OPENFABRICS ALLIANCE

Quantized Congestion Control (QCN) Experiences with Ethernet and RoCE

Liran Liss

Mellanox Technologies

Agenda

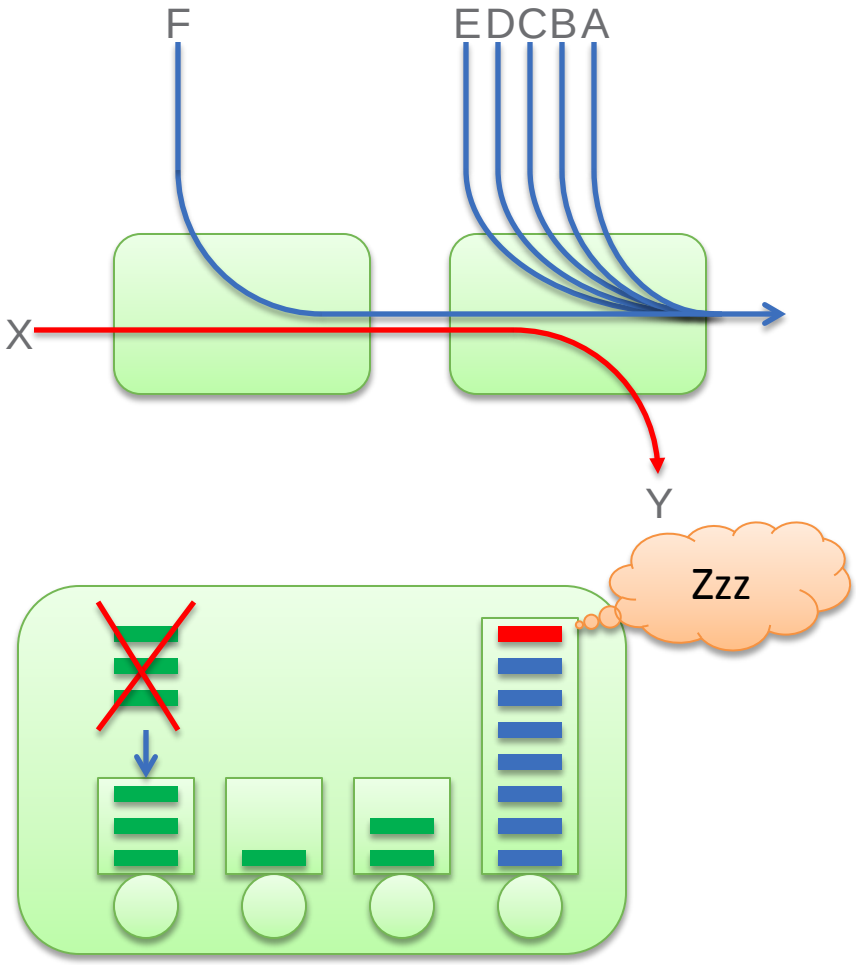
- Introduction
- Why QCN?
- Data plane
- Management
- QCN in ConnectX-3 and SwitchX
- Initial results
- Conclusions

QCN Introduction

- IEEE standard (802.1Qau)
  - Provides congestion control for long-lived flows in limited BW-delay product networks
  - Part of the Data Center Bridging (DCB) protocol suite
    - ETS, PFC, QCN, DCBX
- Conducted at L2
  - Similar to IB FECN/BECN congestion management
  - Independent from L3 congestion control
    - E.g., ECN in TCP/DCTCP
- Suitable for HW implementations
  - Simple
  - Minimal state

Why QCN?

