Technical Overview of Data Center Networks

Joseph L White, Juniper Networks

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- 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.

Abstract

Technical Overview of Data Center Networks

- With the completion of the majority of the various standards used within the Data Center plus the wider deployment of I/O consolidation and converged networks, a solid comprehension of how these networks will behave and perform is essential.
- This tutorial covers technology and protocols used to construct and operate Data Center Networks. Particular emphasis will be placed on clear and concise tutorials of the IEEE Data Center Bridging protocols (PFC, DCBX, ETS, QCN, etc), data center specific IETF protocols (TRILL, etc), fabric based switches, LAG, and QoS. QoS topics will address head of line blocking, incast, microburst, sustained congestion, and traffic engineering.

The Data Center Network is Complex

- L2 Ethernet:
  - VLAN
  - STP
  - DHCP
  - LAG
  - Broadcast/Multicast
  - DCB
    - PFC
    - ETS
    - QCN
    - DCBX
  - TRILL
  - overlay networks
    - VXLAN/NVGRE
- L3/L4:
  - IP
    - TCP
      - NAS
      - iSCSI
    - UDP
    - ICMP
      - ECN
  - transporting Lots of applications!
- FC:
  - SAN Protocol
  - FC as Transport
    - credit flow control
  - FCoE
  - FC Services
- Network Control & Monitoring
  - traditional network management
  - SDN (Software Defined Networks)
- Traffic considerations for the Data Center:
  - queues + buffer
  - head of line blocking
  - incast/microburst
  - sustained congestion
  - latency vs. throughput

Data Center Requirements and Trends

Requirements

High Throughput

High Availability

Wide Scalability

Low Latency

Robustness

Deterministic

Trends

Bandwidth

10G from the server becoming increasingly common
40G, 100G in the works mean network bandwidth available

Traffic Separation

DCB protocols give multiple planes within the same network
Can share the bandwidth without cross traffic interference

Large L2 domains

Enabled by fabric implementations

Latency improvements

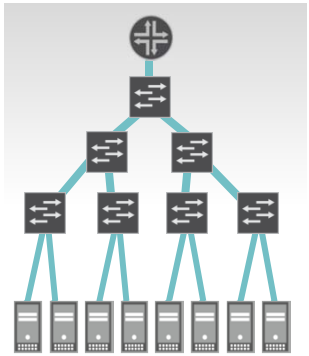
switches and fabrics with optimized forwarding paths

Network Congestion Management

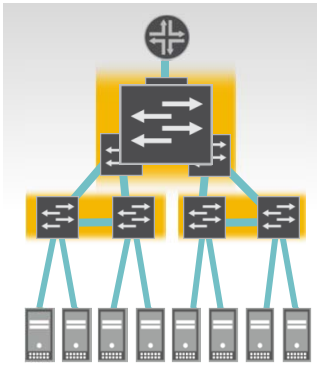
Multiple flow control schemes working at the same time across the physical infrastructure

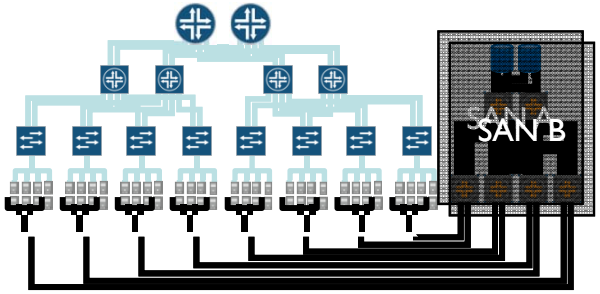
Network Convergence

- Convergence occurring along 2 major themes
  - These are happening at the same time
  - Here we will only focus on the lower case

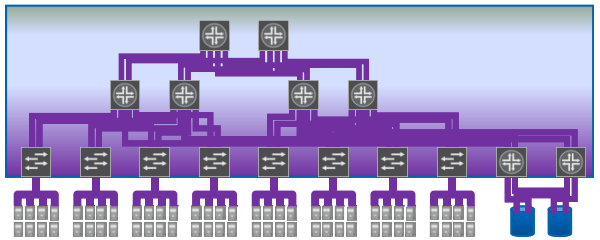


Collapsing Tiers





Converging Infrastructures



Virtualization OF EVERYTHING

- Aggregate up and Virtualize down
  - many examples such as storage arrays, servers, ...
  - avoid Accidental partitioning
  - embrace Deliberate partitioning
- Aggregation
  - Physical and Software
  - Bring together and pool capacity with flexible connectivity
- Virtualization
  - logical partitions of the aggregated systems to match actual need
  - flexibility → fungible resources everywhere
  - Utility Infrastructure with just in time & thin provisioning

THIS IS HAPPENING TO NETWORKS AS WELL

Virtualization Drives Multi-Protocol Connectivity

... because Data Centers are always in flux

Application life cycle

services introduced, updated, retired

Load on servers and networks constantly changing

can be unpredictable

Resource management challenge

- Minimize the need for excess capacity
  - Reconfigure
  - Reclaim/Reuse
- Adding resources is last resort

