Data Center Fabric Trends

Patricia Thaler

Broadcom

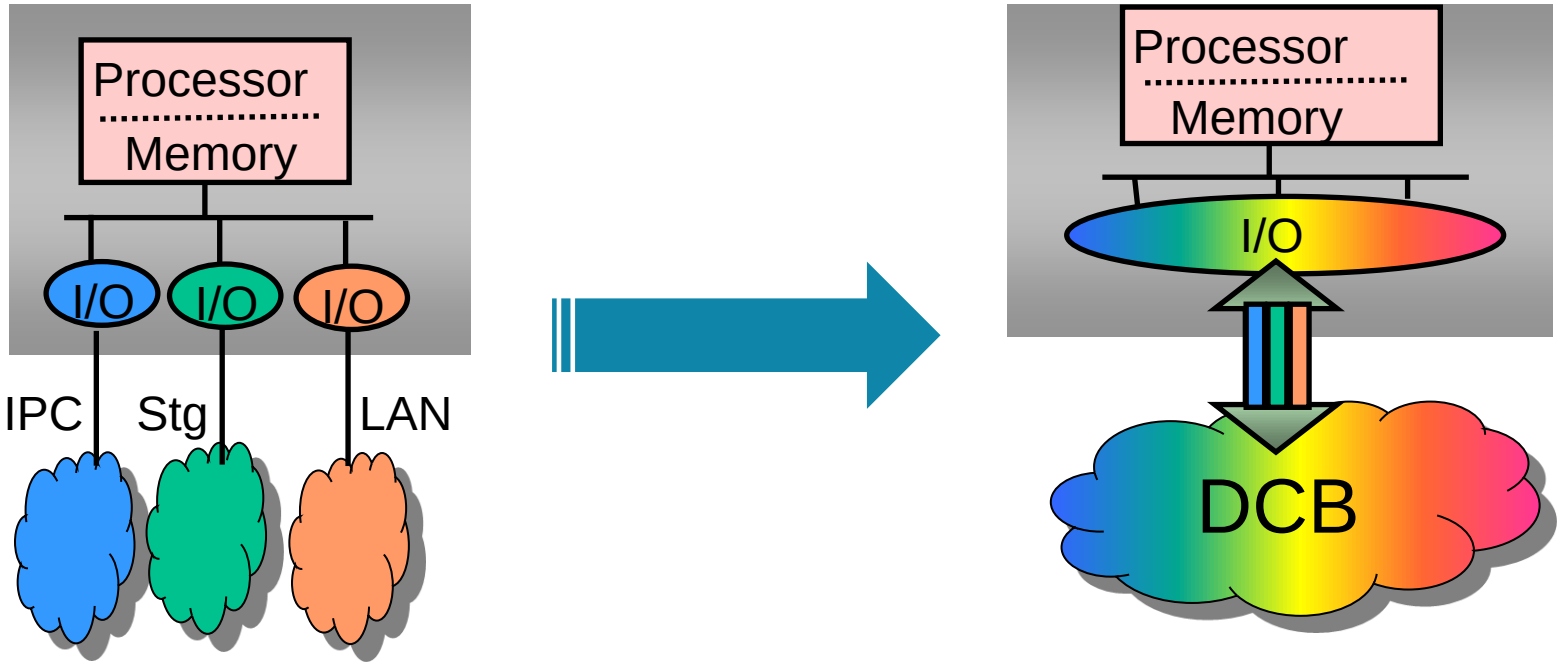
Acknowledgements

- For content and graphics from the Data Center Bridging Tutorial

  http://ieee802.org/802_tutorials/07-November/Data-Center-Bridging-Tutorial-Nov-2007-v2.pdf