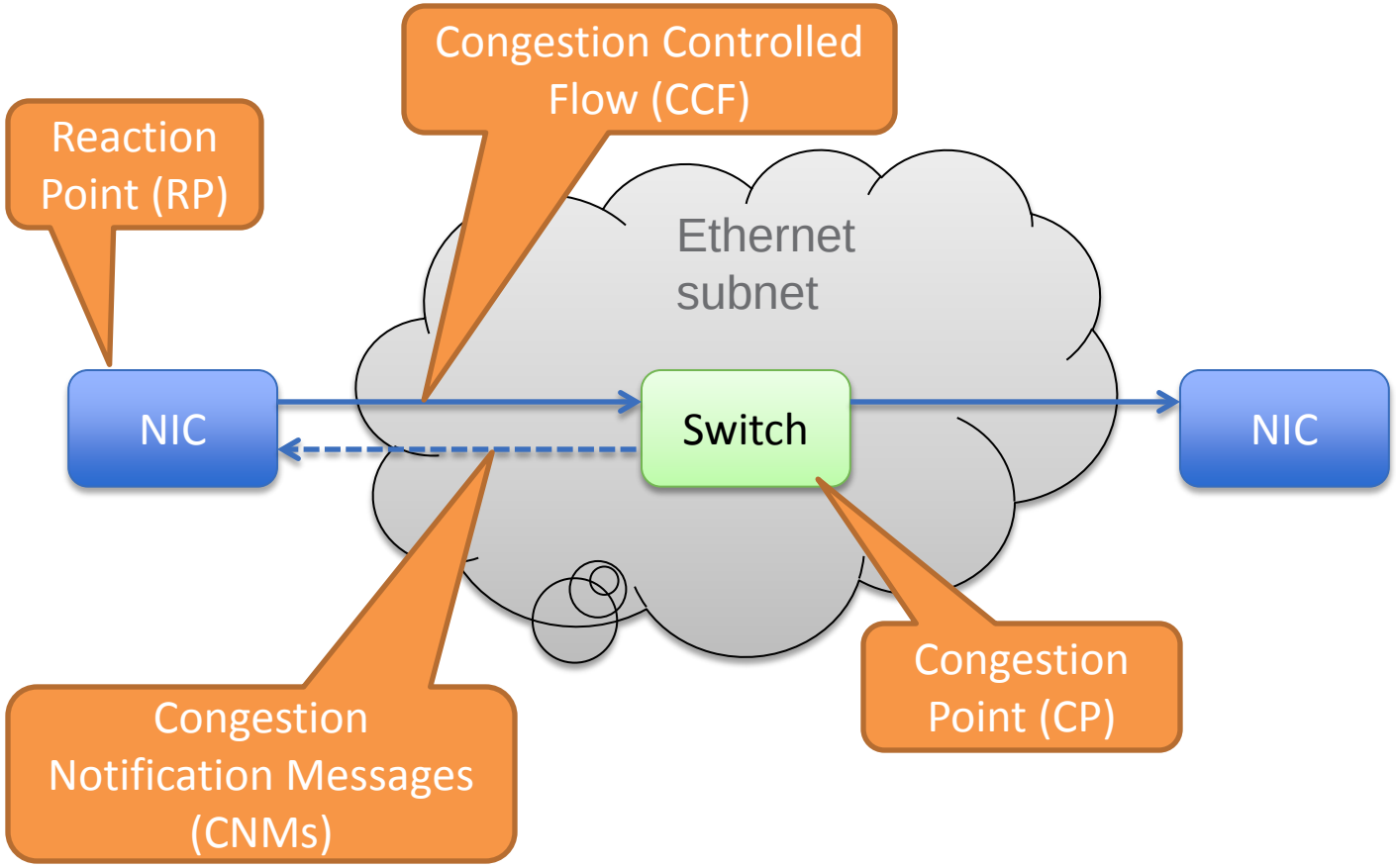
- Prevent congestion spreading
  - Mitigate victim flows
  - Important when PFC is used
- Control switch buffer utilization
  - Less impact on other flows
    - Handle bursts better
    - Mitigate incast
  - Improve latency



Tailored for DCB Environments

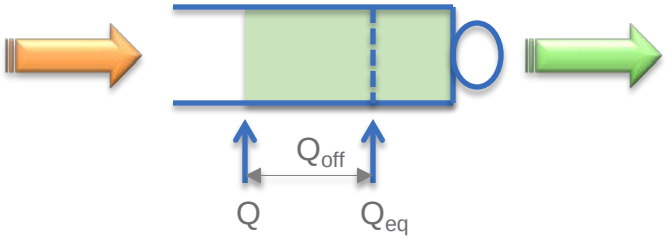
- Diverse traffic patterns
  - Partition-aggregate incast
  - Long running BW-hungry flows
  - IPC
- Diverse traffic classes
  - Best effort lossy traffic
  - Critical lossless traffic

Data Plane



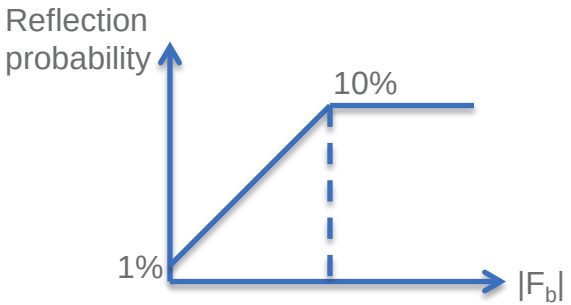
Congestion Point Behavior

- Congestion feedback computed from Switch buffer state
  - Takes into account both the degree and trajectory of congestion
- Feedback sent directly to RP
- Statistical sampling
  - Sampling probability increases with congestion

