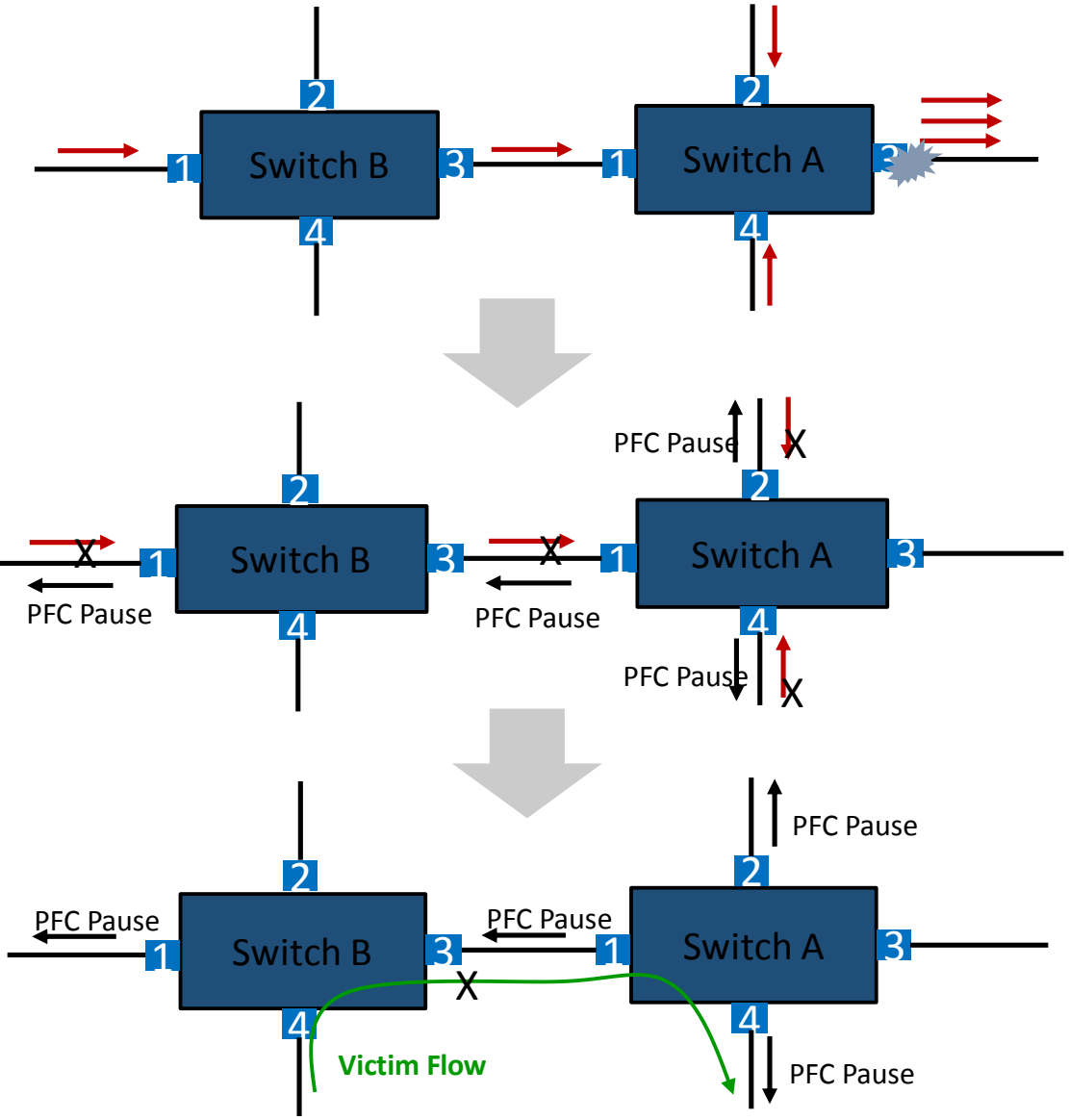
Flow-Based Flow Control(FFC)
Dynamic Virtual Lane(DVL)

