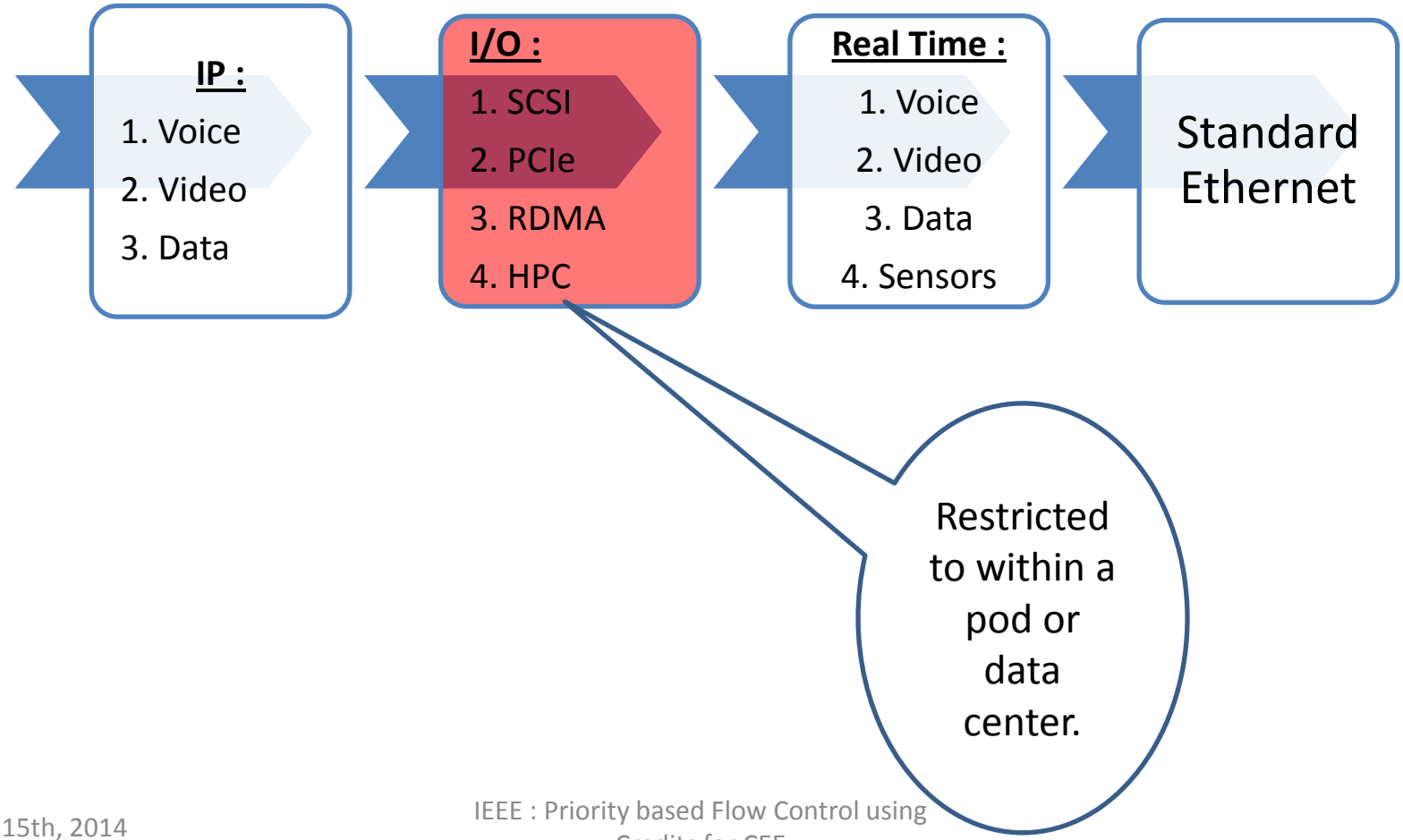
Credit based Link Level Flow Control and Capability Exchange Using DCBX for CEE ports.

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

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 interactive mechanism. Interaction between peer ports before data exchange brings in more certainty and better resource utilization.
- Need to enable CEE and Metro Ethernet to have reliable I/O convergence over longer distances and simultaneously reduce buffering overheads.
- Require very low latency transport for flash based storage protocols.

Convergence of various payload types.