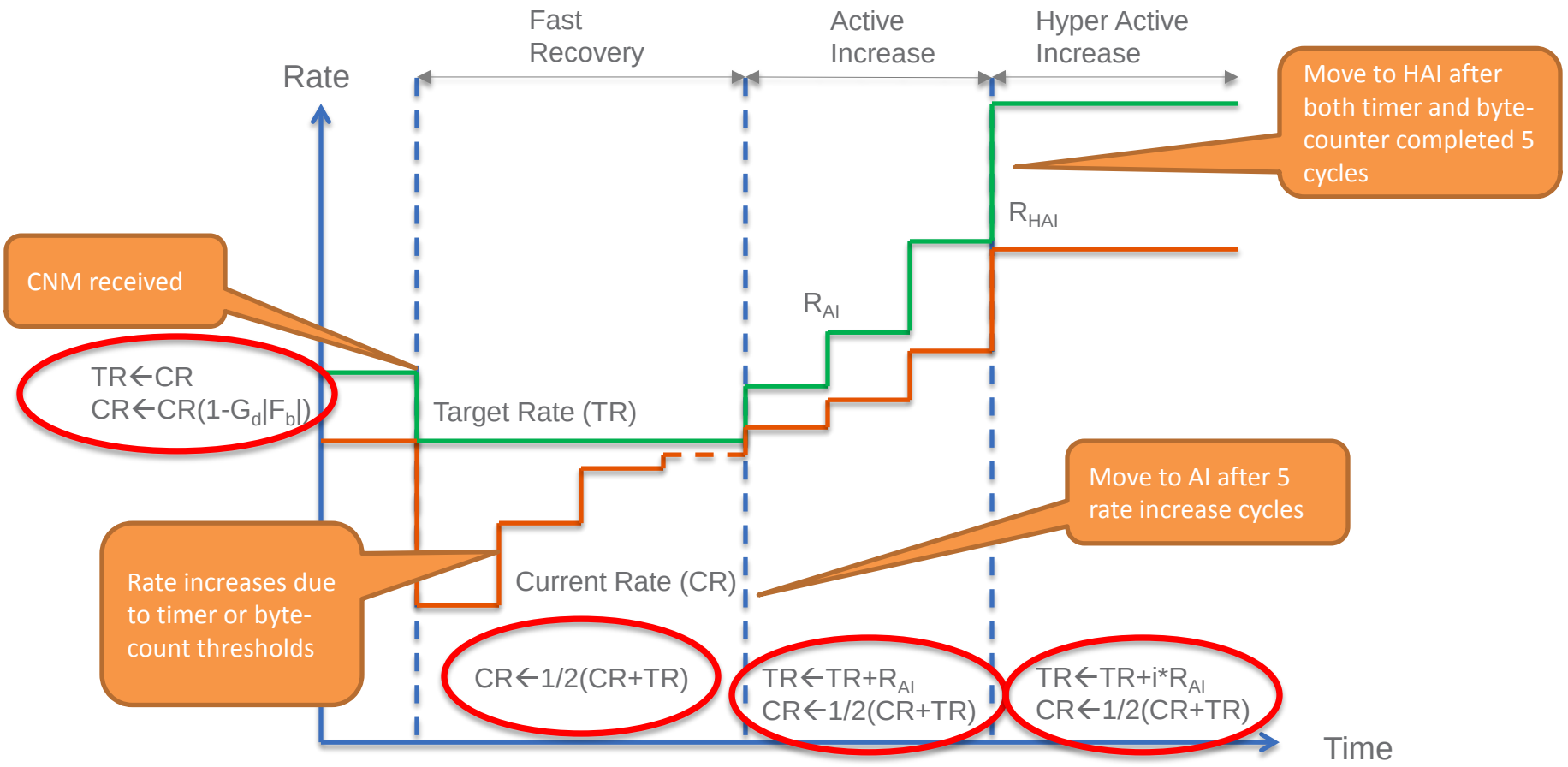


$$Q_{delta} = Q - Q_{old}$$

$$F_b = -(Q_{off} + wQ_{delta})$$



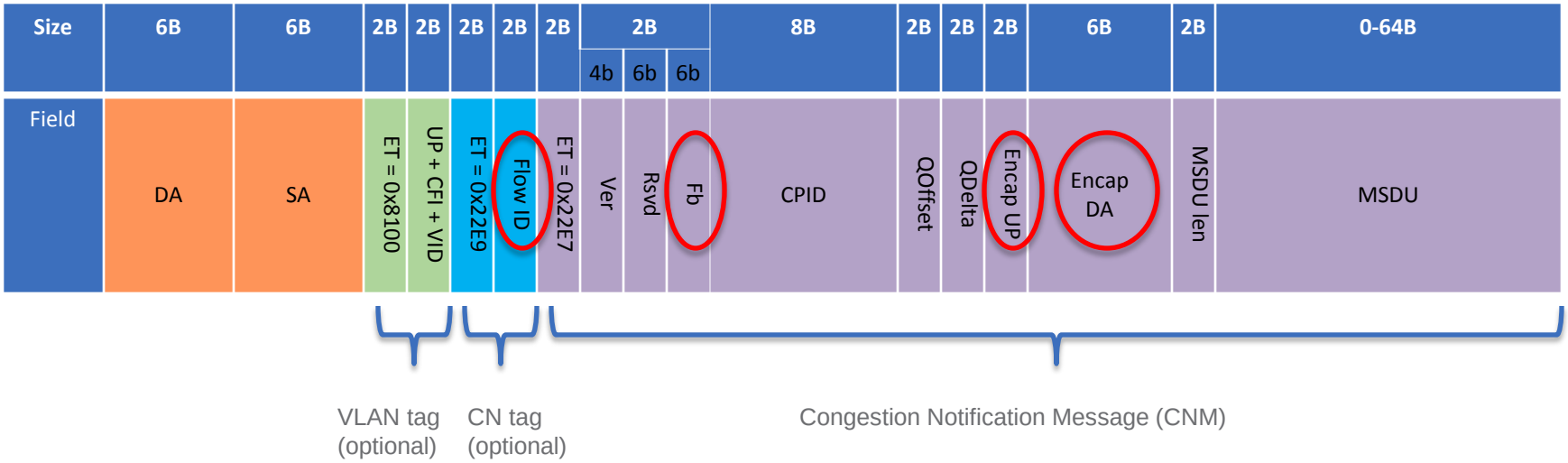
Reaction Point Behavior



Flow Segregation

- Up to 7 Congestion Notification Priority Values (CNPVs)
  - At least one User Priority dedicated for best effort traffic
- An endpoint may support multiple RPs per CNPV
  - Grouped into Reaction Point Groups (RPGs) for configuration
- Congestion Controlled Flows (CCFs) determined solely by the RP
  - By Application