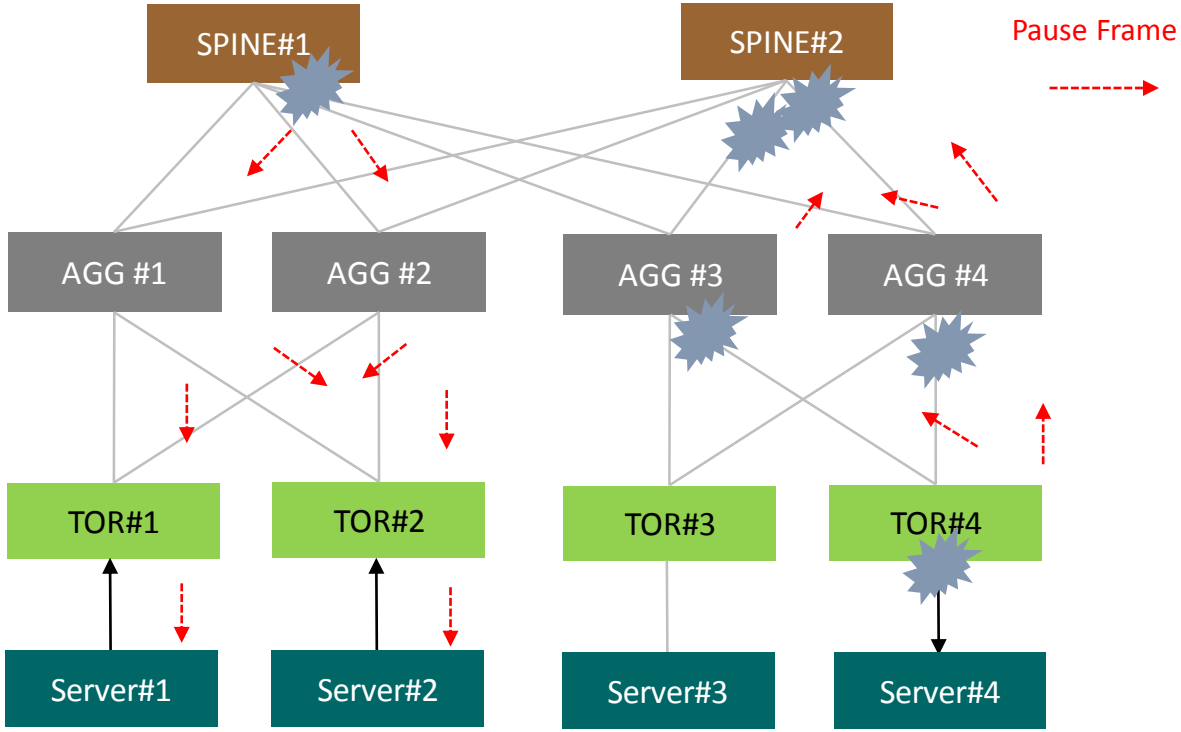
Yolanda Yu

Yolanda.yu@huawei.com

IEEE 802.1 DCB

PFC's drawbacks



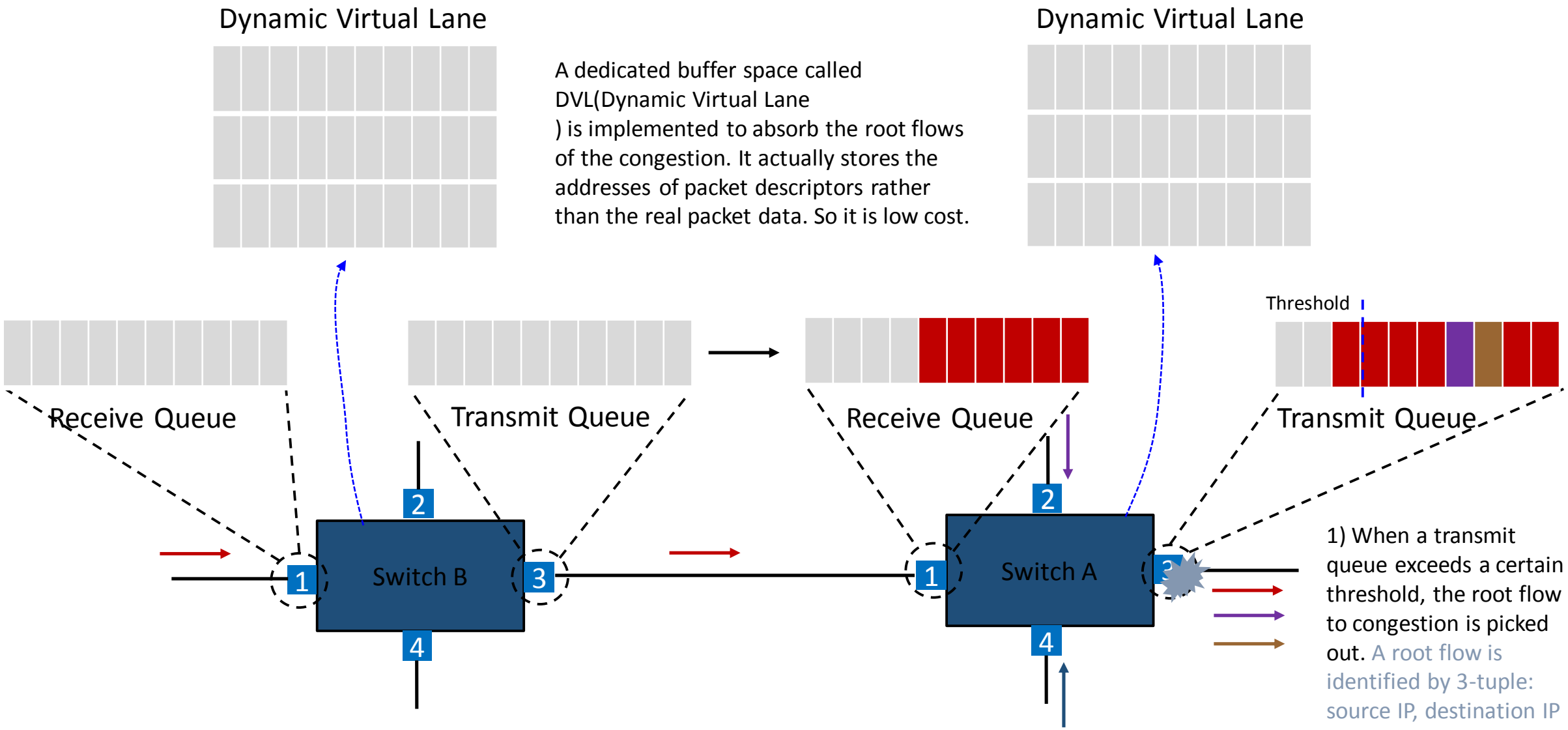


PFC is a coarse-grained mechanism. It will pause the whole traffic of a congested priority.

This will results:

- Head-of-line blocking
- Congestion propagation
- Maybe reduction of the total throughput of the fabric

FFC： Flow-based Flow Control(1/4)



FFC： Flow-based Flow Control(1/4)

