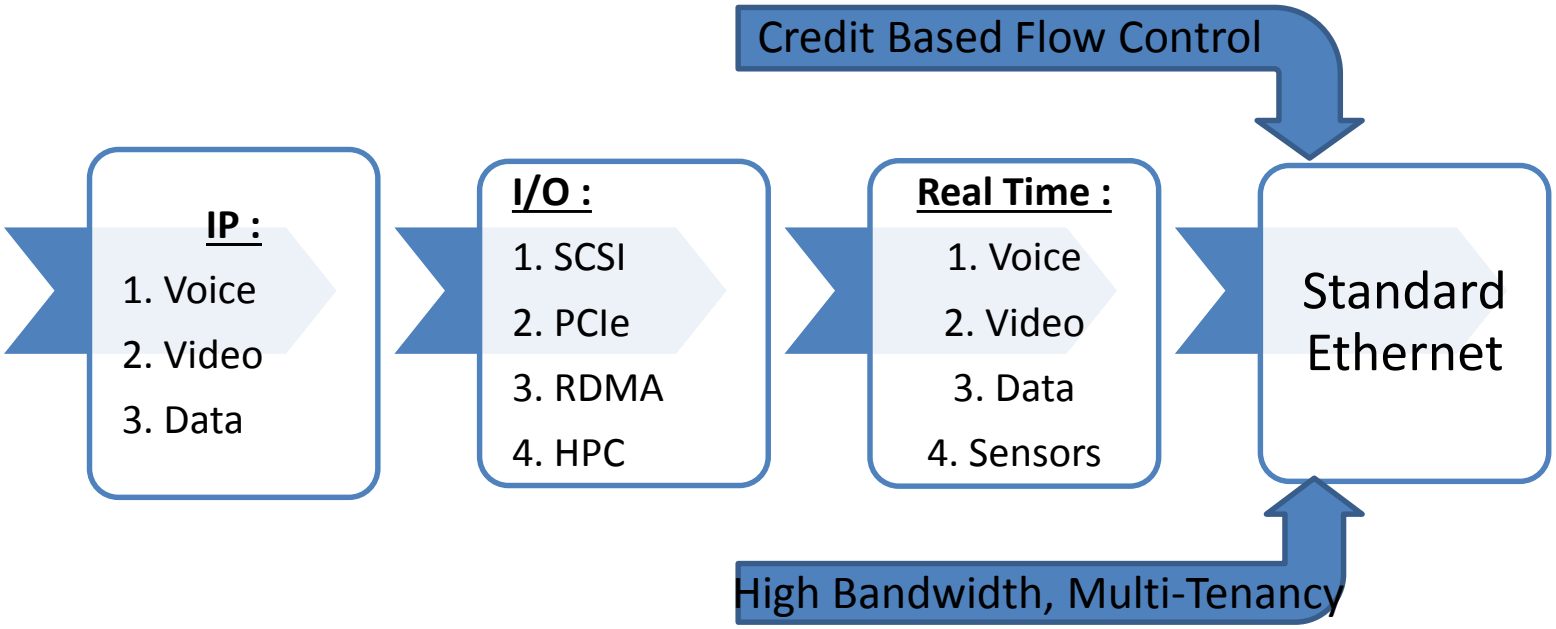
Credit based Link Level Flow Control and Capability Exchange Using DCBX for CEE ports.

Keshav Kamble (kkamble@us.ibm.com)
Jeffery Lynch
Renato Recio
Casimer DeCusatis
Mitch Gusat
Cyriel Minkenberg
Vijoy Pandey
IBM Corporation