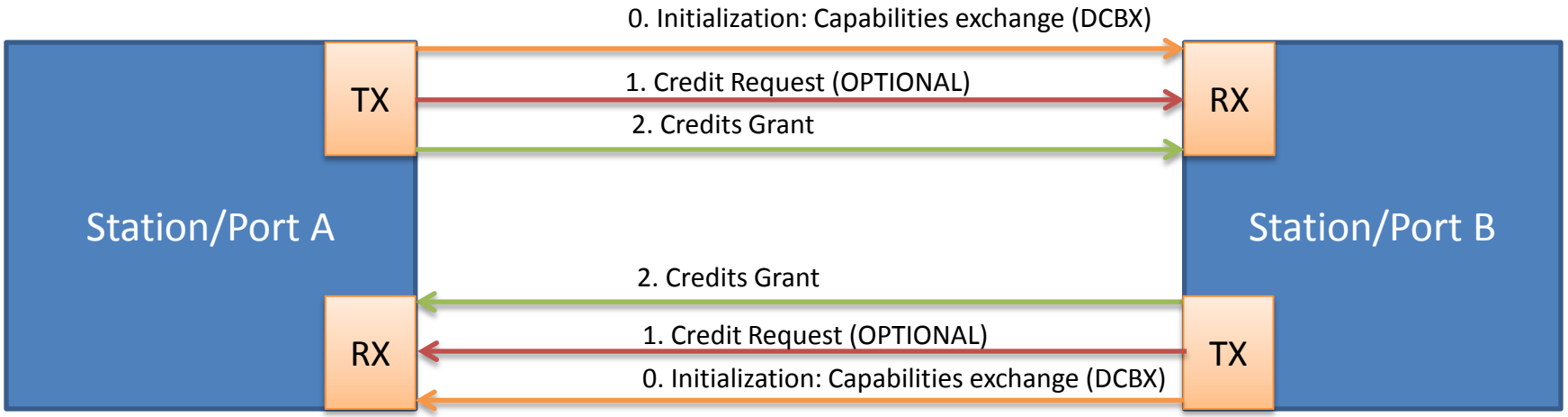
CEE Facts

- 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.