4) Get the root flow info in the BPF and assign a DVL to the root flow. The subsequent packets of the root flow will be put into the DVL .When the threshold is exceeded, a BPF(Back Pressure Frame) is sent to the upstream to suspend transmission of that flow.

Dynamic Virtual Lane

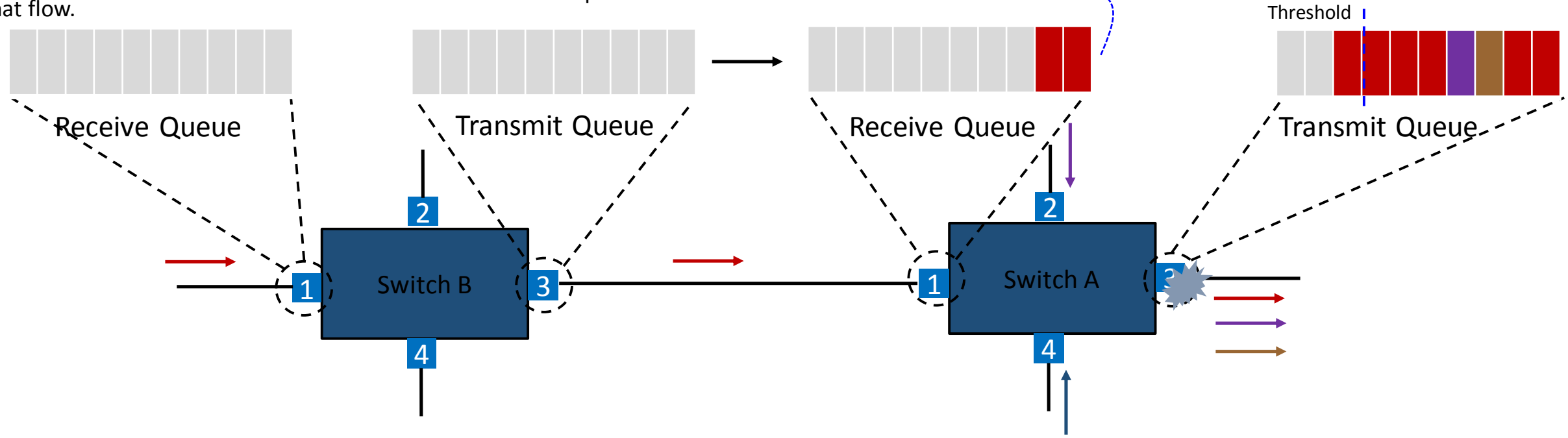
Threshold

3) When the threshold is exceeded, a BPF(Back Pressure Frame) is sent to the upstream to suspend transmission of that flow.

