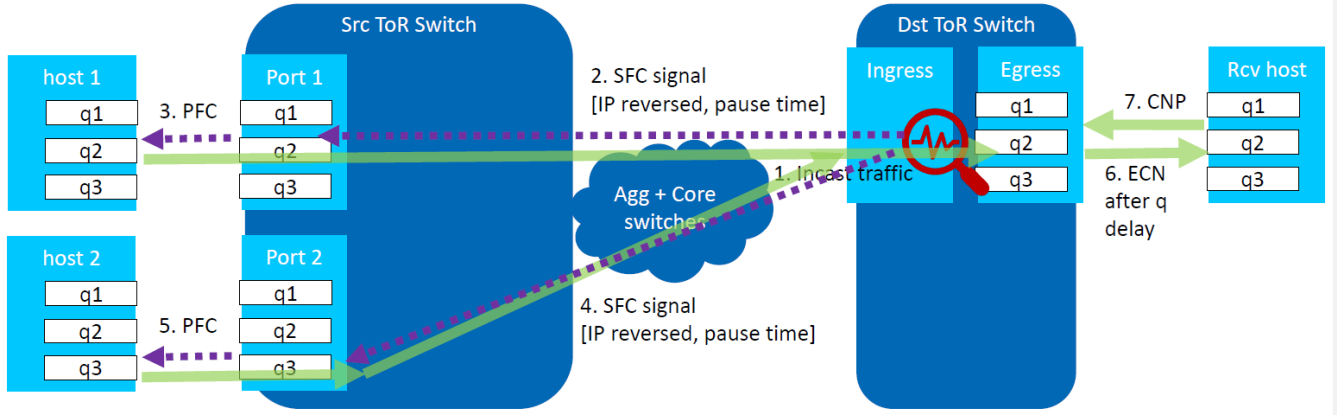
Source(Remote) PFC Test

Lily Lv , Jinfeng Yan, Peiying Tao(Huawei)

Background

- A previous contribution “Source Flow Control (SFC)” was presented in Nendica.
  - https://mentor.ieee.org/802.1/dcn/21/1-21-0055-00-ICne-source-flow-control.pdf
  - The idea is for a congested switch to send a signal to the source TOR, triggering PFC mechanism on the source TOR.

- Unlike legacy PFC, which is triggered locally by the internal threshold (XON/XOFF) of the switch, the new method invokes PFC differently. We call it ‘remote PFC’.

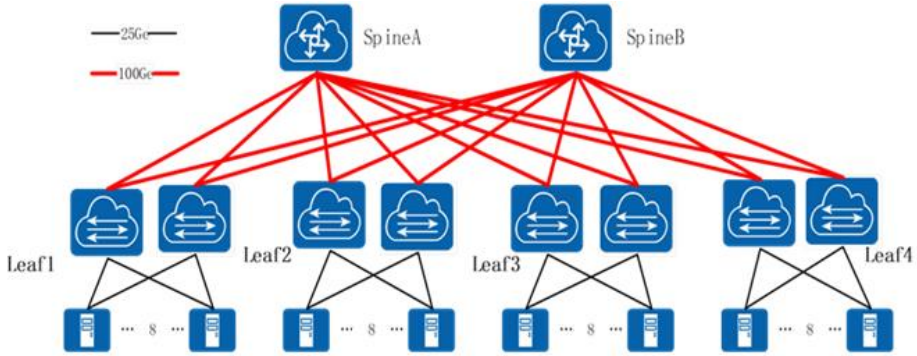


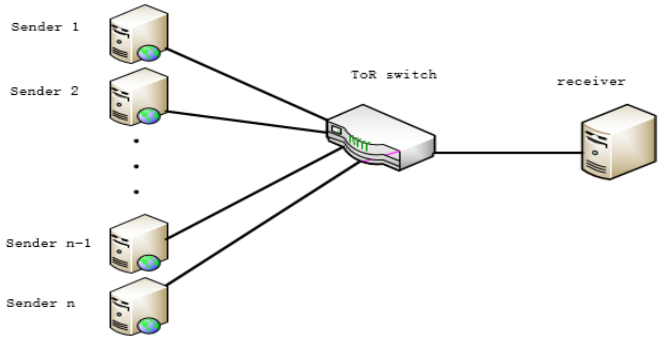
From “1-21-0055-00-ICne-source-flow-control.pdf”

- Field tests have demonstrated the benefits of remote PFC.
- This presentation explains the field test, and our test results.

