Empowered by Innovation NEC

10G QCN Implementation on Hardware

2009 November 17th

Masato Yasuda, Noriaki Kobayashi, Kiyohisa Ichino (NEC)

Abdul Kader Kabanni, Balaji Prabhakar (Stanford University)

Overview

- It’s the world’s first 10G QCN Full implementation
- Features
  - 10G QCN-NIC, QCN-Switch
    - Compliant with Pseudo Code v2.3 [1]
  - Pause Function
- We show results at the following condition
  - 10G QCN
  - 10G QCN with Delay
  - 10G Pause (No QCN)
  - 10G QCN + Pause

[1] QCN Pseudo Code v2.3: http://www.ieee802.org/1/files/public/docs2009/au-rong-qcn-serial-hai-v23.pdf

Background

Previous Implementation (1G Platform)

- IEEE Interim Meeting in May
  - First QCN Full Implementation
- IEEE Plenary Meeting in July
  - QCN + Pause, QCN + TCP

Our Implementation

10G QCN Hardware

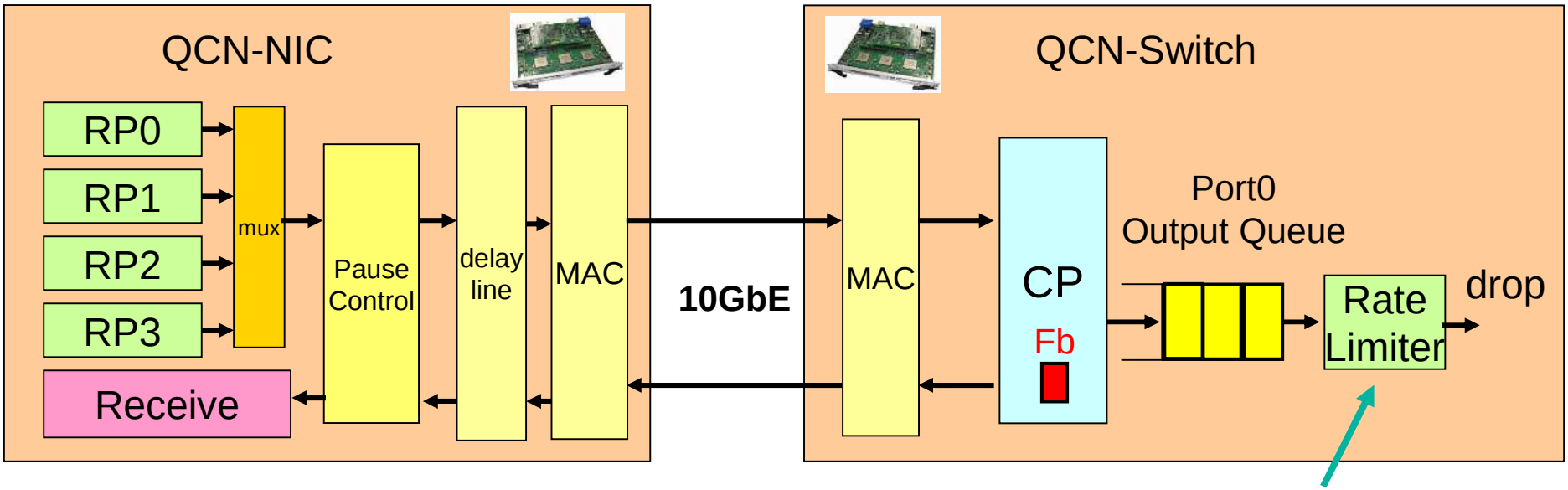
Porting from 1G QCN to 10G

- Changed the data bandwidth on Data PATH
- Control Plane is the same to 1G: Working on the same algorithm

