An Open Congestion Control Architecture with network cooperation for RDMA fabric

draft-zhh-tsvwg-open-architecture-00
draft-yueven-tsvwg-dccm-requirements-00

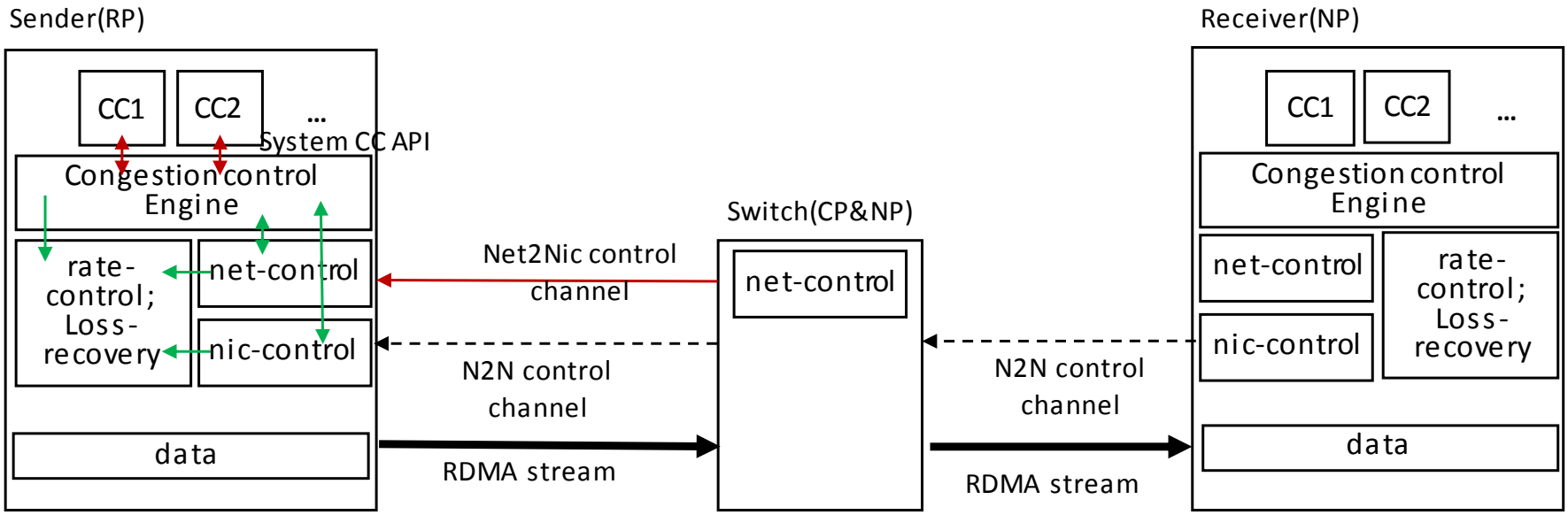
IETF 105, Montreal, Canada

Rachel Huang (Presenter), Yan Zhuang
Yu Xiang, Roni Even
Huawei Technologies

An open congestion control architecture with network cooperation for RDMA fabric

- **Scope**
  - Managed datacenter networks
  - RDMA traffics for applications, such as HPC and storage….requiring low latency, high throughput…
- **Motivation, requirements and use cases**
  - Incast traffic suffers from congestion in the network.
  - Mixture of RDMA traffic and TCP traffics effects each other.
  - More efficient and effective congestion controls are needed to support the scalability and high performance.
- **Objectives**
  - Define an open congestion architecture with network cooperation to enable more effective congestion controls for RDMA fabrics.

Architecture Overview

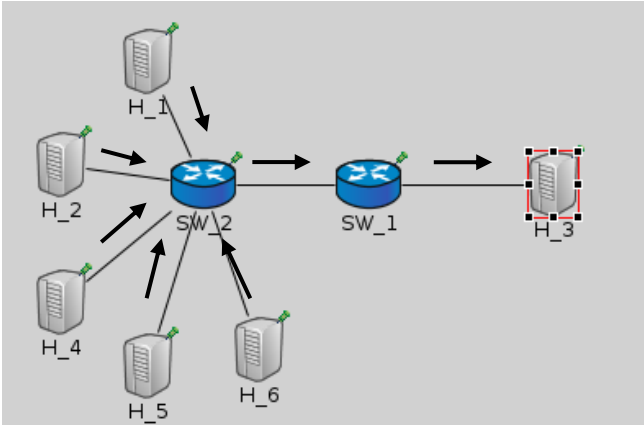


- Open to congestion control deployment and management
- Open to network cooperation

Open for Network Cooperation

- **What?**
  - Net-control module inside network nodes (e.g. switches) can signal back to senders' NIC, and further incorporated into NICs' transmit rate control.
- **Why?**
  - **Fast Convergence**: reduce the CC feedback/control time.
  - **Accurate congestion awareness:** as congestion point, network aware of the degree of the ongoing and expected congestion and can requests for proper moderation of the selected flows.
- **How?**
  - A Net2Nic control message can be used to report congestion information from the network nodes to sender NICs.