Issue in Field Test

- Networking
  - 32 server, 4 groups of TOR
  - 25Gbps link between TOR and server, 100Gbps link between spine and leaf
  - DCQCN is activated
- Traffic Model
  - TCP/ROCE 9:1 mixed traffic
  - 7 to 1 incast or 32 full mesh
- Issue
  - Output port of TOR switch is heavily congested.
  - Latency increases a lot, to ms level.





#bytes	#iterations	t_min[usec]	t_max[usec]	t_typical[usec]	t_avg[usec]
4096	1000	4028.99	4697.27	4691.91	4691.74

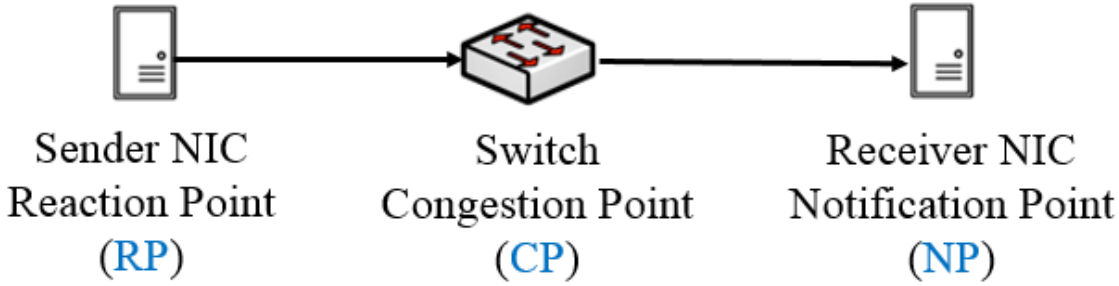
Buffer Usage on each Queue: (cells/KBytes)

QueueIndex	Current	Peak	Average
0	0/0	0/0	0/0
1	0/0	0/0	0/0
2	0/0	0/0	0/0
3	47/11	250/62	39/9
4	0/0	0/0	0/0
5	9709/2427	47046/11761	16743/4185
6	0/0	1/0	0/0
7	0/0	0/0	0/0

<DUT-Leaf1A>

Issue Analysis

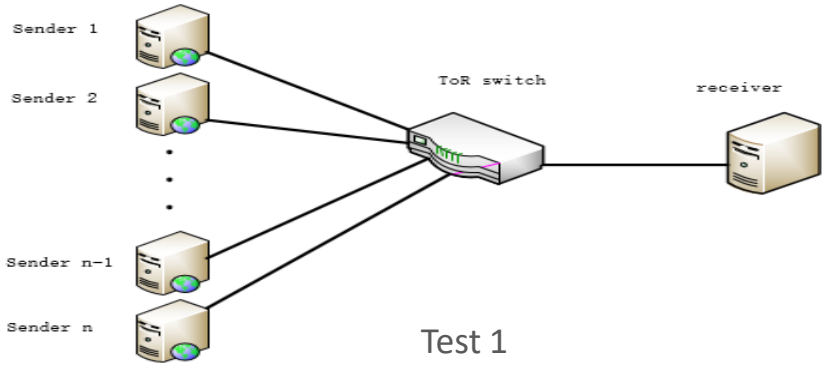
- DCQCN principle
  - CP sets ECN mark in packets when congested
  - NP sends CNP to RP when receiving packets with ECN mark
  - RP reduces the flow rate when receiving CNP
  - RP recovers the flow rate when not receiving CNP for a certain time (timer R)