Hardware Platform : NEC's 10G custom FPGA Board

- Advanced TCA
- FPGA: Altera Stratix II
- Number of 10G Ports: 1

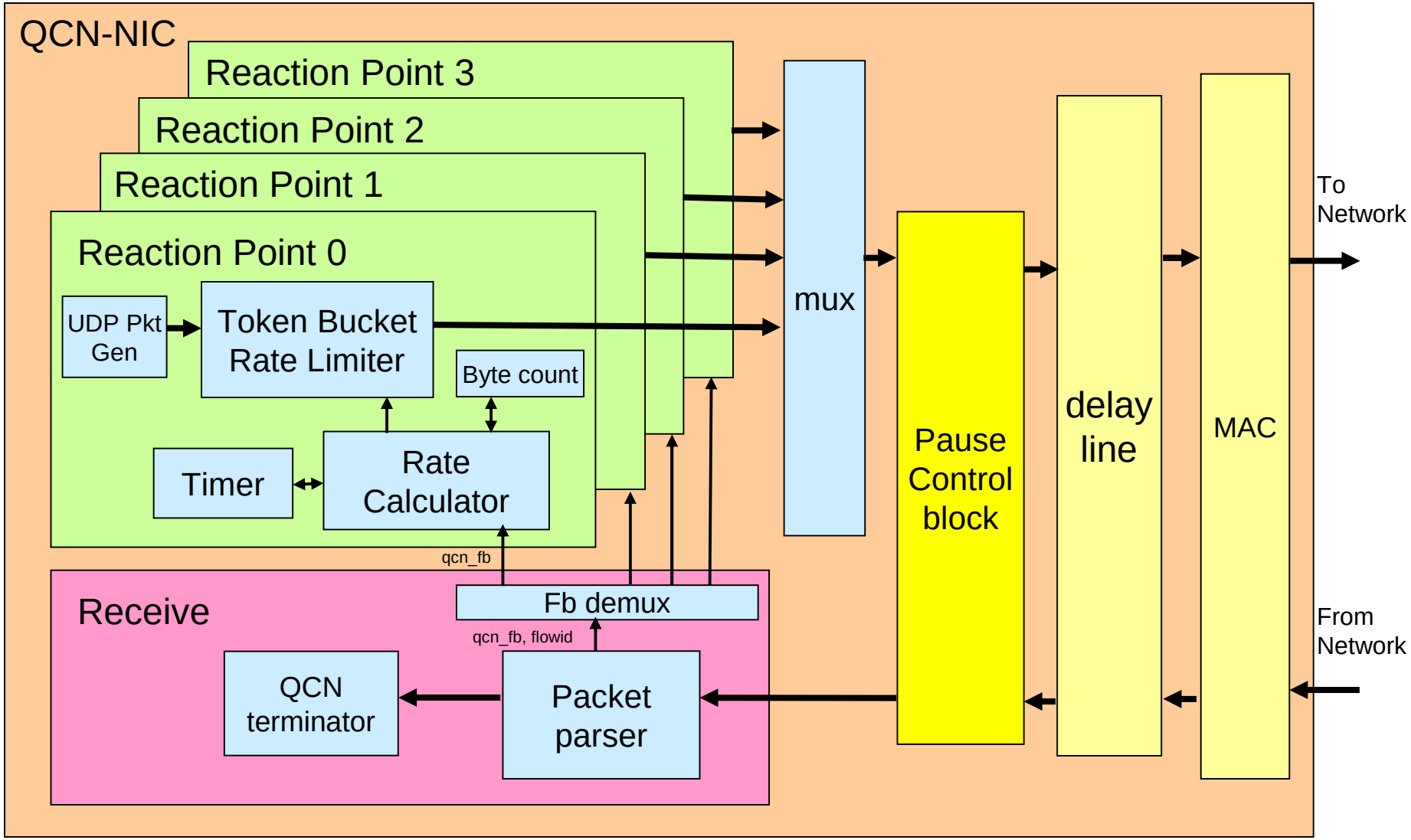