  – Pat Thaler, Manoj Wedakar, Anoop Ghanwani, Joe Pelissier, Anoop Ghanwani

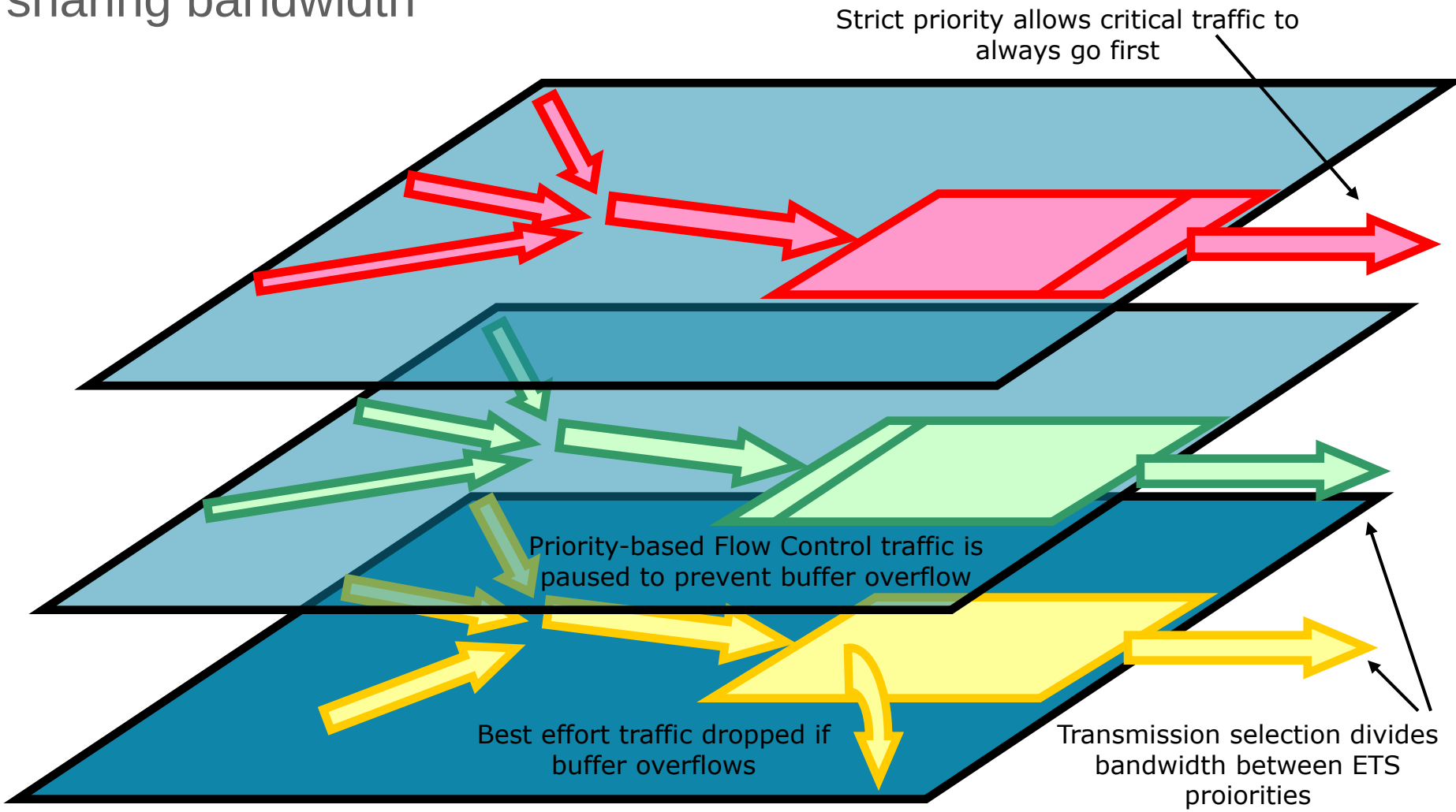
Fabric Convergence



- Convergence:
  - Allowing traffic to migrate from specialized networks, e.g. Fibre Channel and Infiniband, to Ethernet
- Lossless Fabric:
  - Fibre Channel and Infiniband fabrics provide link flow control for effectively no congestion loss
  - DCB provides this over Ethernet while maintaining traditional congestion behavior for other traffic.