Dynamic shared resource pools address these issues

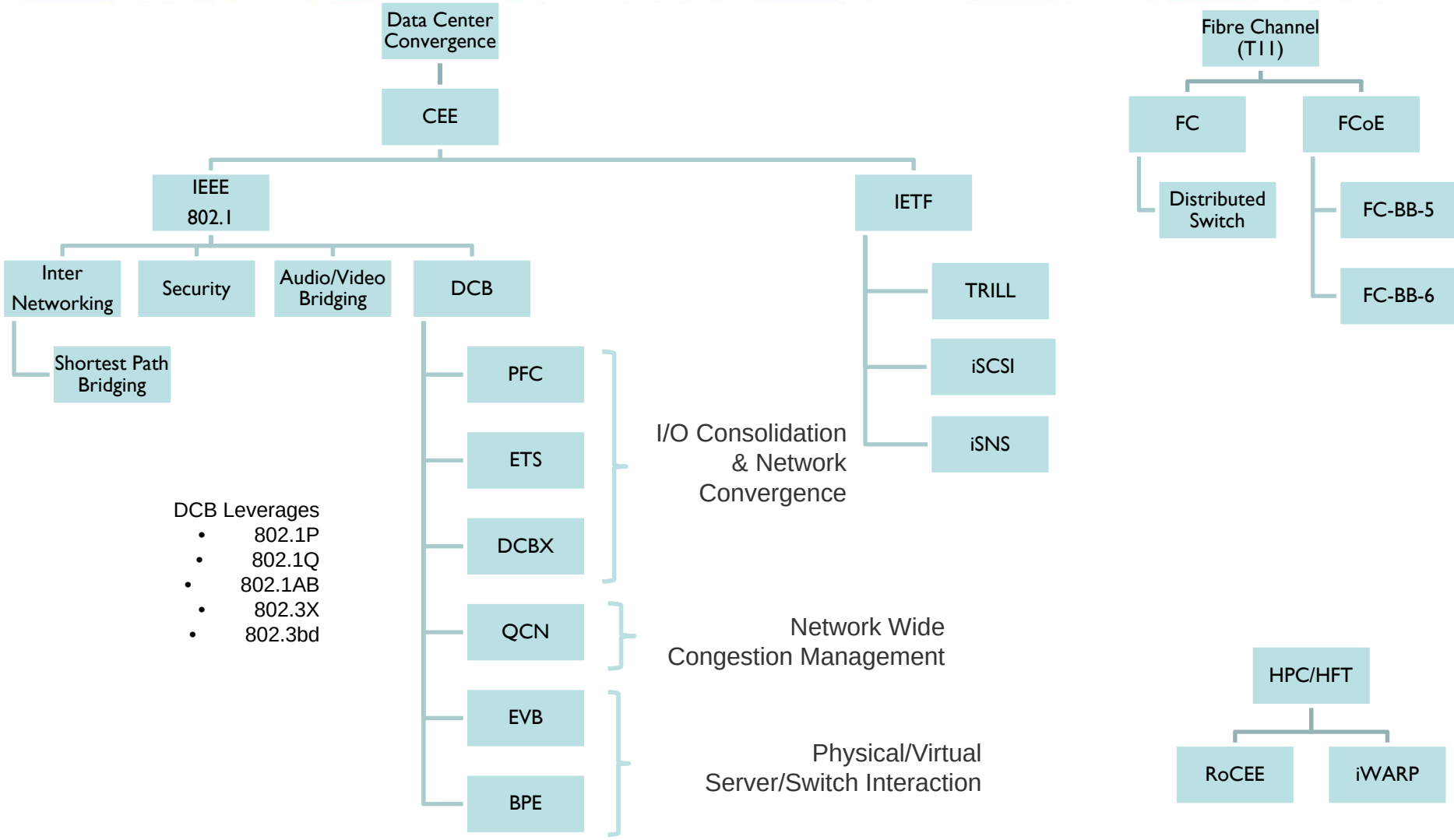
Enabled by Virtualization + Full Connectivity Networks

Any server potentially needs access to any device across a variety of protocols

will drive Drives SAN attach from 20% to near 100%

Do you want a single converged infrastructure or multiple parallel infrastructures?

Protocols Taxonomy



Many Topics

- As you can see this is a huge subject area so for this session we will focus on the following topics:
  - Data Center Bridging Convergence Protocols
    - PFC
    - ETS
    - QCN
    - DCBX
  - Link Level Flow Control
  - Congestion Effects
  - VLANs and Overlay Networks
  - Traffic Considerations

FIBRE CHANNEL OVER ETHERNET (FCoE)

- From a Fibre Channel standpoint
  - FC connectivity over a new type of cable called… an Ethernet cloud
- From an Ethernet standpoint
  - Just another ULP (Upper Layer Protocol) to be transported,
    - but… a challenging one!
  - DCB designed to meet FCoE's requirements

FC-BB-5: VE-VE & VN-VF, FC-BB-6 adds VN2VN, VA2VA

Class 2, 3, and F carried over FCoE

Ethernet Support

Lossless – aka not allowed to discard because of congestion
Transit delay of no more than 500ms per forwarding element
Shall guarantee in order delivery

- Components
  - FCoE/FC Switches (or FCFs)
  - FCoE/FC Gateways (NPIV/NPV/N_Port_Virtualizer based)
  - FCoE Transit Switches (DCB plus FIP-Snooping)