10G QCN Evaluation System



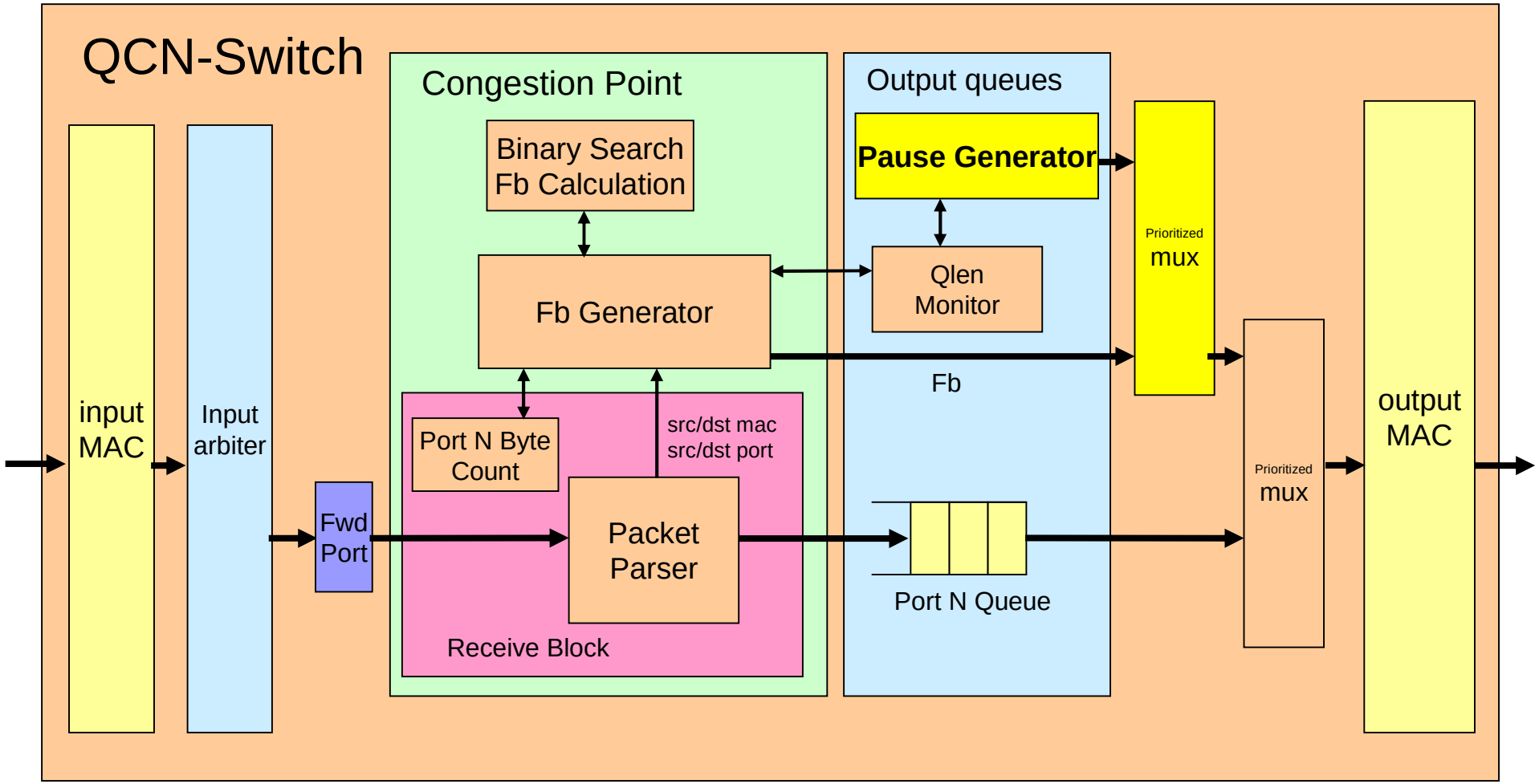
Limit Rate (9.5G→2G→9.5G) to check flow control behavior