Data Center Bridging

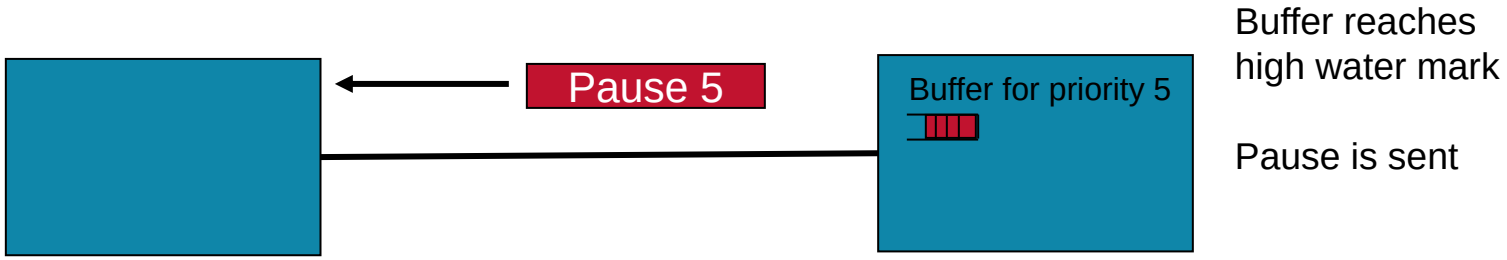
- Differentiated flow control behavior based on priority while sharing bandwidth



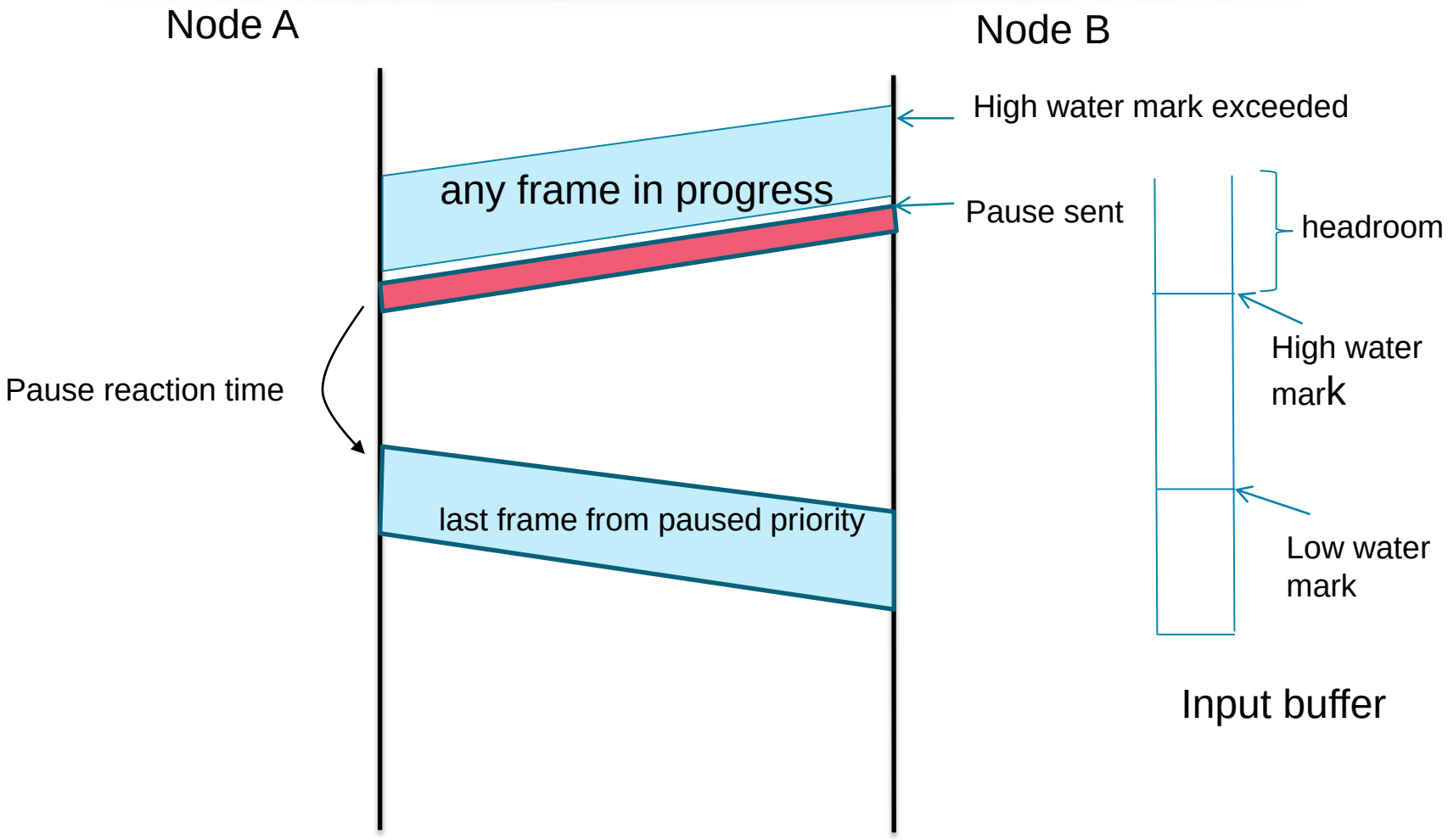
Priority-based Flow Control (PFC)