Convergence



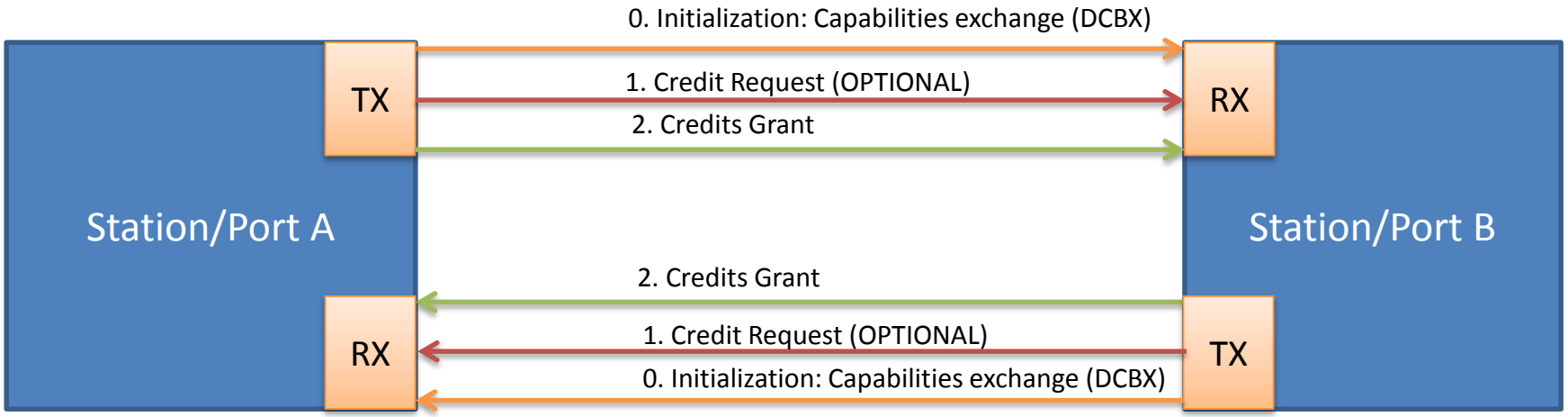
CEE Facts

- Cable Length supported is directly proportional to amount of per port buffering for PFC RTT delay.
- Limits the distance between compute and storage. Limits metro area network connectivity between data centers.
- Flash storage requires low latency and high bandwidth to see the application experience.
- Cost of CEE enabled switches : competition against Infiniband.

Proposed Credit Exchange Logic

- A new frame i.e. Credit Exchange Frame (CE Frame) with a MAC Control Ethernet Type 0x8808. OPCODE 0x0110/0x0111. (Slide 4)
- To ensure loss-less delivery of frames where the frames are of different priorities ranging from 0 to 7, sender (MAC TX) optionally requests credits by sending a new CE frame to the peer port. This frame contains credit requests for multiple priorities per tenant slice (channel Id).
- The peer ports (receivers or MAC RX) receive the CE Frame or CE-Tagged frame and interprets the requests for corresponding priorities / priority.
- Calculated amount of credits are issued to the peer ports by sending a CE frame.
- Upon receipt of credits, sender sends frames for the appropriate / allowed priorities per cannel.
- Unit of credit exchange is 512 bits or each peer port can decide its own unit of credit as per it local TM / Buffer Mgr implementation.
- Suggest extension to the DCBX TLV to exchange capabilities of credit exchange and credit unit selection.
- Credit aging and timers. Consideration to avoid loss of credits.