- Number of Reaction Points (RPs): 4
  - Each RP generates 10Gbps UDP Traffic
- Limit Rate at Switch:
  - 9.5Gbps(3sec) → 2Gbps(3sec) → 9.5Gbps(3sec)

10G QCN NIC Architecture



10G QCN Switch Architecture



Evaluation Condition

Parameters (QCN)

NIC

- FAST_RECOVERY_THRESHOLD = 5
- AI_INC = 5 Mbps
- HAI_INC = 50 Mbps
- BC_LIMIT = 150 KB (30% randomness)
- TIMER_PERIOD = 10msec (30% randomness)
- MIN_RATE = 5 Mbps
- GD = 1/128
- Additional Round Trip Delay (by Delay-Line) = 0 usec or 200 usec

Switch

- Max Queue Size= 150KBytes
- Quantized_Fb: 6 bits
- Q_EQ = 33 KB
- W = 2
- Base marking = 150 KB, and varies according to the lookup table in the pseudo code (30% randomness)

Parameters (Pause)

Pause

- Pause at Watermark_hi = 130KB
- Unpause at Watermark_lo = 110KB

Paused RPs

- Freeze timer & timer_scount
- Rate limiter stops sending packets but keeps adding tokens based on the crate value
- Obey all Fb messages:
  - Reset timer, *timer_scount, tx_bcount, si_count*
  - Decrease crate etc

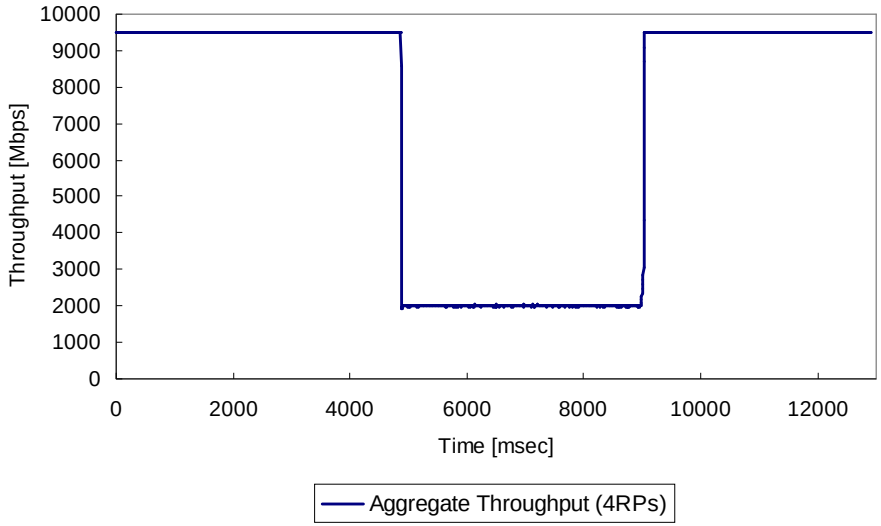
Evaluation Result

10G QCN

Evaluation Result (Hardware)
(4-RPs, QCN, Delay Line : disabled)