DCB – DATA CENTER BRIDGING

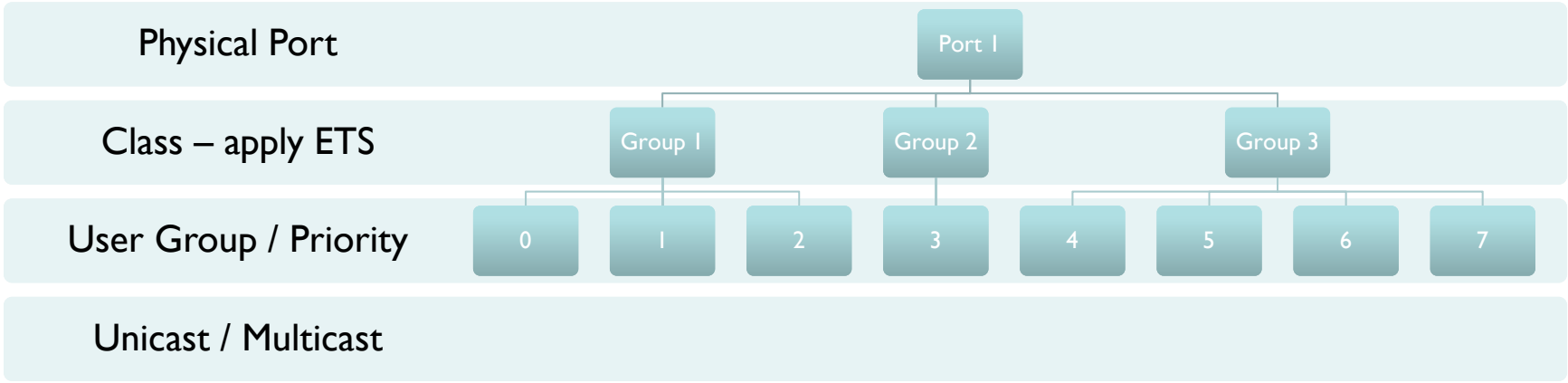
Partitions the network in to parallel planes
- 8 largely independent lanes
- Configurable separately
- Leverages the 3 bit VLAN CoS (aka user priority)

Class groups
- Groups lanes together for ease of configuration

Traffic Classes (Individual Protocols or Protocol Groups)
- Define what can be configured on each lane / class group

- Designed for protocol / service separation
  - NOT VM separation
- Have to bind services to the right 802.1p priority and VLAN
  - aka User Priority or PCP Priority Code point
  - Operating and Hypervisor enablement
  - CNA Mapping

DCB HIERARCHY – CONCEPTUAL VIEW



Implementations may have separation of multicast & unicast

Builds on existing standards – really just small amendments:

- 802.1P User Priorities & VLAN structures leveraged
- 802.1Q Forwarding & Queuing Functions as a basis
- 802.3X Flow Control as a basis
- 802.3bd MAC Control Frame for PFC amended
- 802.1AB LLDP leveraged

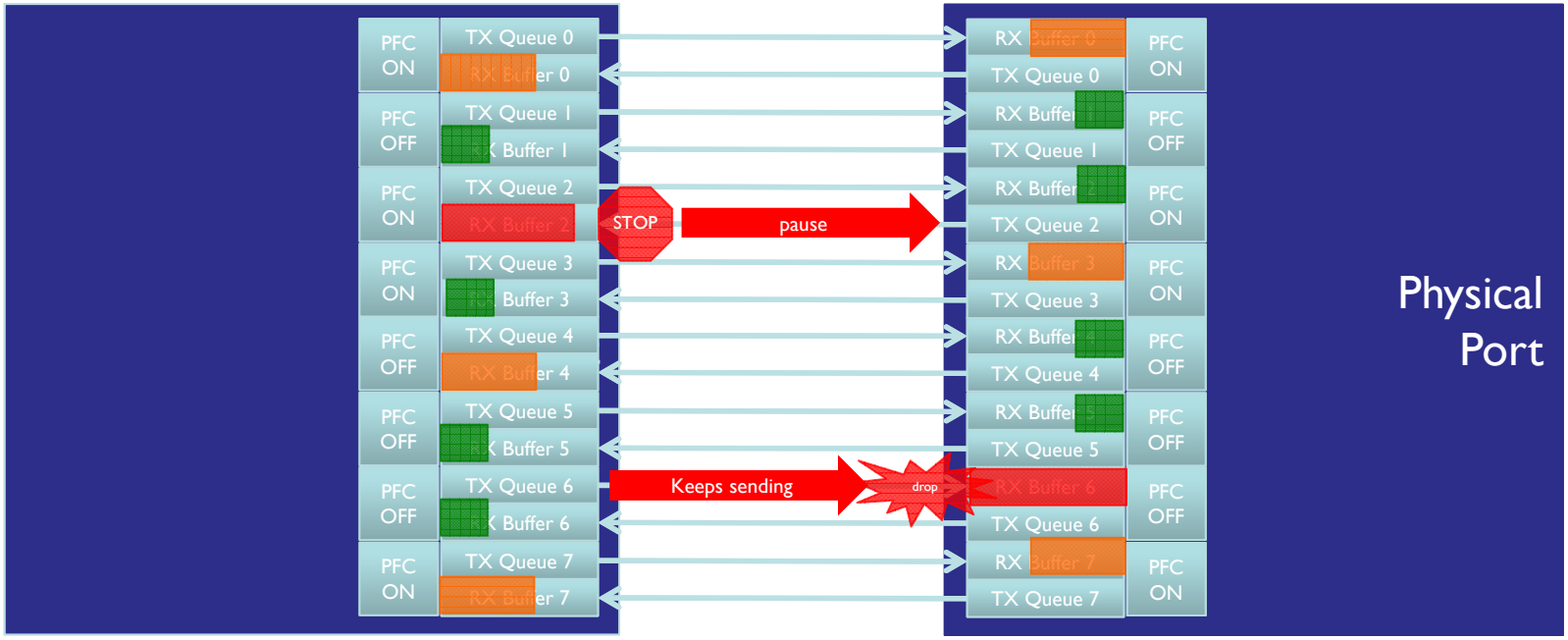
Link Level Flow Control

- What Traffic Requires or Benefits from Lossless Networks?
- Techniques
  - Pause
  - Priority Flow Control
  - Pause vs Credits
- Complications around Link Level Flow Control
  - Head of Line Blocking
  - Congestion Spreading
  - link level configuration vs end to end loss behavior
    - internal forwarding paths vs external links
    - for end to end need vswitch,

Who benefits from lossless networks?

