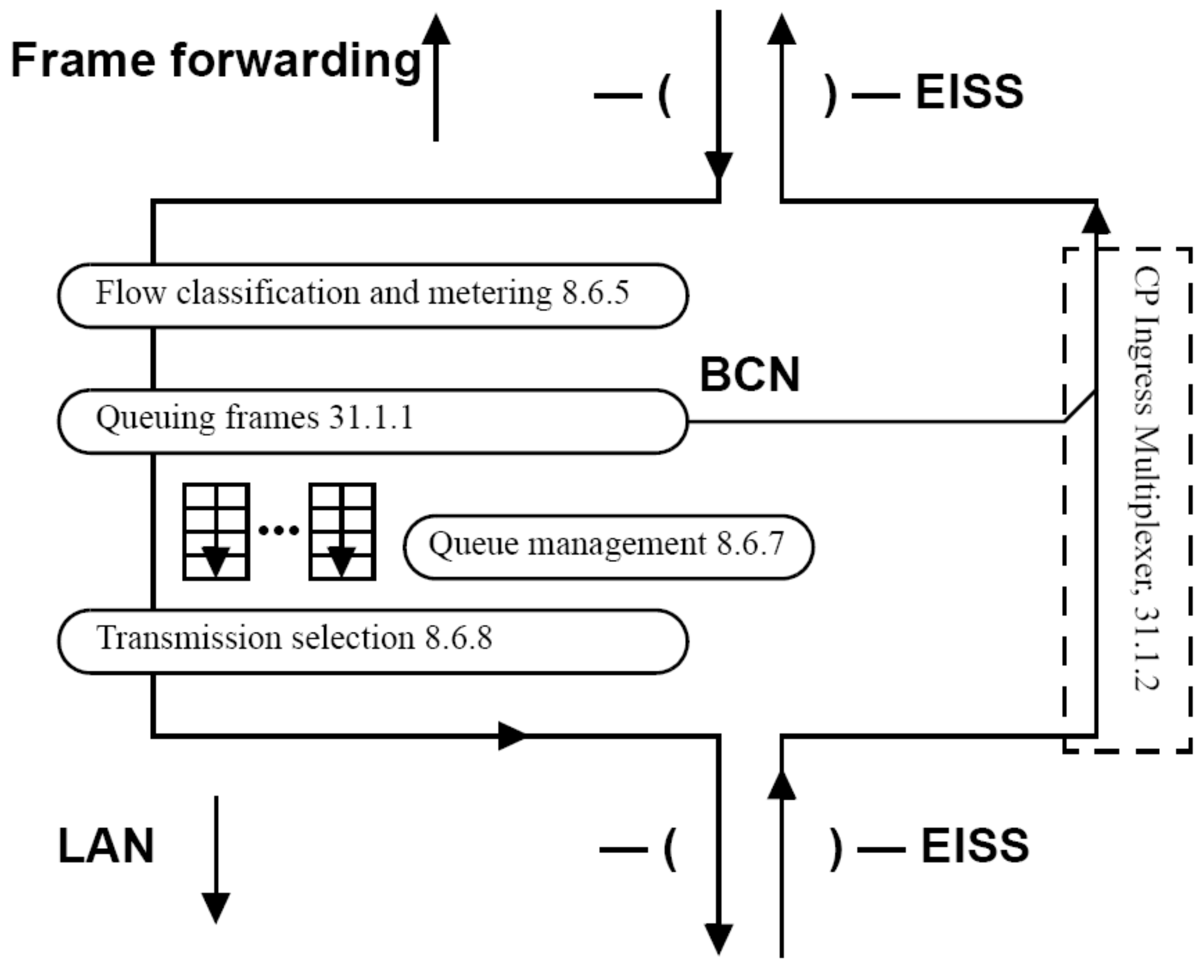


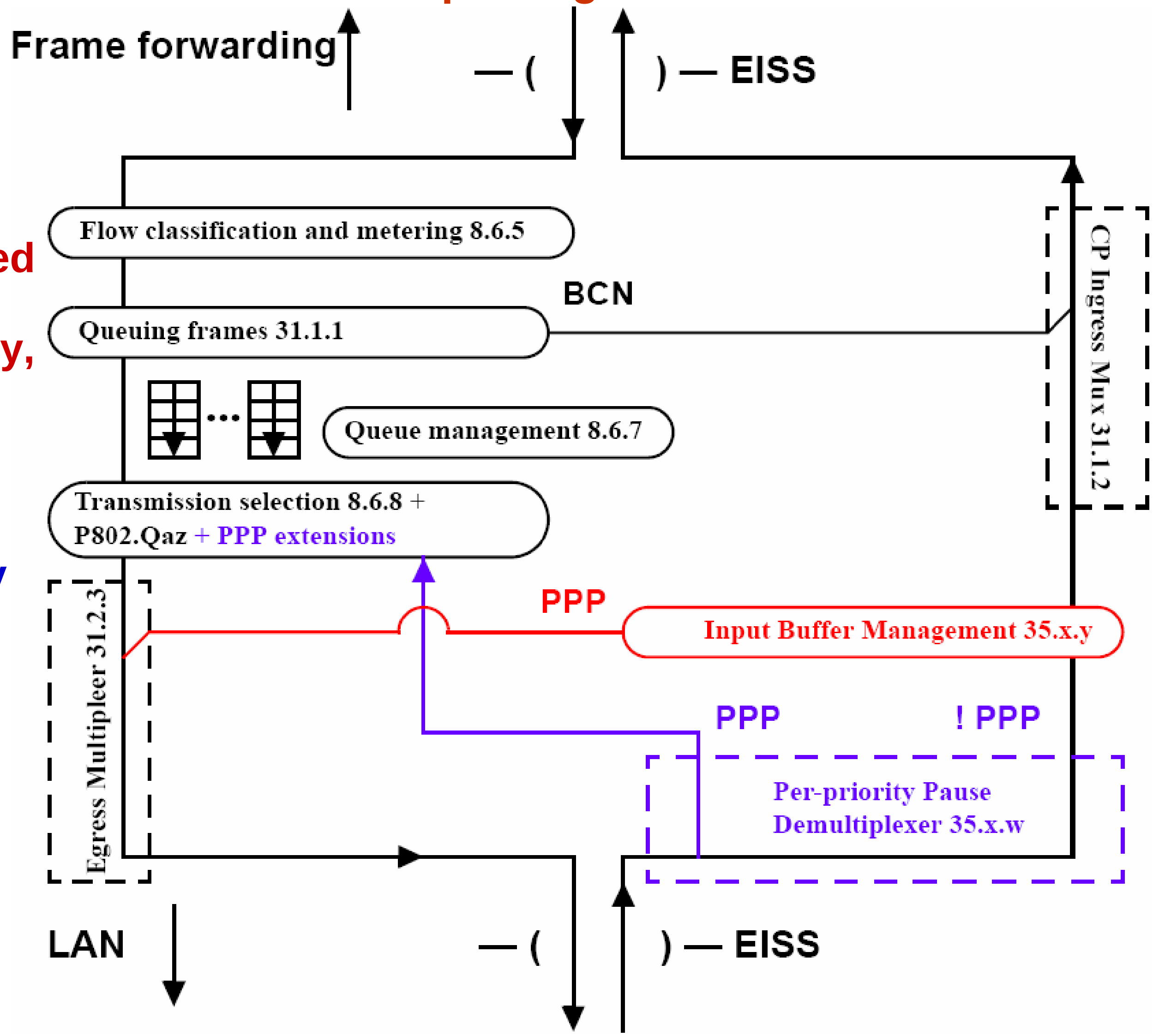
Figure 31-1—A Congestion Point in a Bridge Port

Proposed extension for PPP in Bridges

RED: PPP generation based on space available in per-port, per-controlled-priority, input buffers.

BLUE: PPP reaction, only for BCN controlled priority queues.