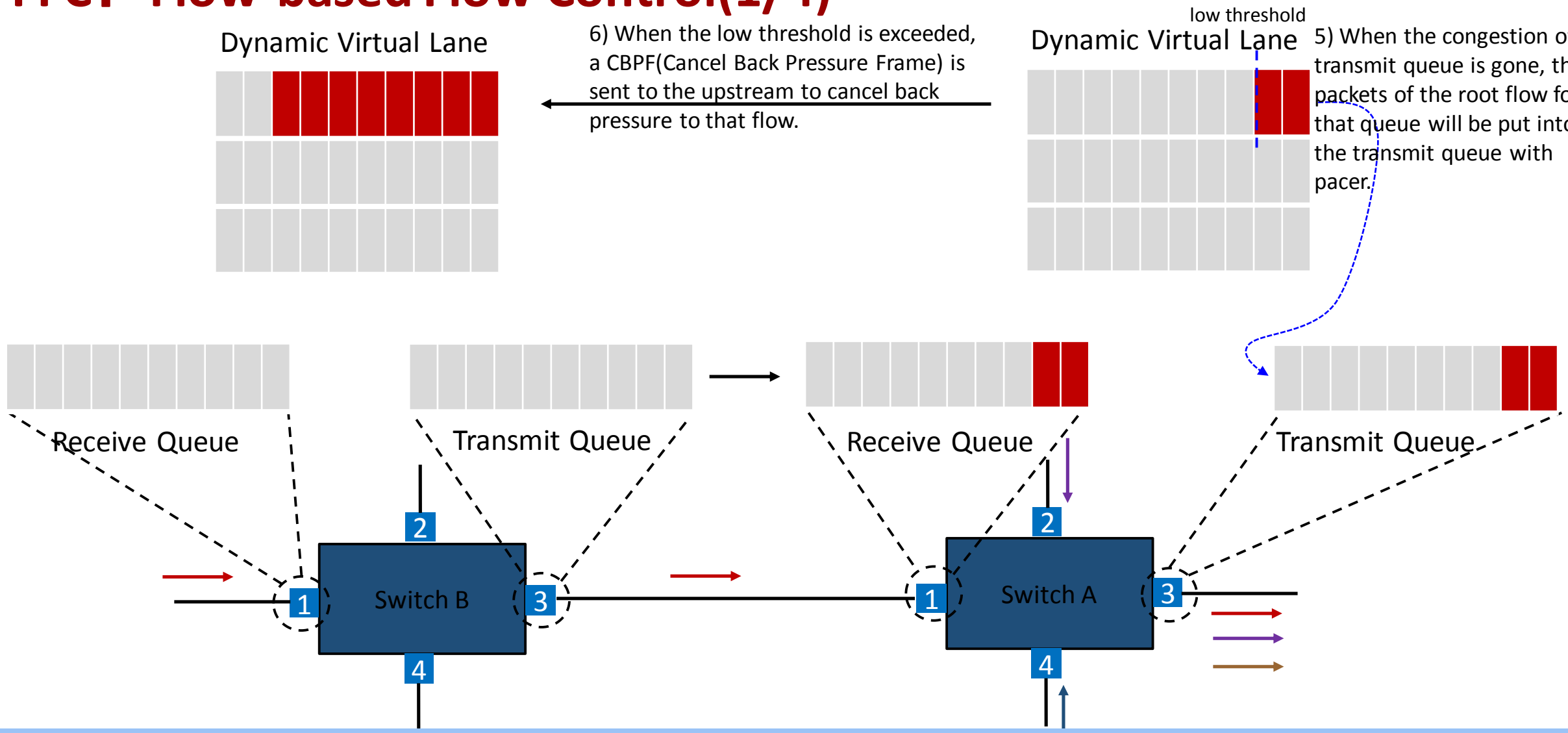
Dynamic Virtual Lane

Threshold

2) One DVL will be assigned to that root flow. The subsequent packets of the root flow will be processed by the chip normally, but it will be put into the DVL, instead of the transmit queue.