Priorities can be individually configured with PFC to provide the no-drop behavior desired for FCoE and RoCE

- Uses a MAC Control message similar to 802.3 PAUSE
- PFC PAUSE frame carries
  - Vector with a bit for each priority to say whether the message applies to it
  - A time value for each priority
    - A zero means that the priority is unpaused
    - A non-zero value indicates the time to pause
    - Usually used as X-ON, X-OFF by sending the maximum time value when high water mark is exceeded and released by sending zero value



PFC



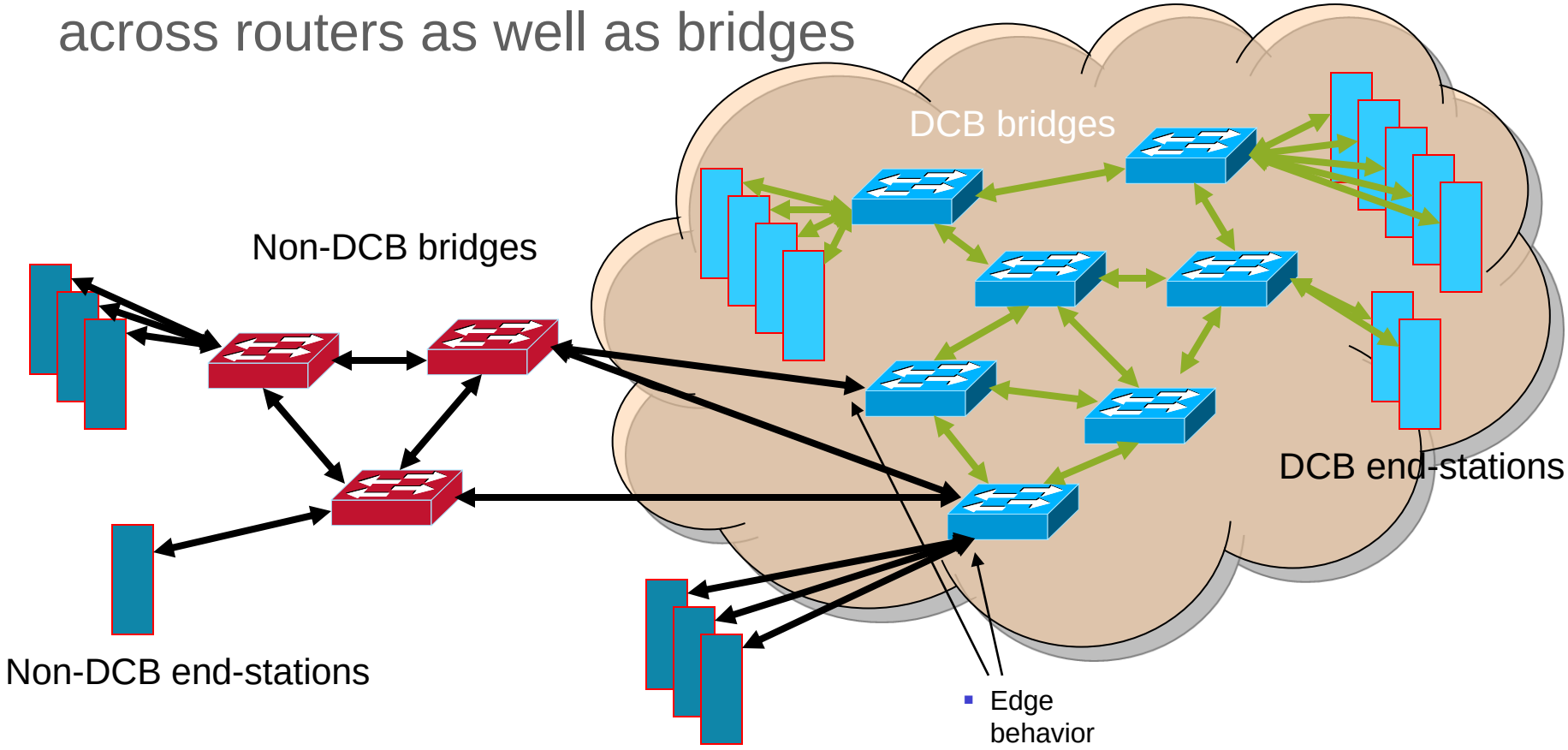
For no drop behavior, must have enough headroom to absorb packets before Pause takes effect.