- Lossless Required
  - FCoE
- Other Lossless Candidates
  - iSCSI local traffic
  - LAN Backup
  - Vitrual Machine Mobility
  - Cluster
  - HPC
- Lossy Candidates
  - Management
  - Campus
- *Notes*
  - *may want or need more than one priority for some protocols*
  - *TCP loss recovery mechanisms inject latency and deliberately reduce throughput. For some applications this is undesirable.*

Priority Flow Control (PFC)



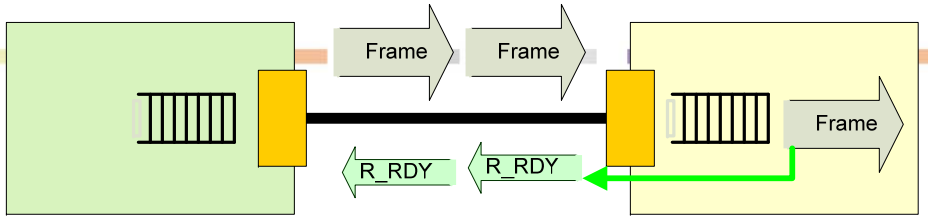
Requirements

- **Enabling PFC on at least one priority**
- On PFC Priorities, PFC M_CONTROL.requests
- On PFC Priorities, PFC M_CONTROL.indications
- Abide by the PFC delay constraints
- **Provide PFC aware system queue functions**
- Enable use of PFC only in a DCBX controlled domain

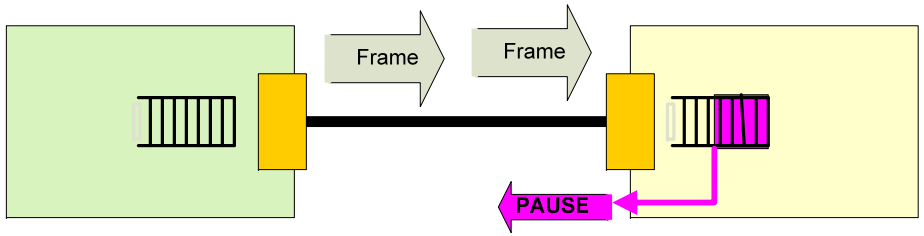
Optional

- **Enabling PFC up to eight priorities**
- **Completely independent queuing**
- Supporting the PFC MIB
- **Buffer optimized by MTU per priority**
- **Configurable by RTT**
- **Buffer pooling across switch ports**

Credit vs Pause Based Flow Control



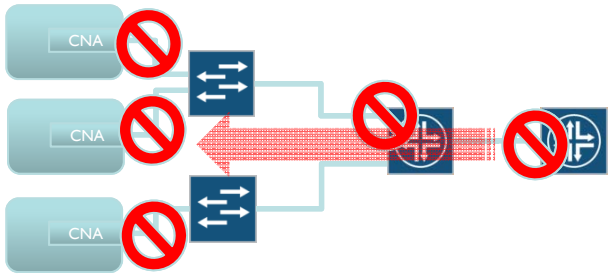
- **FC Credit based link level flow control**
  - A credit corresponds to 1 frame independent of size
    - (1 credit needed per 1 km separation for 2G FC)
  - Sender can only xmit frames if he has credit sent from the receiver (R_RDYs)
  - Amount of credit supported by a port with average frame size taken into account determines maximum distance that can be traversed
    - ***If the credit bandwidth delay is exceeded the maximum possible sustained throughput will be reduced***



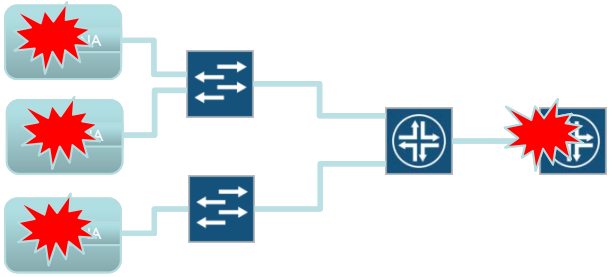
- **Pause Frame Based Link Level Flow control**
  - When the sender needs to be stopped the receiver sends a frame to notify the sender
  - For lossless behavior the receiver must absorb all the data in flight
  - This puts a hard limit based upon the receiver buffer on the distance for storage traffic across a direct connect Ethernet
    - ***If the buffer is overrun then frames can be dropped***

Link Level Pause Complications

- Head of Line Blocking and Congestion Spreading
  - For lossless networks FC based SANs have always worked like this



- Lossless without QCN
  - Congestion spreading



- Lossy without QCN
  - Congestion causes loss
  - TCP Congestion controls are dominant behavior

Traffic Management

- Traffic management is a blanket term for protocols and algorithms that attempt to control bandwidth and congestion throughout the network.
  - Note that these is considerable interaction with link level flow control
  - Applications can also directly manage their characteristics but we are not considering that here
- Output Restrictions
  - Rate Limiting
  - ETS (Enhanced Transmission Selection)
    - settings within DCBX which express configuration for output rate limiters
- End to End Congestion controls
  - TCP/IP inherent
    - driven deliberately by WRED (weighted random early detection/discard) and related algorithms (policers)
  - QCN
  - ECN
  - ICMP source quench