Capabilities and Credit Exchange

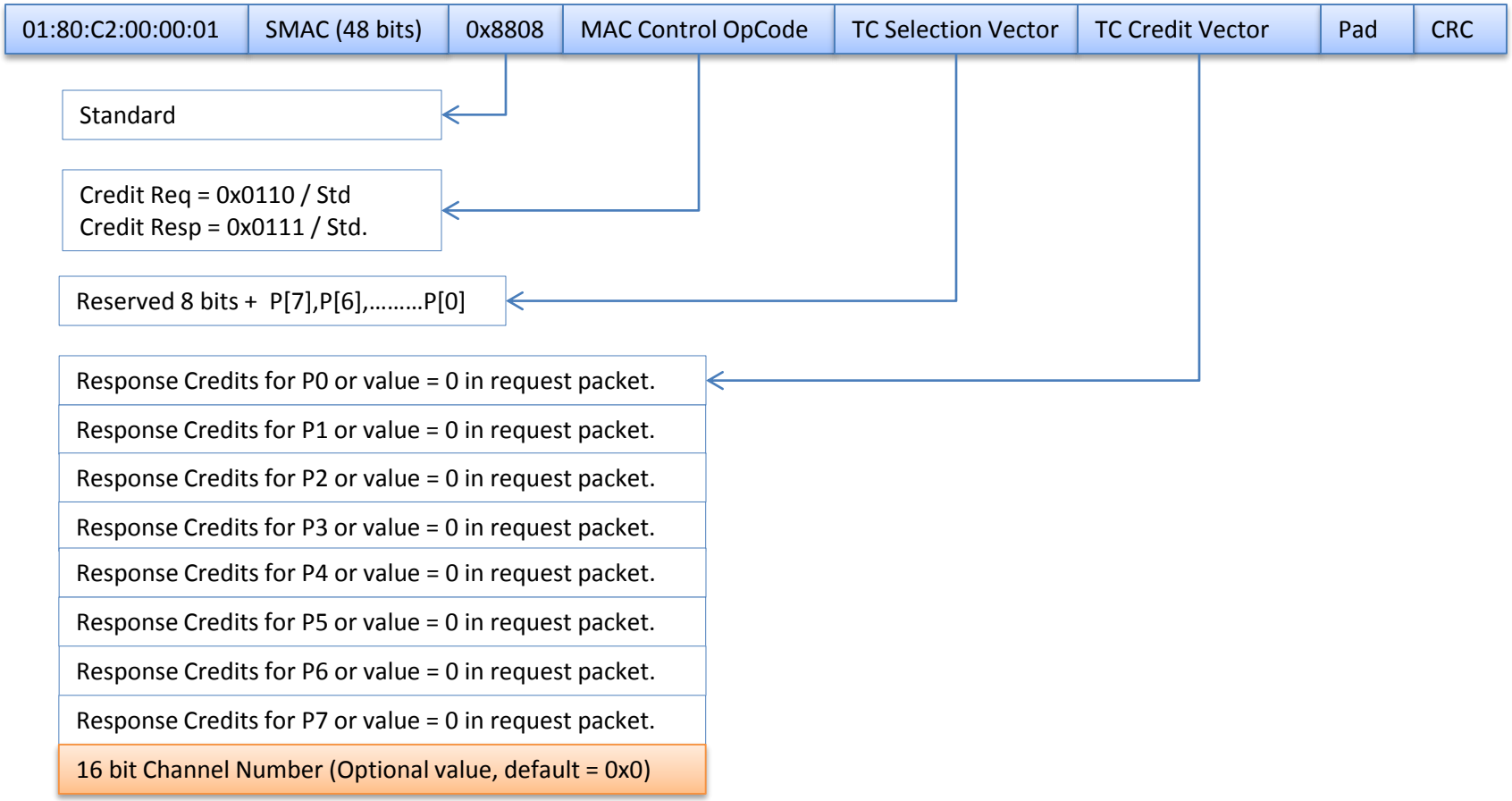


Credit Exchange

- Credit Requested = F(egress queue size or number of frame/s to be transmitted, port speed, tenant slice bandwidth, RTT delay, prediction factor, etc.).
- Priority Value = IEEE 802.1P value of the frame to be transmitted OR COS queue priority value.
- Channel Id = F(S-Tag or other multi-tenancy unit).
- Credits granted = F(credits requested, RX buffers available, credits available for xchange, port shaper settings, global prediction factor, etc.)
- Credit Units = Min 512 bits to max of MTU size. (Programmable)

- Algorithmic Steps :
  1. Exchange Capabilities (Port A, B).
  2. Request Credits (Optional).
  3. Grant Credits.
  4. Transmit data.
  5. Receive data.
  6. Accounting.