Enhanced Transmission Selection (ETS)

- IEEE 802.1Q initially only specified strict priority
  - Lower priority traffic is only sent when the higher priorities have no traffic
  - A credit shaper algorithm was added for audio and video traffic
  - IEEE 802.1Q also allowed proprietary transmission selection algorithms
- Network consolidation required bandwidth allocation to traffic classes
  - If traffic in a PFC priority was higher priority than non-PFC, it could lock the non-PFC traffic out of the link and if the non-PFC traffic was higher priority, the PFC traffic could be locked out
  - Most data center hardware already supported some form of weighted round-robin transmission selection
- ETS was added to provide a uniform way to manage bandwidth sharing for data center network consolidation

DCB cloud

- PFC and ETS are provided over a cloud of bridges that support them
- For routable RoCE traffic, this behavior should be maintained across routers as well as bridges

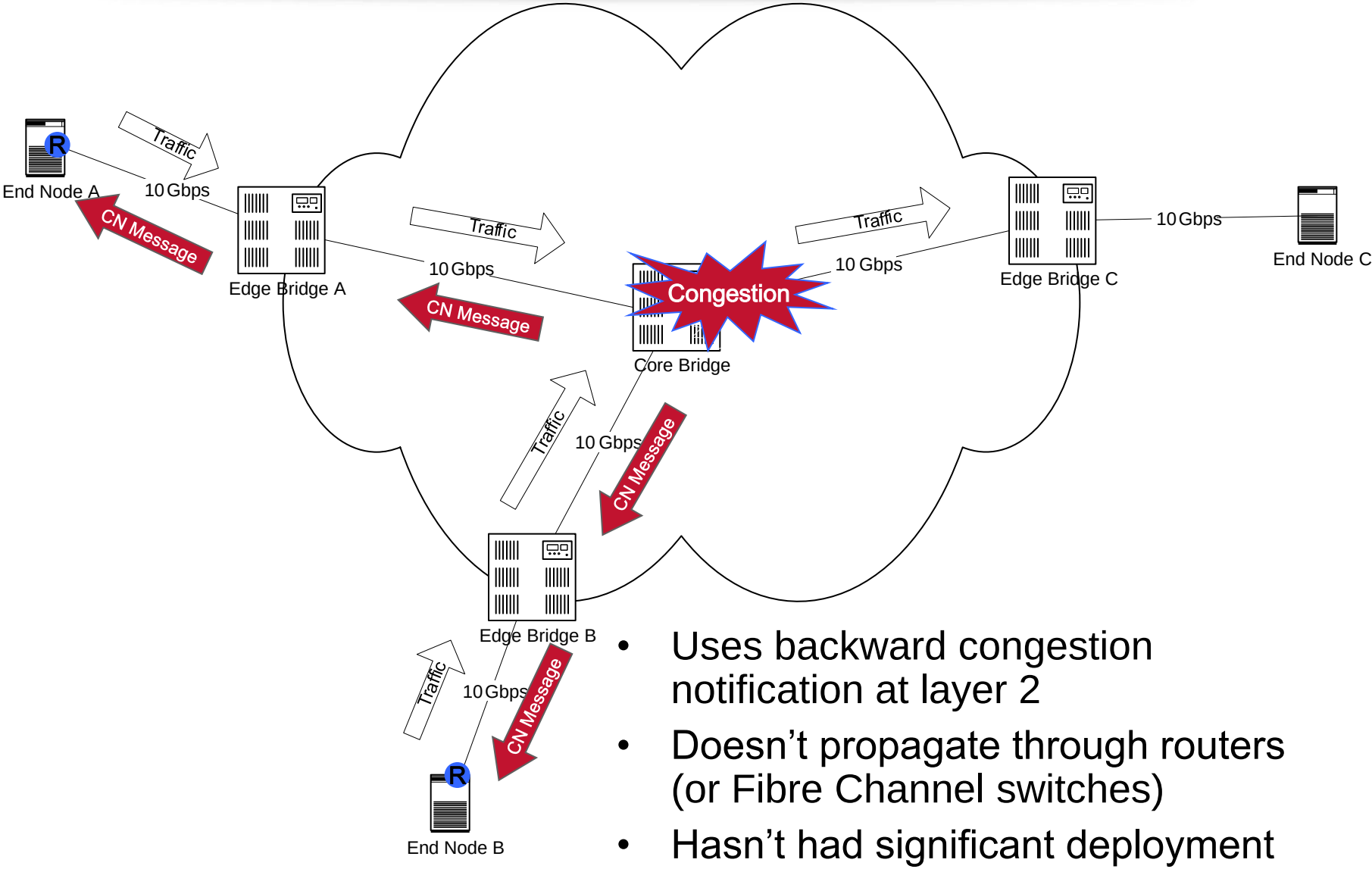


Data Center Bridging Exchange Protocol (DCBX)

- Runs over IEEE 802.1AB LLDP
- Allows a “willing” device to learn the DCB configuration from its link partner