Output Rate Limits and ETS

- Output Rate Limiting
  - Sets bandwidth limits on a per class/group/queue basis
  - Allows control of the amount of traffic on a given link
- ETS (Enhanced Transmission Selection)
  - settings within DCBX which express configuration for output rate limiters
  - way to define groups and assign bandwidth to each group
  - assign priorities to groups
- Most implementations also have multi-layered/tiered queuing and scheduling

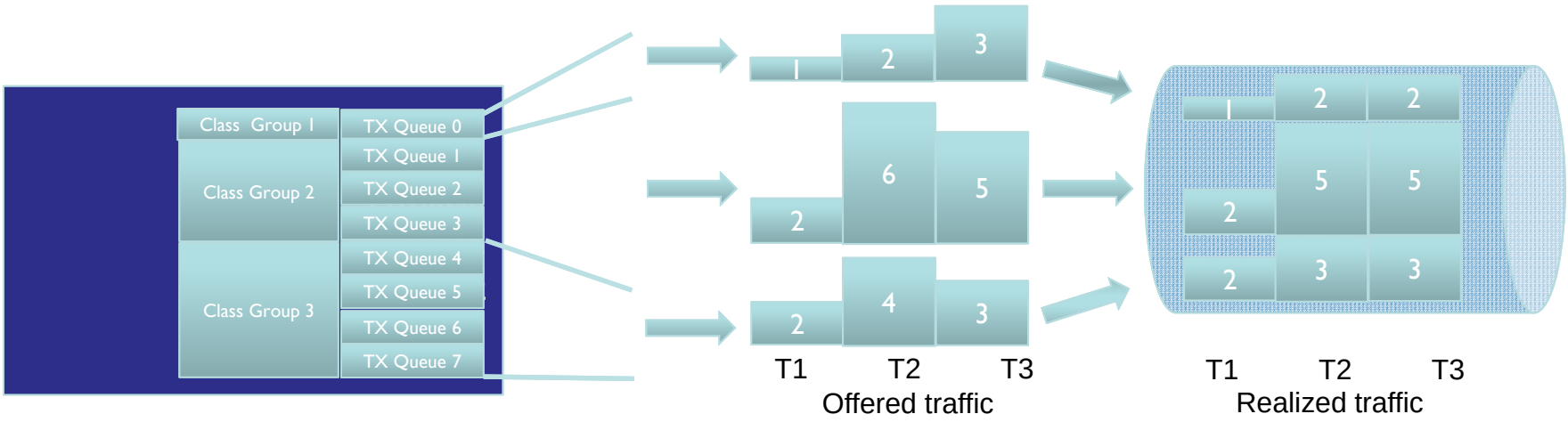
ETS Details

From a link perspective

- ETS controls bandwidth
  - **Applies per class group**
    - **A class group is one or more classes**
    - **Normal fairness within a given group**
- Applies to both lossy and lossless
  - For lossless results impacts timing of pause
  - For lossy impacts when frames dropped
  - Applies after strict priority traffic accounted for

Requirements

- **Support at least 3 Traffic Classes *and up to 8***
  - 1 for priorities with PFC Enabled
  - 1 for priorities with PFC Disabled
  - 1 for Strict Priority
- Support bandwidth allocation at a granularity of 1% or finer allocating from the bandwidth remaining after that used by other algorithms
- Support a transmission selection policy such that if one of the traffic classes does not consume its allocated bandwidth, then any unused bandwidth is available to other traffic classes (optional enable / disable)



Characteristics of TCP

- For Storage Networking TCP is Critical (FCIP, iSCSI, NFS, pNFS, SMB)
- Connection Oriented
  - Full Duplex Byte Stream (to the application)
  - Port Numbers identify application/service endpoints within an IP address
  - Connection Identification: IP Address pair + Port Number pair ('4-tuple')
  - Well known port numbers for some services
  - Reliable connection open and close
  - Capabilities negotiated at connection initialization (TCP Options)
- Reliable
  - Guaranteed In-Order Delivery
  - Segments carry sequence and acknowledgement information
  - Sender keeps data until received
  - Sender times out and retransmits when needed
  - Segments protected by checksum
- Flow Control and Congestion Avoidance
  - Flow control is end to end (NOT port to port over a single link)
  - Sender Congestion Window
    - responds to packet drops and reordering
  - Receiver Sliding Window

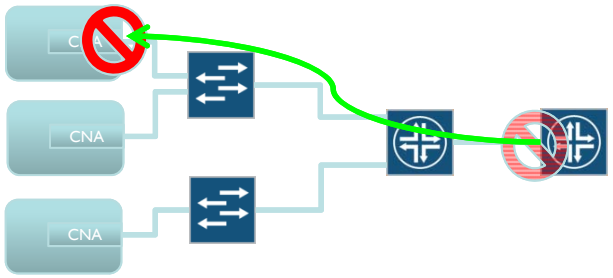
Congestion Controls: QCN

QCN operates on a per priority basis… Reaction points configured per port…
'somewhat like TCP congestion management but L2 and for DC distances'

- Proactive
  - Devices don't have to drop packets to trigger host slow down
  - Instead can back pressure based on congestion
  - Transmitting device knows its congestion and backs off
- How well does QCN Actually work ?
  - How fast is traffic flow changing ?
  - What is the reaction time from actual event to adjusting flow rate ?
  - What is the granularity ?