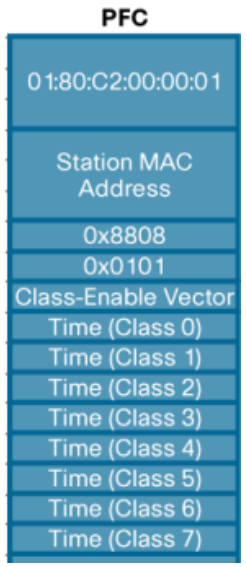


FFC： Flow-based Flow Control(1/4)



- Flow-based, fine-grained, solve head-of-line blocking and congestion propagation problem
- Isolate the congestion out the physic transmit queue, no influence of latency-sensitive small flows and no reduction of the total throughput of the fabric

Flow-based Flow Control(4/4)



Flow-based Flow Control Back Pressure Frame

01:80:C2:00:00:01
Station MAC Address
Ethertype = 0x8808
Control Opcode = 0x0111
Flow Count
Flow Info
State
Flow Info
State
…

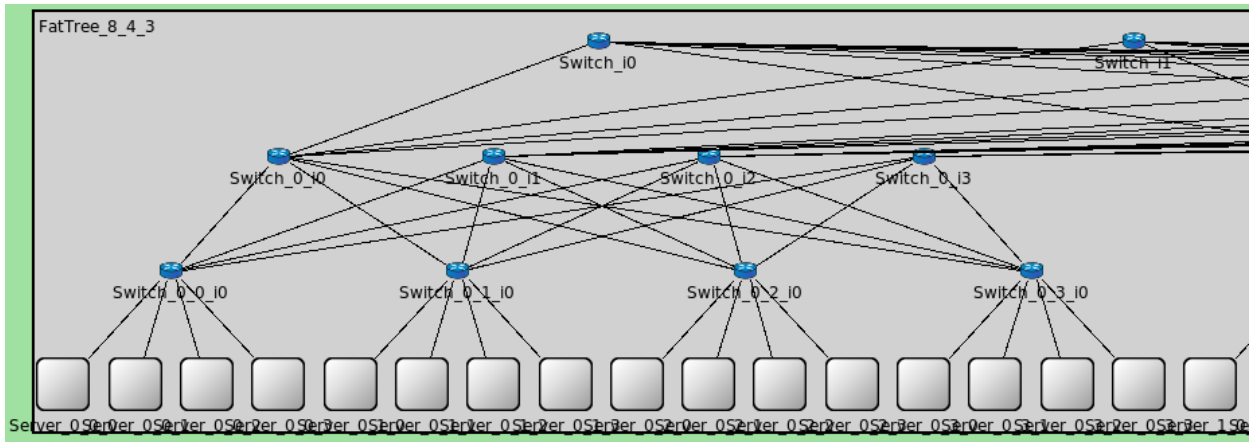
Indicate FFC BPF

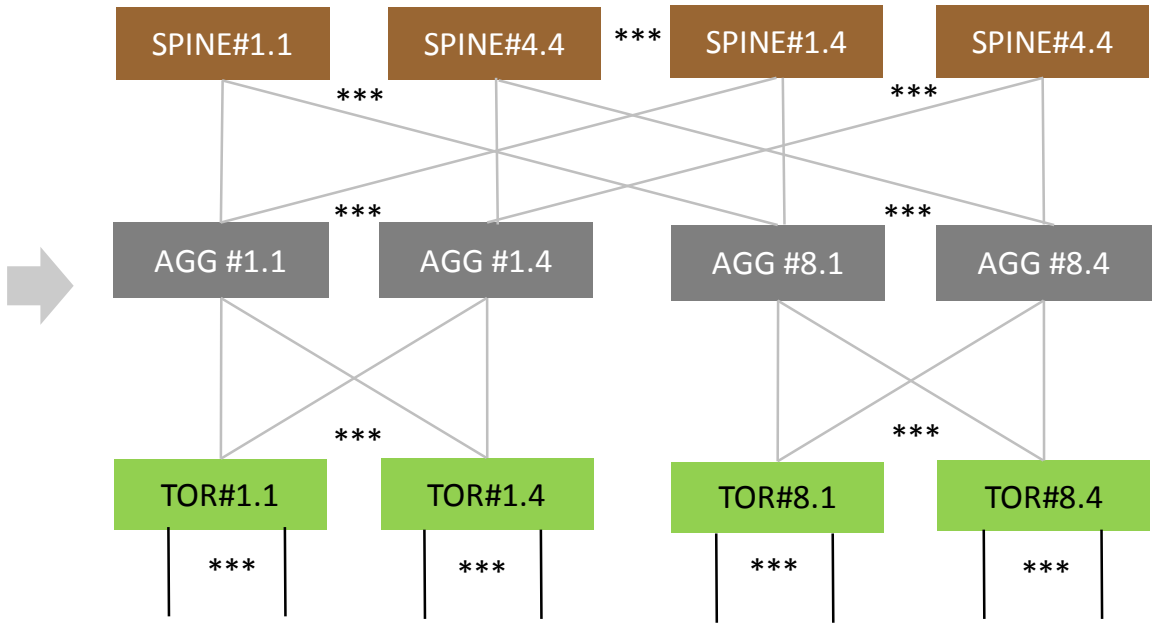
Indicate the flow count in the BPF. If different flows use back pressure at the same time, transfer in one BPF