  - Typically used to allow end nodes to learn the network configuration
- Allows bridges to learn whether their link partner has a compatible DCB configuration

- Includes
  - PFC and ETS configuration
  - Application to priority mapping
  - Work is underway to add application to VLAN ID mapping

Congestion Notification (CN) (aka Quantized Congestion Notificaiton: QCN)



- Uses backward congestion notification at layer 2
- Doesn't propagate through routers (or Fibre Channel switches)
- Hasn't had significant deployment

Link flow control potential future work

- Concerns that PFC will require too much buffer headroom on links operating faster than 100Gb/s
- Consideration of a credit-based mechanism has been suggested
- Pre-PAR* discussion is underway in IEEE 802.1 DCB Task Group

* IEEE 802 to IETF translation: pre-PAR - pre-charter

References

- IEEE Std 802.1Qaz Enhanced Transmission Selection
  - Includes DCBX and ETS
- IEEE Std 802.1Qbb Priority-based Flow Control
- Note – a revision is underway to incorporate these amendments into IEEE 802.1Q – the main content will be in 8.6.8, Clauses 36 to 38
- IEEE 802.1Q-2011 Clauses 30-33 (origionally IEEE 802.1Qau) Congestion Notification
- IEEE P802.1Qcd Application VLAN TLV

Thank you