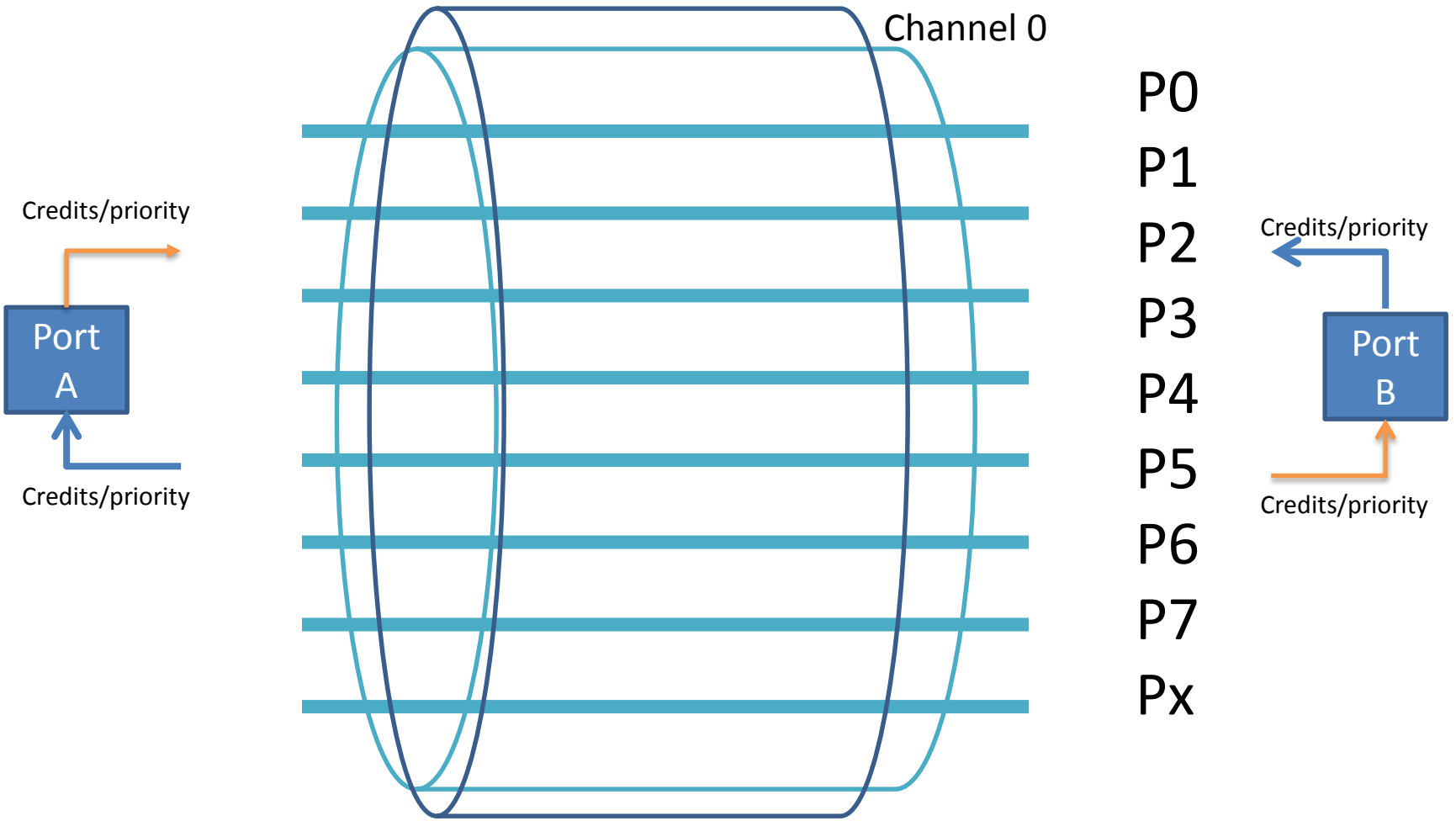
CE Frame Format



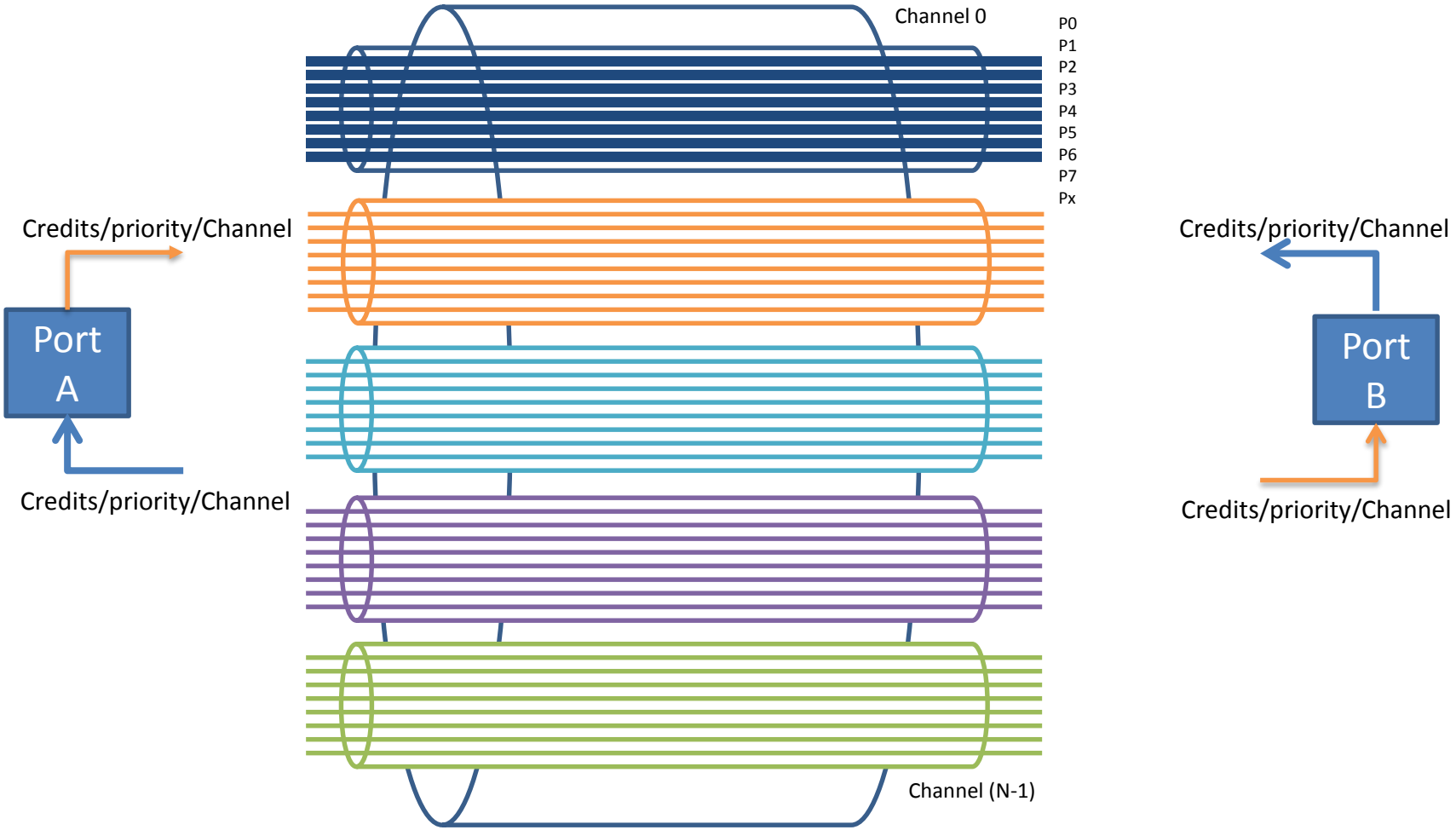
Flow Control Methods

Port Speed	IEEE 802.3X Flow Control	IEEE 802.1Qbb PFC	Credit based Flow Control (CFC)	Comments
10Mbps-1Gbps	YES	YES	OPTIONAL	
10Gbps	YES	YES	OPTIONAL	Priorities 0-7. Programmable Flow Control method per priority.
40Gbps	YES	YES	OPTIONAL	Same as above.
100Gbps	YES	YES	OPTIONAL	Same as Above. Additionally, consideration to 16 bit tenant or bandwidth slice.
400Gbps	YES	OPTIONAL	YES	Priorities 0-7. Programmable Flow Control method per priority.
1000Gbps (or 1.6Tbps ?)	YES	OPTIONAL	YES	Same as above.

Bandwidth Multi-tenancy (Default setting)



Bandwidth Multi-tenancy (programmable)