  - By DMAC
  - By 5-tuple hash
  - By QP
- Optional Flow ID may be injected by RP using Congestion Notification tags (CNtag)
  - Reflected back in CNMs

CNM Frame Format



QCN Management

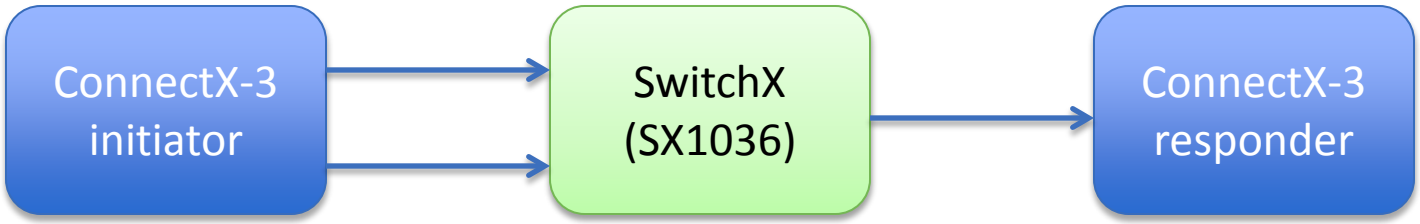
- Congestion Notification Domains (CNDs)
  - A connected subset of switches and endpoints configured to support a CNPV
- CNDs must be defended
  - Accomplished by remapping CNPVs to non-CNPV priorities
- End to end configuration required
  - May be configured manually or by DCBX