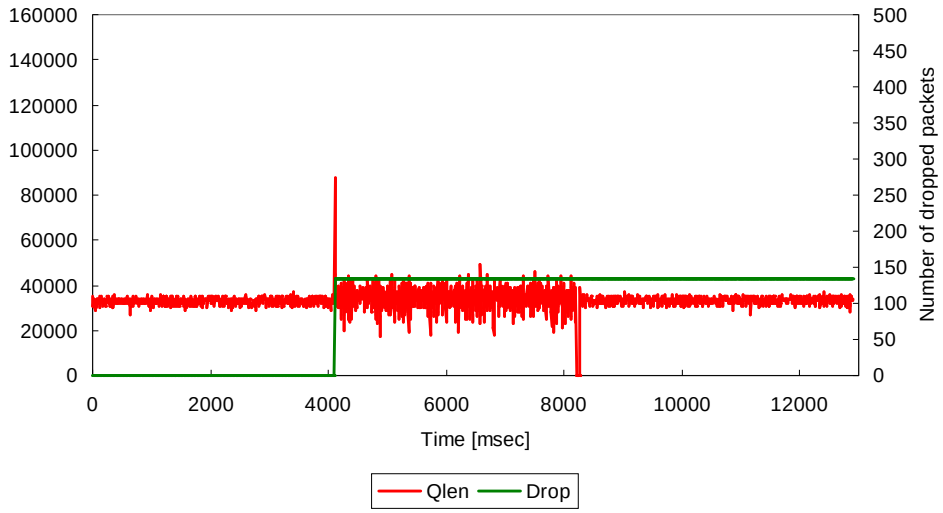
Data sampling interval in HW = 5msec

NIC: Aggregate Throughput



Recovery Time: 70 msec

Switch: Queue Length & Dropped packets



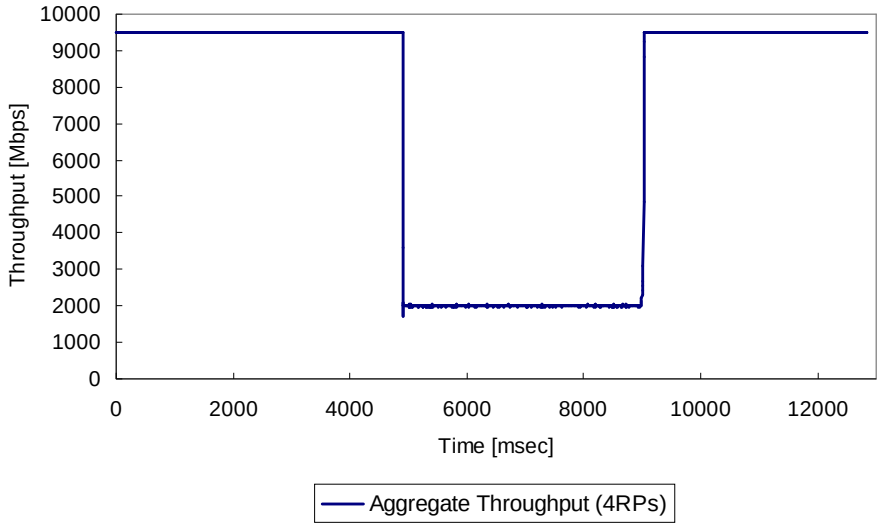
(Qeq=33KB, Max Qlen=150KB)

10G QCN with Delay

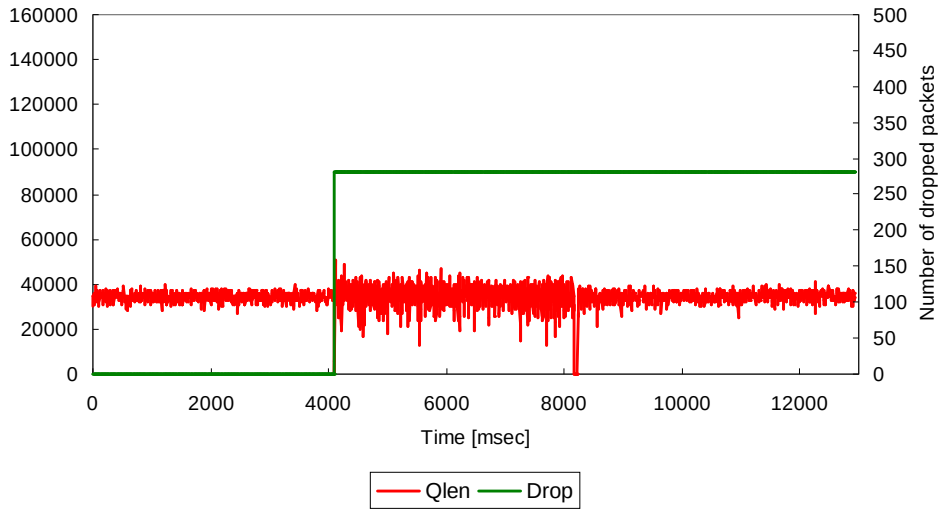
Evaluation Result (Hardware)
(4-RPs, QCN, Delay Line: 200usec roundtrip delay)