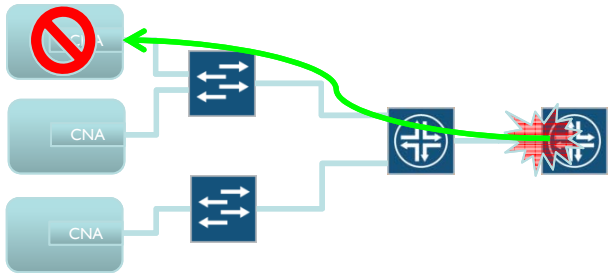
Using QCN with Lossless

- Reduces congestion spread



Using QCN with Lossy

- Reduces packet loss



Congestion Controls: ECN

- ECN (Explicit Congestion Notification)
  - This is an extension to IP or TCP
  - uses bits in the headers to signal capability & indicate detected congestion
    - for IPv4 header ECN uses bits in the DiffServ Field
      - 00: ECN not supported
      - 10: ECN Capable Transport
      - 01: ECN Capable Transport
      - 11: Congestion Encountered
    - along with the TCP header bits
      - NS (Nonce Sum)
      - ECE (ECN-Echo)
      - CWR (Congestion Window Reduced)
  - the operation is as follows
    - The endpoints indicate that they are capable
    - an intermediate hop along the path to marks capable packets that would have been dropped with Congestion Encountered
    - The TCP/IP receiver echoes the fact that the congestion was encountered by setting the ECE flag for packets sent on the connection
    - The TCP/IP sender reacts to getting an ECE by reducing his congestion window and setting CWR to let the receiver know his request was honored and can stop setting the ECE flag
  - Note this scheme requires ECN capability in the endpoints as well as intermediate network

Network Virtualization

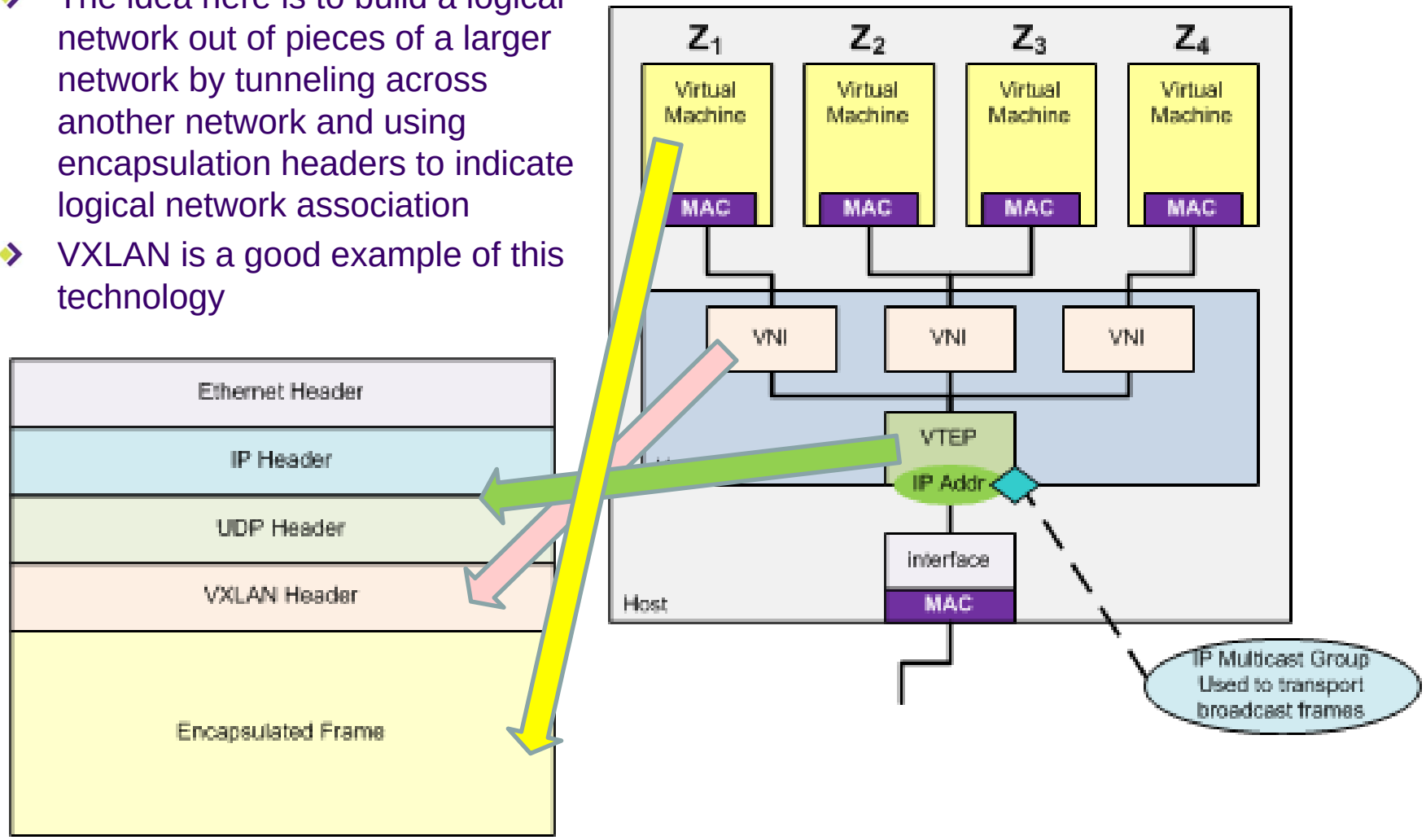
- Network Virtualization allows multiple simultaneous logical connectivity domains across a single physical infrastructure
- VLANs
  - these certainly virtualize the network
- Overlay Networks and Tunnels
  - VXLAN, NVGRE
  - EVPN
  - L2 Multipath
- Link Aggregation