Initial Experiment on Open Network Cooperation



Simulation Environment：

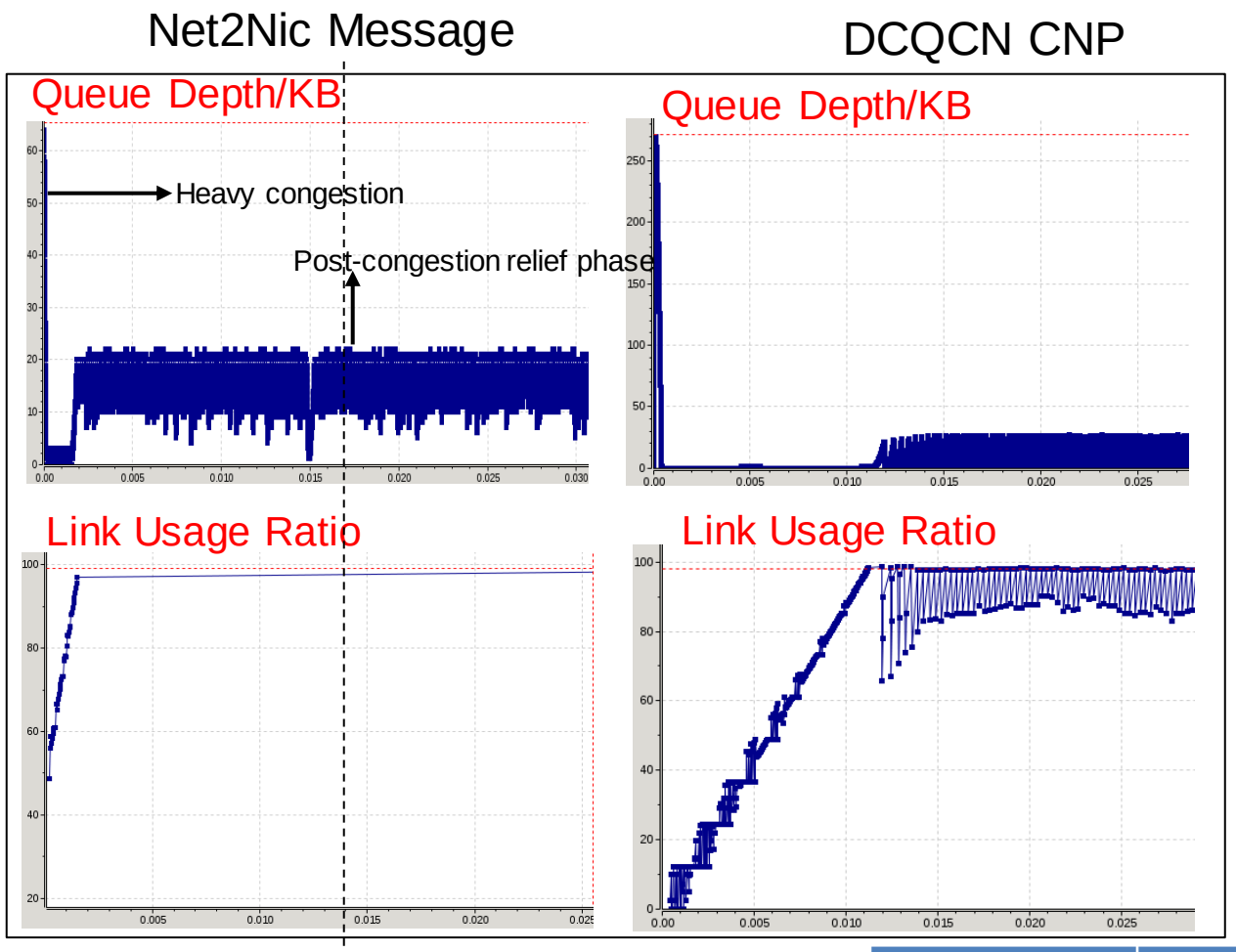
- H_1, H_2, H_4, H_5, H_6 each sends 10MB data to H_3 simultaneously.
- Each port of SW_2 is 10Gbps
- ECN Threshold：20Kb PFC Threshold：300Kb

- Net2Nic Message：SW_2 Sends messages from network. DCQCN is CC algorithm.
- DCQCN CNP：DCQCN mechanism. Sending CNP from H_3.

■ Parameters Setting：

Mechanisms	Parameter settings	
	Feedback Message Interval /us	Sending Rate increase interval /us
Net2Nic Message	10	15
DCQCN CNP	50	55

Result Comparison



Result:

- Heavy congestion: Net2Nic Message can prevent overshoot (prevent triggering PFC). After congestion is relieved, the rate can be quickly restored to ensure throughput and bandwidth utilization.
- Slight congestion: Net2Nic Message prevention of microburst issues (Excessive suppression, hard to guarantee throughput, and high bandwidth usage fluctuation)

Mechanism	Average FCT(us)
Net2Nic Message	40296.02
DCQCN CNP	47434.08

- Net2Nic Message FCT gains **15%**

Open for Congestion control deployment and management

- **What?**
  - Deploy/manage congestion control algorithms in a common way based on the traffic patterns as well as the network resources regardless of the detailed hardware implementation.
- **Why?**
  - **More flexibility:** Traffic patterns may differ in CC choices.
  - **Easy to deployment in HW**: New CC algorithms are suggested to be implemented in hardware.
- **How?**
  - Provide a system CC interface to the operators to deploy CCs through a common platform and then be mapped to local actions/functions.
  - Local functions related to congestion controls can be implemented as function blocks and interact with each other through internal interfaces to achieve the final congestion controls.

Next Step

- Solicit more feedbacks/comments/interests on this open architecture.