QCN in ConnectX-3 and SwitchX

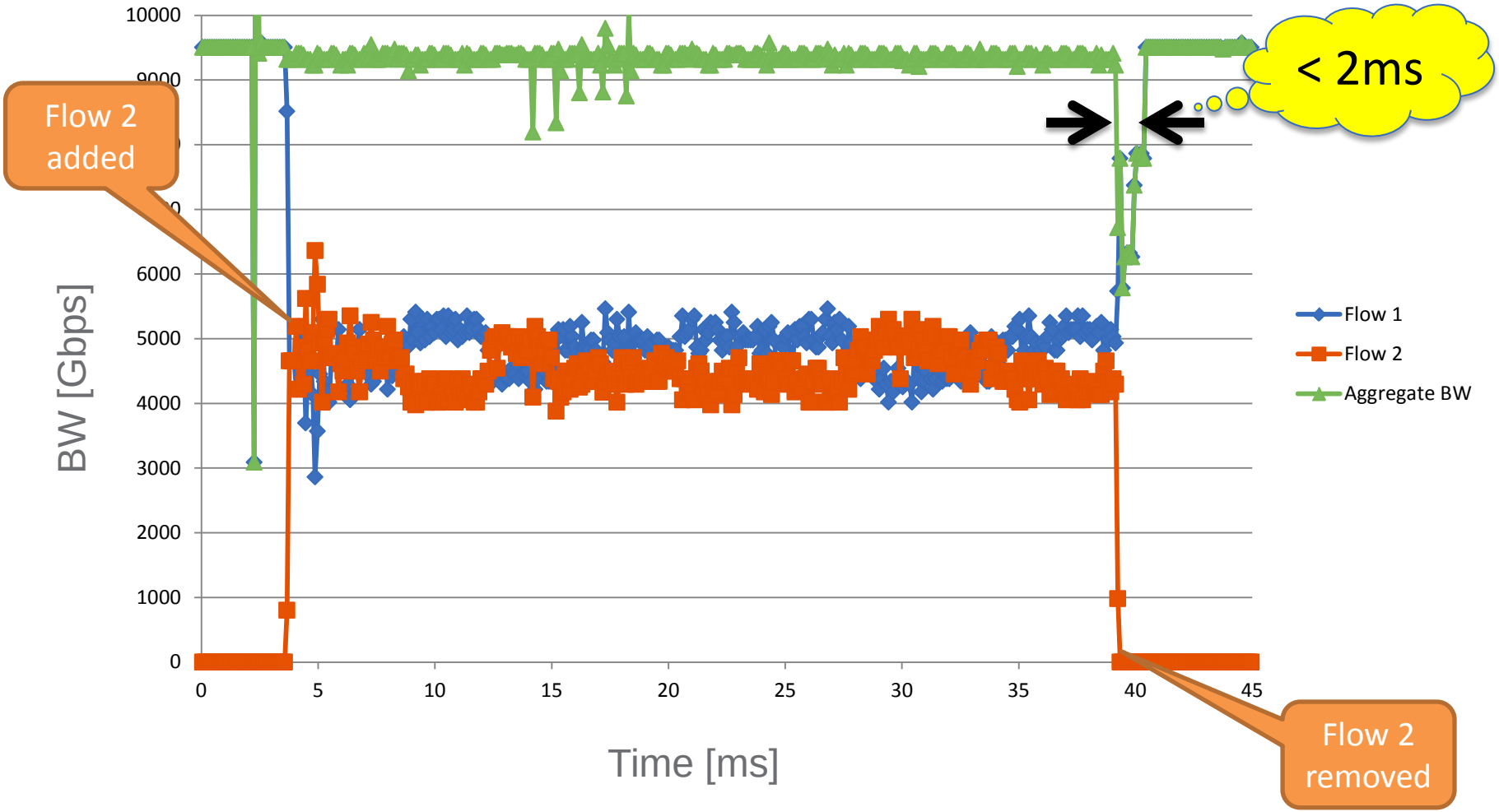
- ConnectX-3
  - HW RP implementation
  - Reaction to CNMs within usecs
  - Operation transparent to SW
    - A must for offloaded protocols such as RoCE
  - Thousands of RPs
- SwitchX
  - CP Active queue management
- Ethernet driver
  - Multiple flow queues per priority
  - Selected based on L2/3/4 hash
- RoCE driver
  - Each QP constitutes its own flow queue

Initial Testing

- Experimental setup
  - 2 ConnectX-3 HCAs
  - SX1036 switch (SwitchX)
- Manual CND configuration
- Traffic pattern
  - Both ports of requestor HCA target the same responder port
  - Tested the dynamic response to congestion



QCN Convergence



Observations

- Rapid convergence
  - Under 2ms
- Stable
  - Small oscillations
  - Negligible underflow
    - >99% utilization of congested link
  - No overflows in steady state
    - No pause frames while enabling PFC/global-pause
    - No packet loss while disabling pause

Conclusions

- QCN offers important benefits to DCB environments
  - Lossless and lossy flows can happily live together
- Efficient HW implementation in ConnectX-3 and SwitchX
  - Promising initial results
- Next steps
  - Multi-flow behavior
  - Complex setups
  - Measure real application benefits

Thank You

OPENFABRICS ALLIANCE