VLANs

- A VLAN sets an Ethernet L2 visibility and broadcast domain
  - indicated in frames by addition of 12 bit field to the Ethernet header
  - most switches allow for a port level default VLAN
  - endpoints are allowed to put the VLAN tag on themselves
  - note that flow control is on a priority not VLAN
- VLANs can help in a multi-protocol environment
- Allows a physical Ethernet infrastructure to logically separate traffic
  - for example FCoE can be on its own VLAN(s) separate from traditional traffic

Overlay Networks

- The idea here is to build a logical network out of pieces of a larger network by tunneling across another network and using encapsulation headers to indicate logical network association
- VXLAN is a good example of this technology



Configuration and setup

- Many protocols at link and local network level
  - STP (many flavors)
  - LACP
  - **DCBX –** *we'll talk about this one in a little detail*
  - others as well...
- Other Protocols for Network Management and Monitoring
  - you might even want to add SDN into this area as it is a super form of configuration
  - These are huge technology areas all on their own

DCBX

Point to Point Link Auto negotiation of capability

- Configures the characteristics of each link
- Mandated by PFC ETS and QCN
- DCBX MIB is mandatory
- based on LLDP

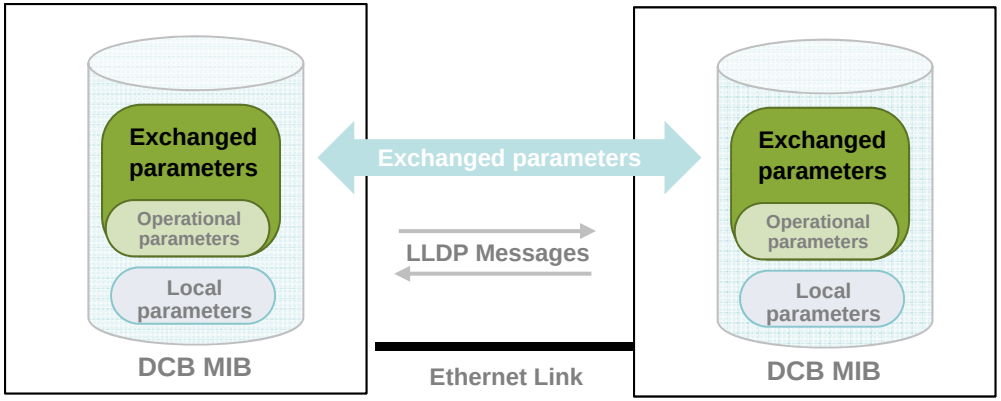
Also protocol & application TLV to define configuration

- DCB Protocol profiles
- FCoE & iSCSI profiles
- Other application profiles

Priority Group Definition
Group bandwidth allocation
PFC enablement per priority
QCN enablement
application assignment to a priority
logical down

Multiple Versions

1.00 (pre-standard)
1.01 (commonly deployed)
IEEE (official DCB standard)



DCBX Applications Examples

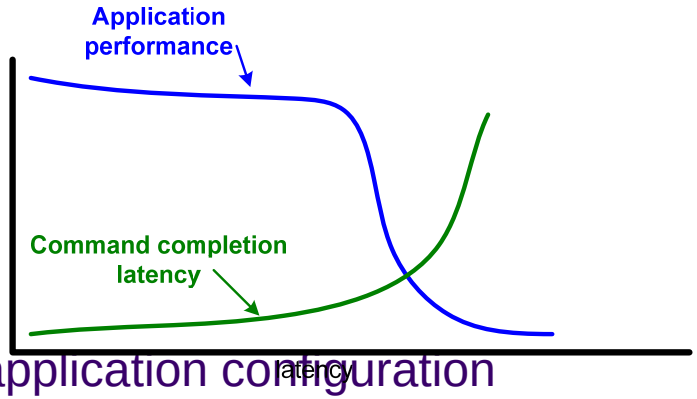
Application	TCP / UDP / Ethernet	Port / Ethertype
HPC - RoCE	Ethernet	8915
HPC - iWARP		
NAS – NFS (old)	UDP	2049
NAS – NFS (new)	TCP	2049
NAS – CIFS/SMB	TCP	445
FCoE		8906
FIP		8914
iSCSI	TCP	3260

Traffic Effects

- Head of Line Blocking and congestion spreading
  - discussed previously
- Latency
  - latency vs throughput
- Incast
- micro-burst, milli-burst, latency bubbles
- speed mismatch effects
- different flow control at different protocol levels
  - TCP/IP over Lossless Ethernet for example
- mixed flow control
  - some part of the network has link level flow control and part has no flow control for example

Latency

- Command/Transaction Completion Time is important
- Contributing Factors: (sum 'em all up!)
  - Distance (due to 'speed of light') - latency of the cables
    - ($2x10^8$ m/s gives 1 ms RTT per 100Km separation)
  - 'Hops' – latency through the intermediate devices
  - Queuing delays due to congestion
  - Protocol handshake overheads
  - Target response time
  - Initiator response time