Subclause TBD – modified queuing entities



Subclause 31 Congestion notification entity operation

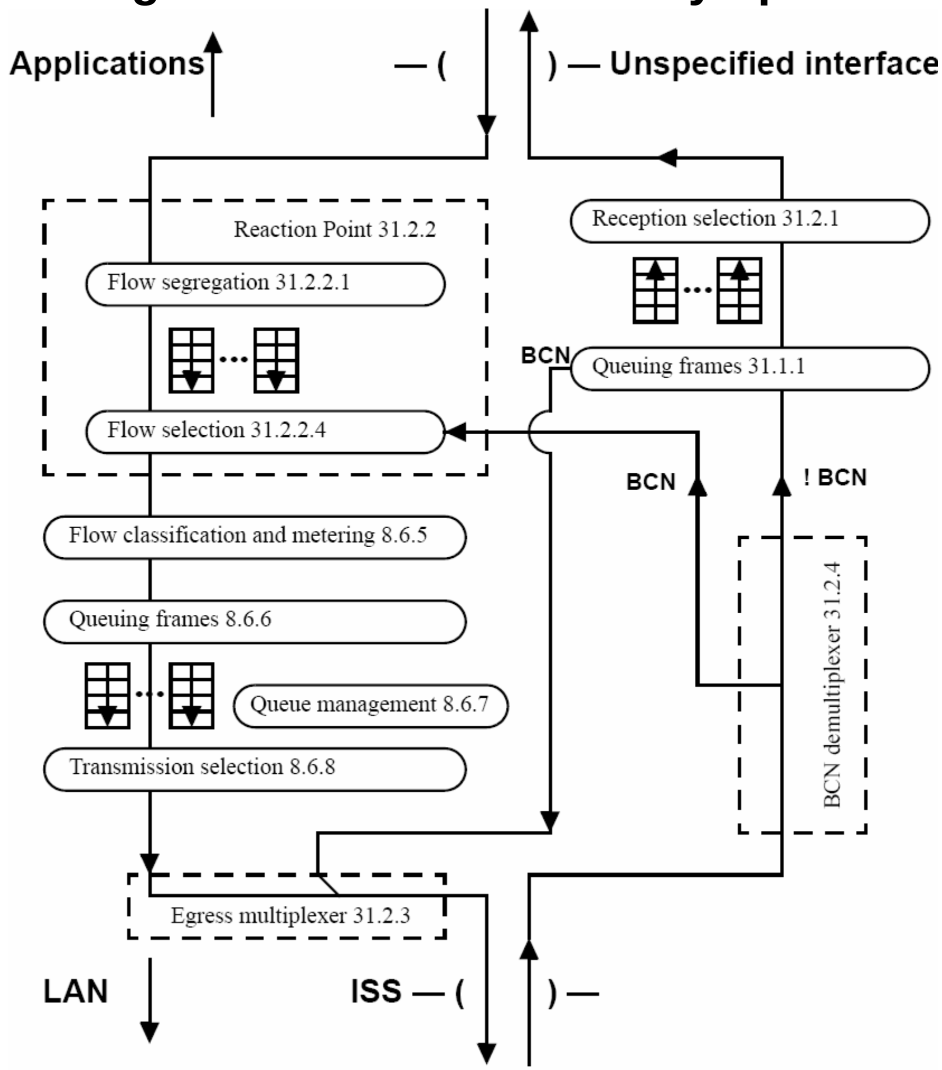


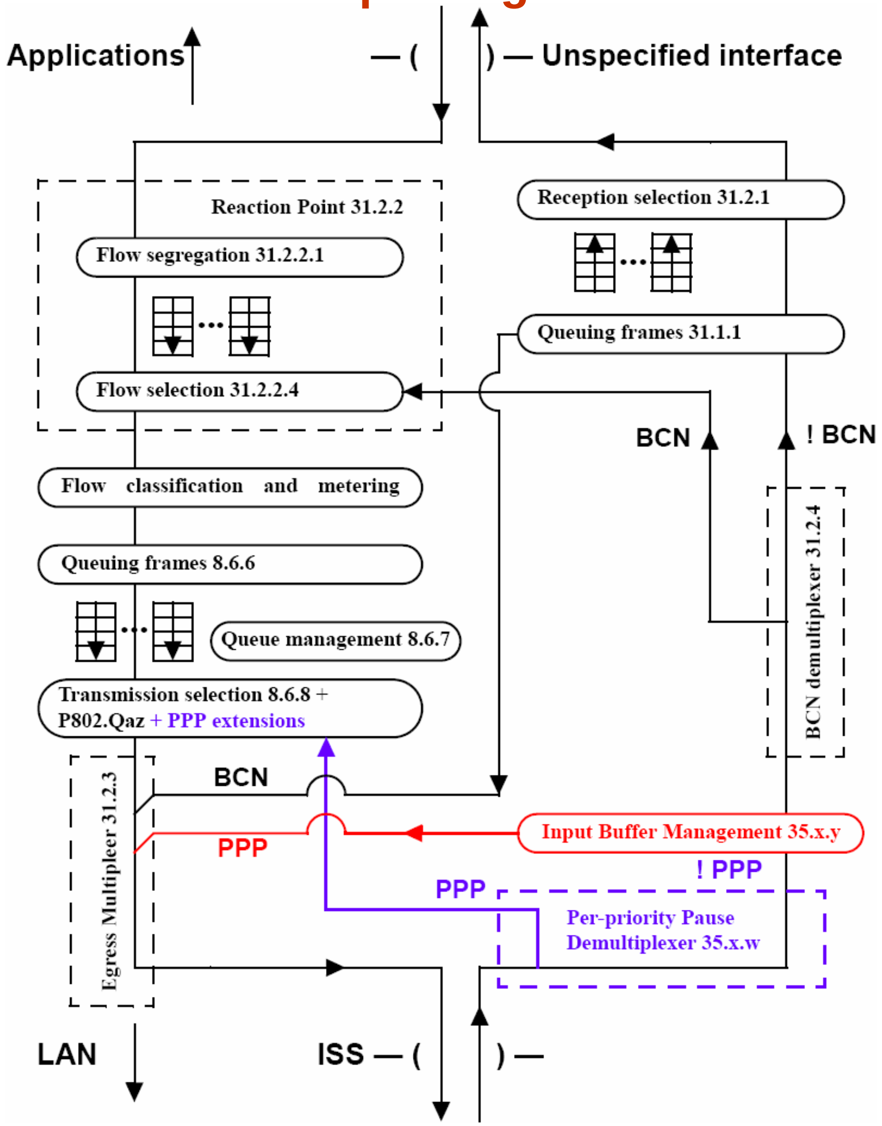
Figure 31-2—Congestion aware station

Proposed extension for PPP in Stations

Subclause TBD – modified queuing entities

RED: PPP generation based on space available in per-port, per-controlled-priority, input buffers.

BLUE: PPP reaction, only for BCN controlled priority queues.



Thank You!

Questions and Answers

- **802.1Qau Congestion Notification**
  - In draft development
- **802.1Qaz Enhanced Transmission Selection**
  - PAR submitted for IEEE 802 approval at this meeting
- **Priority-Based Flow Control**
  - Congestion Management task group is developing a PAR