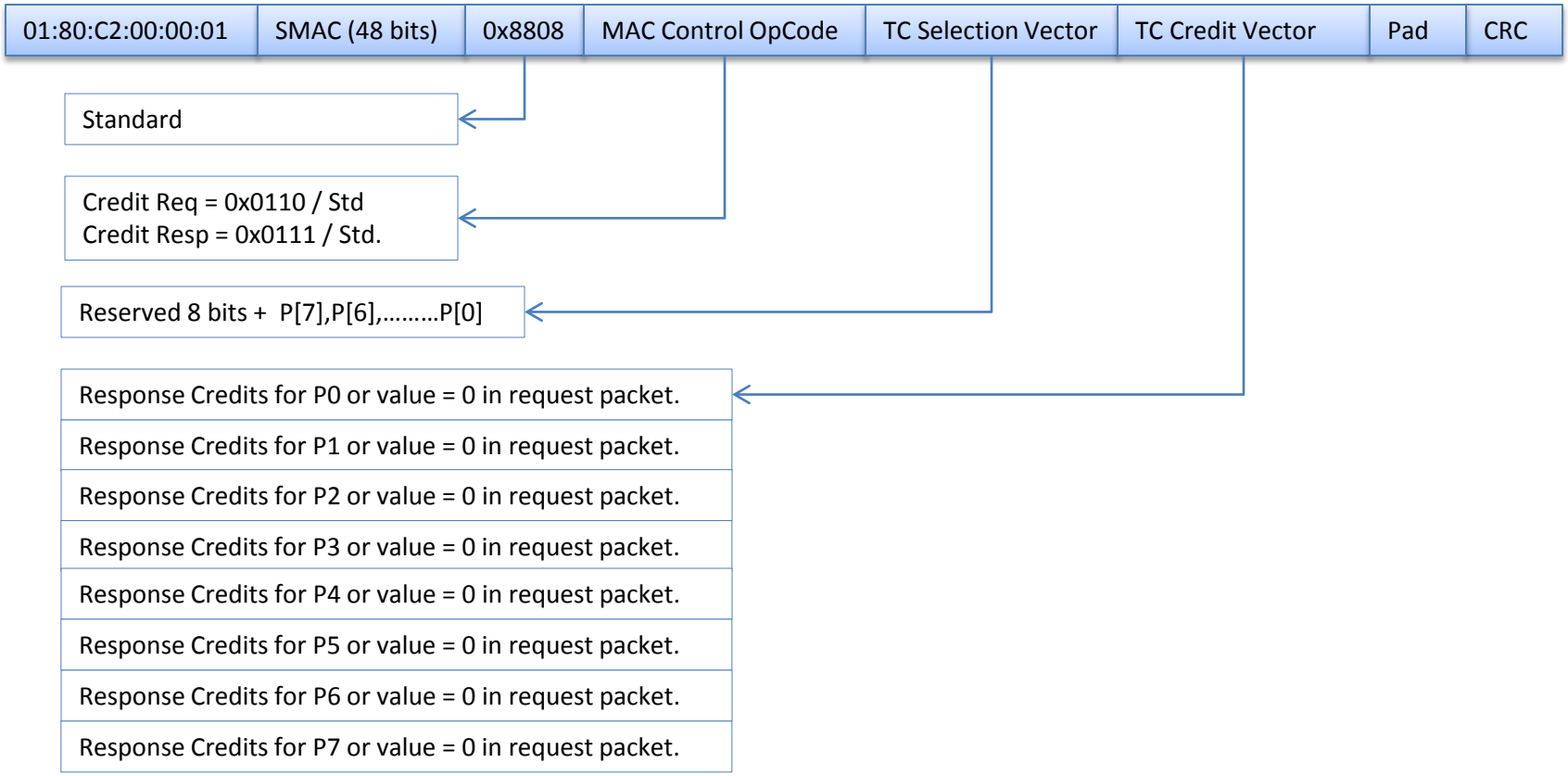
Proposed Credit Exchange Logic

- A new frame i.e. Credit Exchange Frame (CE Frame) with a MAC Control Ethernet Type 0x8808.
- The peer ports (receivers or MAC RX) receive the CE 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.
- Unit of credit exchange is 512 bits or each peer port can decide its own unit of credit as per its local Buffer Manager 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



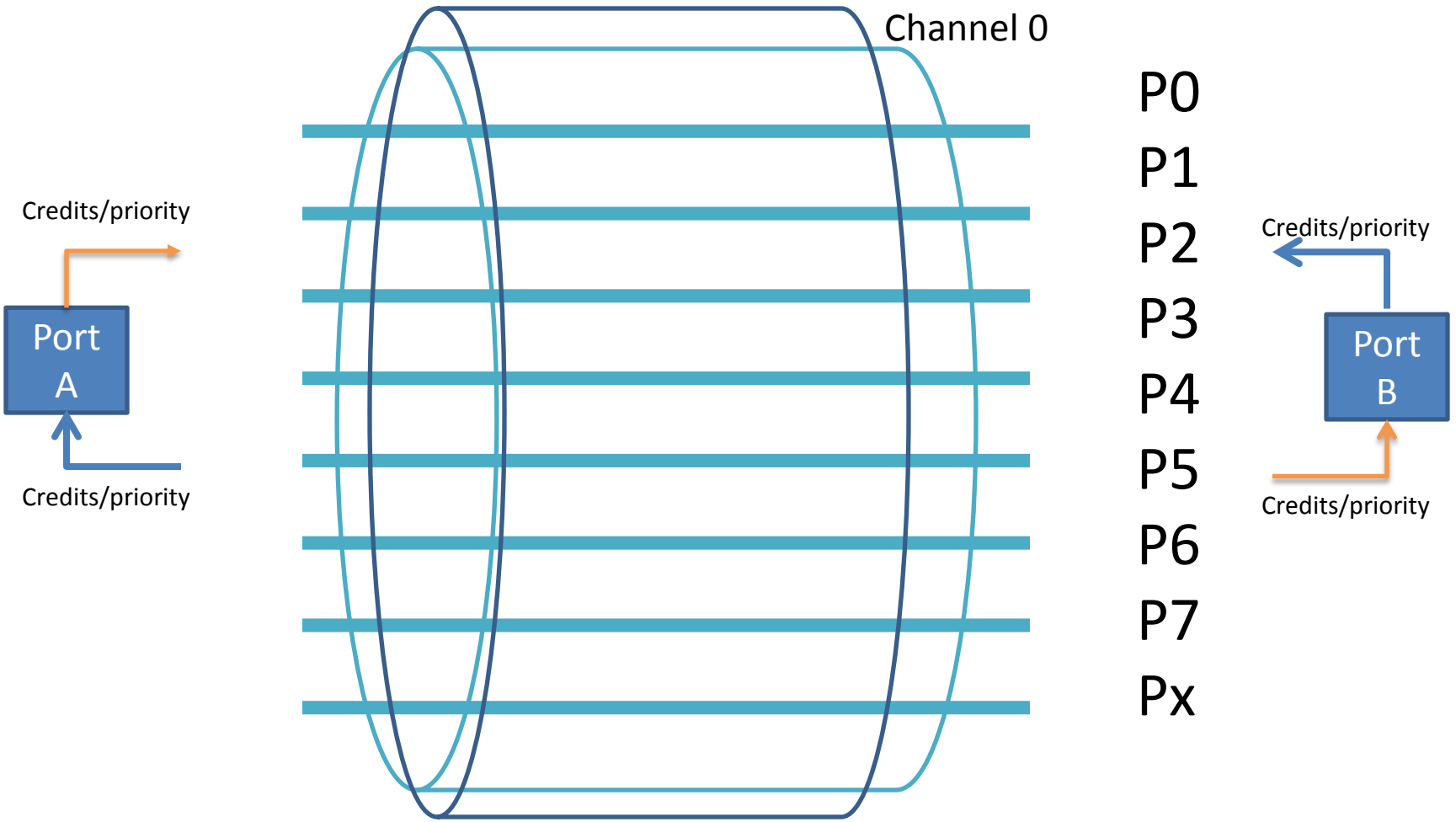
CE Frame Format



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.
400Gbps	YES	OPTIONAL	YES	Same as above.
1000Gbps (or 1.6Tbps ?)	YES	OPTIONAL	YES	Same as above.

Bandwidth Multi-tenancy (Default setting)



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