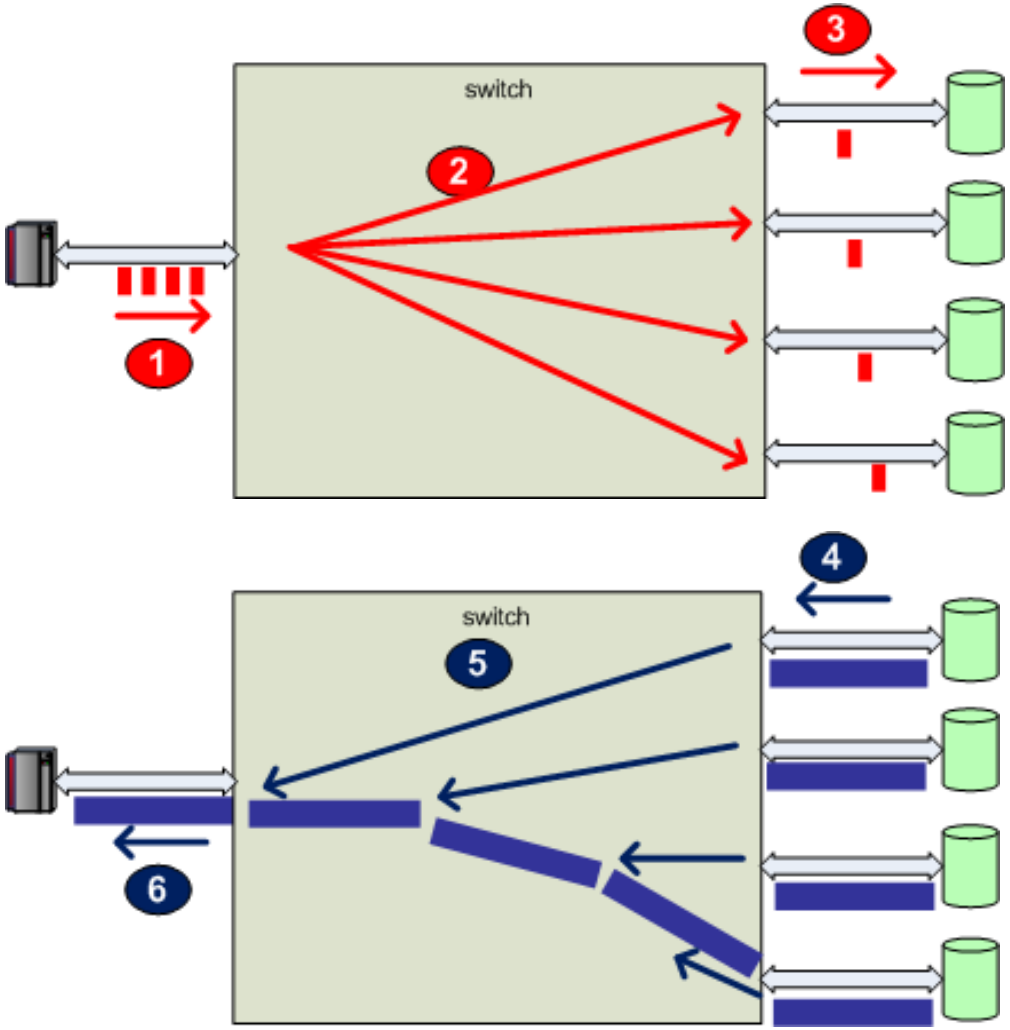


- A complicating factor is the I/O pattern and application configuration
  - Some patterns and applications hide latency much better than others
    - Good: File Servers and NAS
    - Bad: transactional database with heavy update loads

Latency vs Throughput

- Latency and throughput have a complex relationship
  - the statements are broad generalizations to illustrate the concepts
- Throughput generally has two flavors
  - IOPS (Input Output Per Second) – command or request completion rate
  - MB/s (Megabytes per Second) – Data traffic throughput
- High Throughput of both types can still be achieved under high latency
  - multiple outstanding requests/commands hide latency as long as the requirement is not on any specific command
  - streaming data sources
- Latency hurts throughput for single request at a time transactional systems
  - if each request has to wait for the previous one to finish then the more latency can kill performance (see articles on 'buffer bloat in the internet')
- SOME systems need low latency even when they might also have multiple outstanding streaming data sources
  - High rate financial data distribution

Incast



- Incast occurs when a request for data or set of commands to multiple destinations results in a large burst of data/traffic back from each of those destinations to the requestor closely correlated in time
- This time correlated burst of data can cause congestion related backup or tail drops

Microburst

- This is transient congestion across a link or at an output port due to more data arriving for a given output from multiple sources than can be buffered.
  - drops or congestion spreading can occur in this case even if the average traffic rate is such that no link in the system is oversubscribed
  - Can be random
    - typically more frequent if there are a large number of flows
  - can be due to incast and incast like correlation

Final Thoughs

- We covered many topics
  - and many more were not covered!
- Education and Awareness are Key
  - The large number of protocols in operation leads to many complex interactions
  - However the scale and flexibility gains are large so we have to deal with these interaction
  - The type of complexity is outlined here is manageable with proper designs
  - Several of the protocols are specifically targeted at mitigating negative effects

Attribution & Feedback

The SNIA Education Committee thanks the following individuals for their contributions to this Tutorial.

Authorship History

Original Author:
Joseph L White, August 2012

Updates:
March 2012

Additional Contributors

Simon Gordon

Please send any questions or comments regarding this SNIA Tutorial to ***tracktutorials@snia.org***