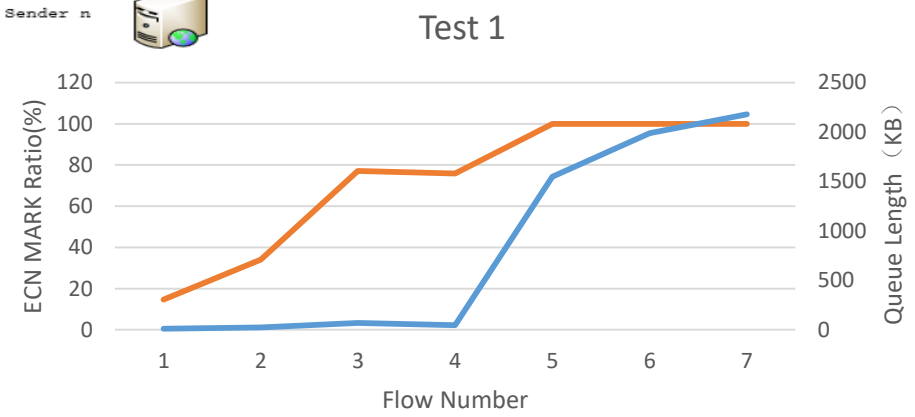


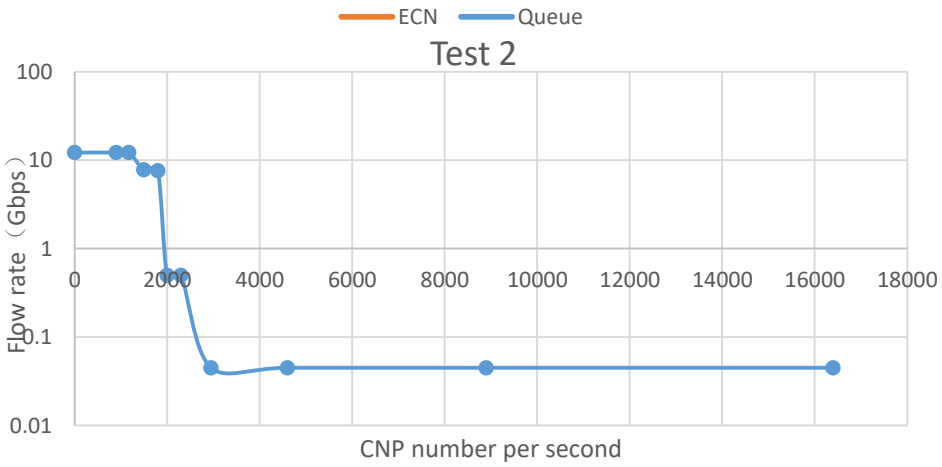
- If RP does not receive CNP in time due to large scale traffic (large number of flows), making the interval of CNP bigger than timer R, the flow causing congestion will increase the rate which creates more congestion.
  - Example: 32 full mesh traffic, 4KB size packet
    - Flow rate = 25G x 10% x 1/31 = 80Mbps
    - Every 400us (4KB/80Mbps = 400us), there is one ECN/ CNP to control the flow rate
    - Default value of timer R is 300us, which pressures the CP and causes the latency issue.
    - If increase timer R, the low speed of recovery may cause throughput issue. Hard to find proper timer R value.
- Other factors which may cause the issue
  - NP NIC capability of CNP generation
    - Constrained by hardware and software implementation, CNP generation speed is limited, e.g. 1us
  - RP NIC capability of flow rate control
    - Constrained by hardware and software implementation, lowest rate of each flow is limited, e.g. 45Mbps

ECN/CNP Adjustment Does Not Help in Large Scale

- **Observed phenomenon 1:** the relationship between queue length, number of flows at each sender and the ECN mark ratio
  - Precondition
    - 7 to 1 incast
    - Adjust flow number and ECN mark ratio
  - ECN mark ratio is increased with increase of flow number.
  - When flow number is 5, ECN mark ratio reaches 100%. ECN does not help to control queue length.
- **Observed phenomenon 2**: the relationship between number of CNPs and flow rate
  - Precondition
    - 2 to 1 incast
    - Sending CNP to sender 1
  - Flow rate is decreased when more CNP is sent per second
  - Increase CNP number per second cannot help further reduce flow rate when flow rate is decreased to 45Mbps