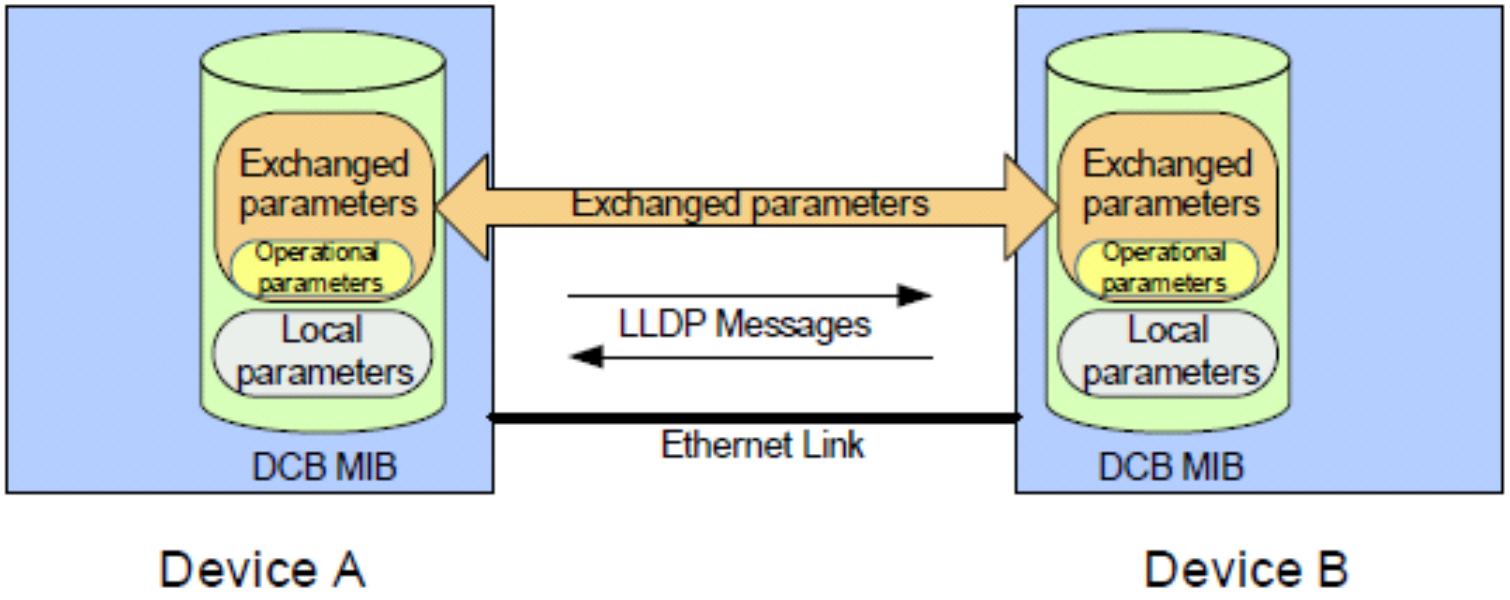
Channel Id Mapping

- Channel Id = F(C-Tag)
- Channel Id = F(S-Tag)
- Channel Id = F(VN-Tag)
- Channel Id = F(Flow)
- Channel Id = F(Tunnel)
- Channel Id = F(programmable).

Exchange of CET Capabilities

- DCBX protocol extension to exchange :
  - Credit Exchange Capabilities for flow control between peer ports.
  - Unit of credit exchange. The unit value can be in number of chunks of bits where a chunk length can be 512 bits to MAX_CR_BLK.
  - MAC_CR_BLK can be MTU size supported or 2548 bytes or any other number > 512 bits. 2548 bytes are adequate to accommodate FCOE frame with security encryption.
  - Capability to exchange channeling capabilities which enables adding 16 bit channel Ids (e.g. like SVID) in credit exchange frame or CE-Tag.