Data sampling interval in HW = 5msec

NIC: Aggregate Throughput



Switch: Queue Length & Dropped packets



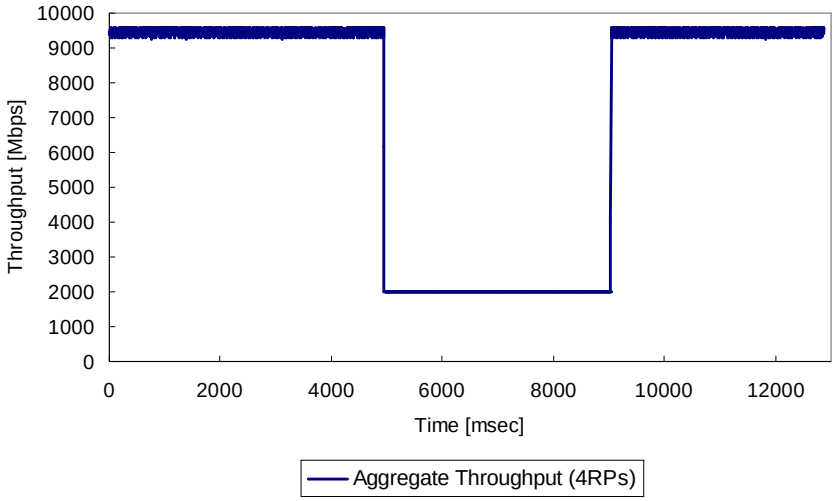
(Qeq=33KB, Max Qlen=150KB)

10G Pause (No QCN)

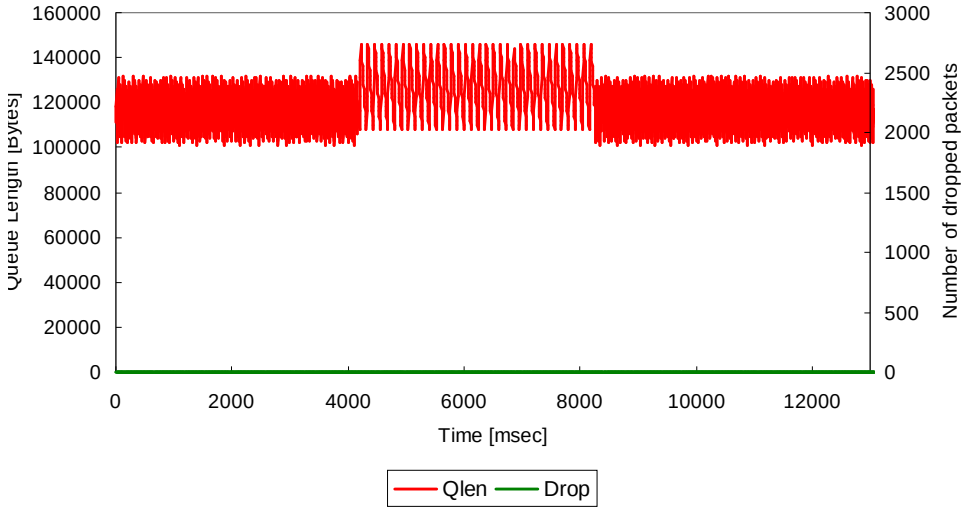
Evaluation Result (Hardware)
(4-RPs, Pause (No QCN), Delay Line: disabled)