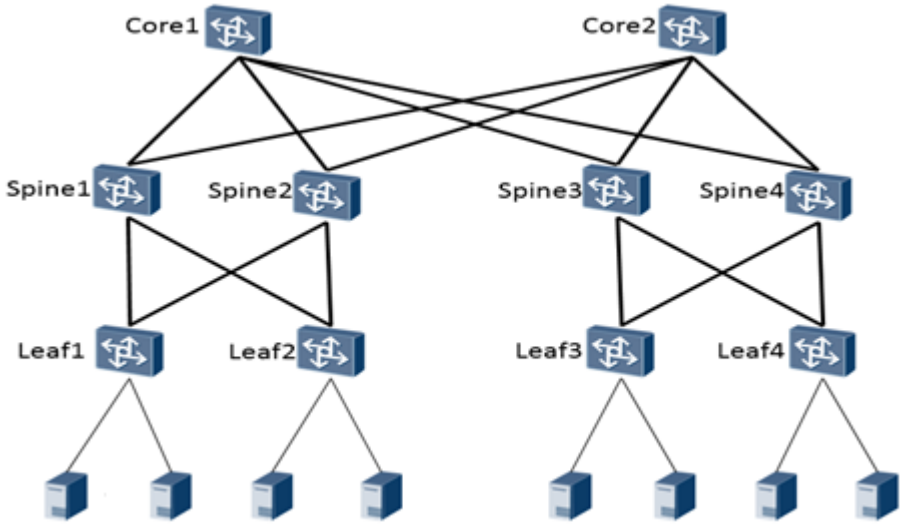






Remote PFC Test

- Networking
  - 10 switches, 8 servers, full bisection, 25G link
  - Switch buffer: 32MB, dynamic threshold
  - Congestion control: DCQCN, ECN kmin=7k， kmax=750k，droprate=10%
  - Remote PFC threshold = 2* kmax
- Test method
  - Generate background traffic
    - TCP and RoCE mixed traffic
    - TCP traffic and RoCE traffic map to 2 different queues
    - 7 sender servers to 1 receiver server (7 to 1 incast)
  - Send one message from a sender to the receiver iteratively
  - Measure average latency under different conditions
    - Message size
    - TCP and RoCE traffic ratio
    - Flow(QP) number



Remote PFC Test

- Test result
  - Remote PFC performs better when increasing message size or increasing flow number
  - Remote PFC has similar performance when message size is small, or flow number is small

Without Remote PFC

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	1024.00	8*7	7.99
7:3	1024.00	8*7	16.99
5:5	1024.00	8*7	6.38
3:7	1024.00	8*7	10.26
1:9	1024.00	8*7	50.17

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	1024.00	256*7	708.31
7:3	1024.00	256*7	919.01
5:5	1024.00	256*7	1301.85
3:7	1024.00	256*7	2182.31
1:9	1024.00	256*7	6616.12

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	4096.00	8*7	10.99
7:3	4096.00	8*7	13.86
5:5	4096.00	8*7	13.56
3:7	4096.00	8*7	18.51
1:9	4096.00	8*7	7067.3

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	4096.00	256*7	759.36
7:3	4096.00	256*7	995.37
5:5	4096.00	256*7	1402.04
3:7	4096.00	256*7	2359.13
1:9	4096.00	256*7	7100.04

With Remote PFC

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	1024.00	8*7	13.51
7:3	1024.00	8*7	9.14
5:5	1024.00	8*7	7.36
3:7	1024.00	8*7	8.9
1:9	1024.00	8*7	43.16

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	1024.00	256*7	160.6
7:3	1024.00	256*7	148.32
5:5	1024.00	256*7	144.64
3:7	1024.00	256*7	62.3
1:9	1024.00	256*7	182.67

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	4096.00	8*7	14.21
7:3	4096.00	8*7	12.46
5:5	4096.00	8*7	14.72
3:7	4096.00	8*7	14.33
1:9	4096.00	8*7	58.74

roce:tcp	Size(B)	Flow number	Avg Latency
9:1	4096.00	256*7	89.1
7:3	4096.00	256*7	117.85
5:5	4096.00	256*7	118.03
3:7	4096.00	256*7	80.98
1:9	4096.00	256*7	664.01

Summary

- In large scale DC network, incast traffic causes a latency issue. Current congestion control, like DCQCN, does not help.
- Remote PFC mitigates the congestion issue, keeping end to end latency low.
- Support standard work of source(remote) PFC.