Flow Control Capability Exchange



DCBX Extension / Brief Algorithm

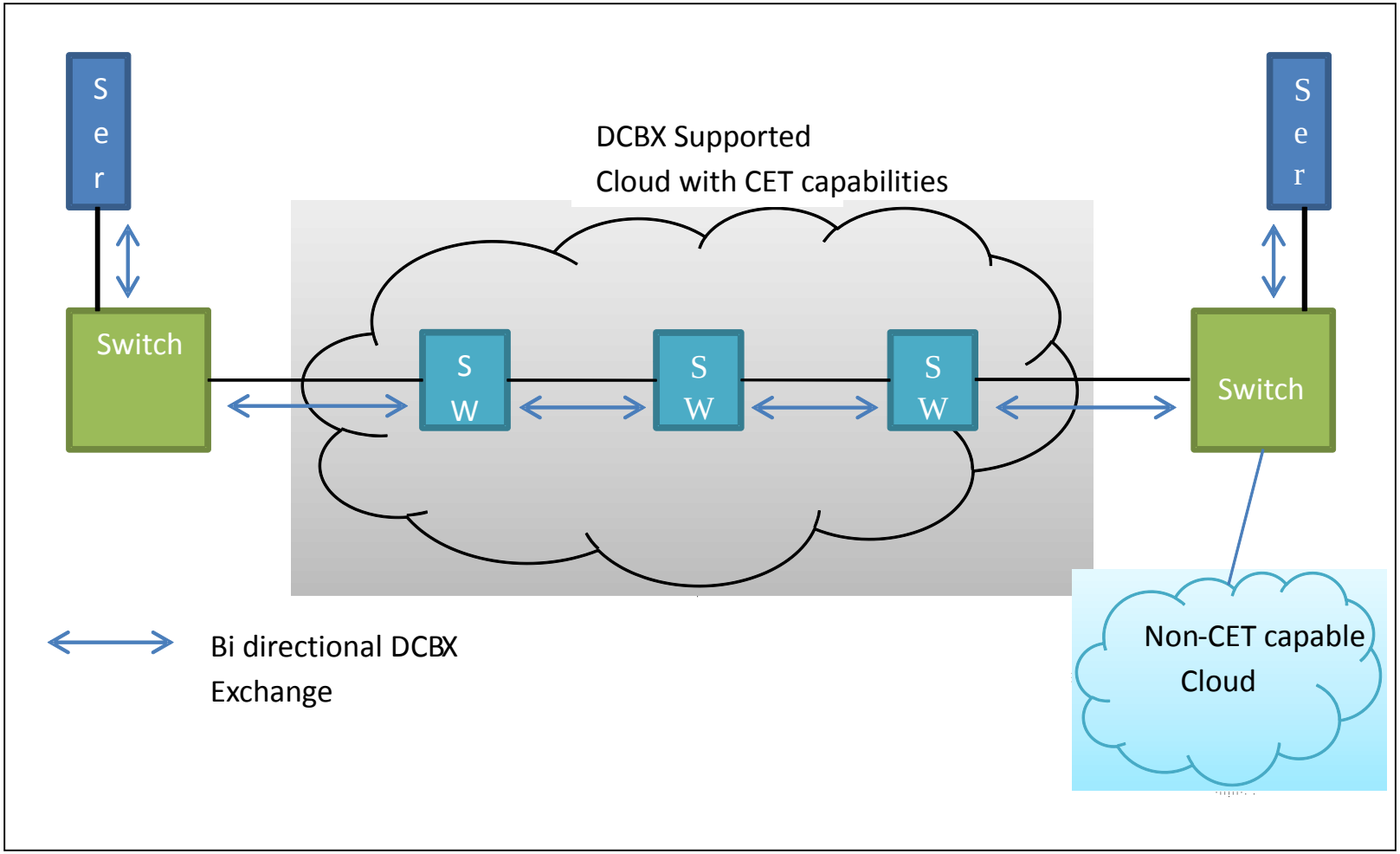
- The DCBX protocol can be extended to advertise the Credit Exchange capabilities.
- New Application Protocol TLV, “CET” should be defined.
- This TLV should be originated by a physical port at a peer to peer basis.
- Switching devices down the line should exchange such TLV messages on all of their DCBX member ports to their peer devices.
- Thus, after complete convergence, a complete path from source to the destination can understand the credit exchange capabilities of their peer ports.

BACKUP

Problem Statement

- Current CEE port based flow control works based on post queuing status on ingress.
- Ethernet needs more predictable and dynamically controllable flow control and frame acceptance mechanism with request-response mechanism.
- Interaction between peer ports before data exchange brings in more certainty and better resource utilization. Ethernet needs such a mechanism for purpose specific applications and convergence of Infiniband over Ethernet.
- Ethernet needs credit based flow control mechanism to bring in more certainty for transmission.
- Enable CEE and Metro Ethernet to have reliable I/O convergence over longer distances and simultaneously reduce buffering overheads.
- Capability to exercise PFC or CFC (Credit Based Flow Control) for select priorities.

E2E Exchange of CET Capabilities



Optional CE-Tag

1. Optionally, a new layer 2 tag named CE-Tag or Credit Exchange Tag. (Slide 5)
2. To ensure loss-less delivery of frames of one specific priority, the sender (MAC TX) adds a CE-Tagged frame or piggy backs on the data frame in transmission logic. CE-Tag contains request for credits for single priority and has low overhead. Optional !
3. Calculated amount of credits are issued to the peer ports in a CE frame or adding a CE-Tag on next data frame being transmitted.
4. Upon receipt of credits, sender sends frames for the appropriate / allowed priorities per channel.
5. Unit of credit exchange is 512 bits or each peer port can decide its own unit of credit as per it local TM / Buffer Mgr implementation.
6. Proposing an extension to the DCBX TLV to exchange capabilities of credit exchange and credit unit selection.
7. Credit aging and timers. Consideration to avoid loss of credits.

Frame Format with CE-Tag