Data sampling interval in HW = 5msec

NIC: Aggregate Throughput



Switch: Queue Length & Dropped packets



Zero Packet Drop

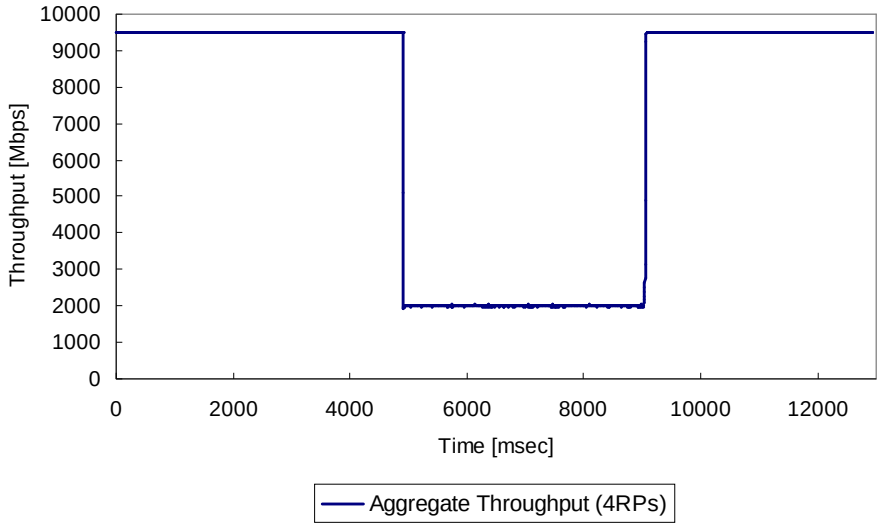
(Watermark_hi= 130K, Watermark_lo= 110K)
(Max Qlen=150KB)

10G QCN + Pause

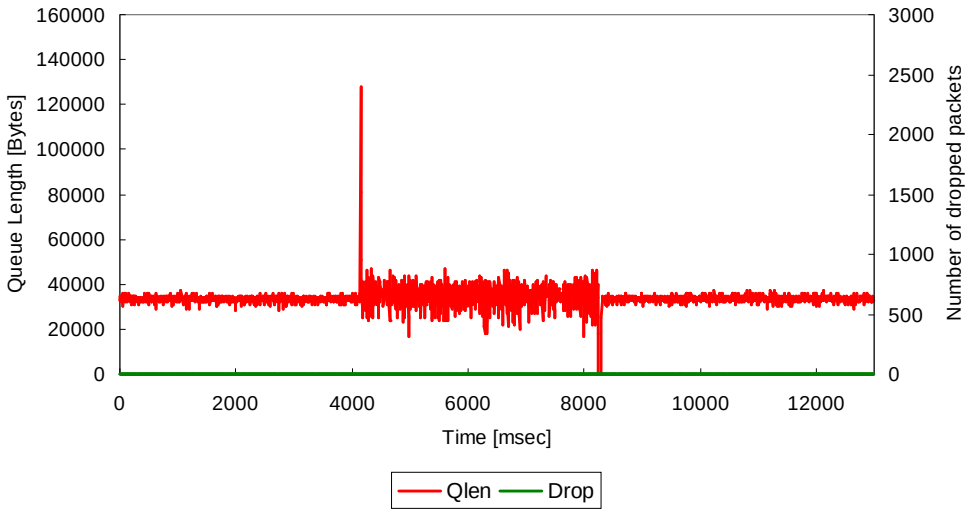
Evaluation Result (Hardware)
(4-RPs, QCN + Pause, Delay Line: disabled)

Data sampling interval in HW = 5msec

NIC: Aggregate Throughput



Switch: Queue Length & Dropped packets



Zero Packet Drop

(Watermark_hi= 130K, Watermark_lo= 110K)
(Qeq=33KB, Max Qlen=150KB)

Conclusion

- We have finished 10G QCN implementation and showed the result on real Hardware.
- We confirmed the following function are working fine.
  - 10G QCN
  - 10G QCN with Delay (RTT: 200usec)
  - 10G Pause (No QCN)
  - 10G QCN + Pause
- We confirmed Zero-loss delivery at QCN + Pause.
  - It will also be effective for Congestion Spreading.