Indicate the info of the root flow: 3-tuple(Src IP, Des IP, QoS). 5-tuple is better, but 3-tuple is enough.

XON or XOFF, just like traffic lights. Because we can’t calculate the exact pause time for one flow, so we just use the two state.

Simulator Environment

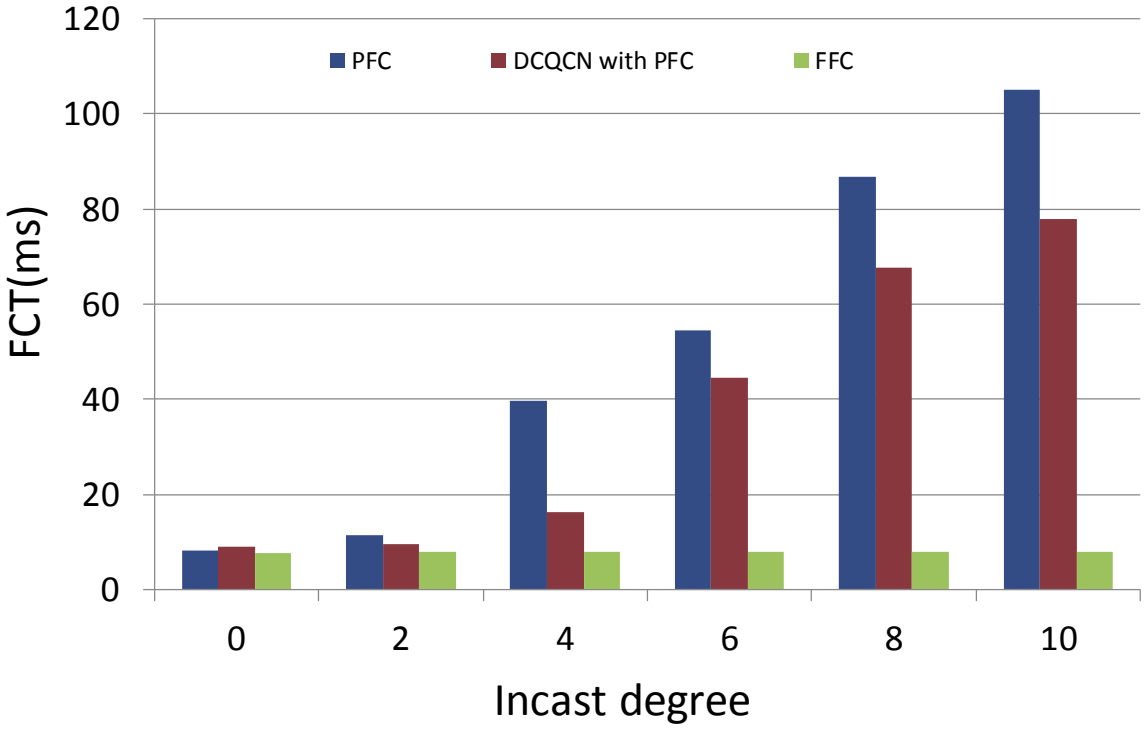




- Platform: OMNET++
- Topology: Fat-tree
- Link capacity： 40G
- Link delay： 200ns (40 meters)

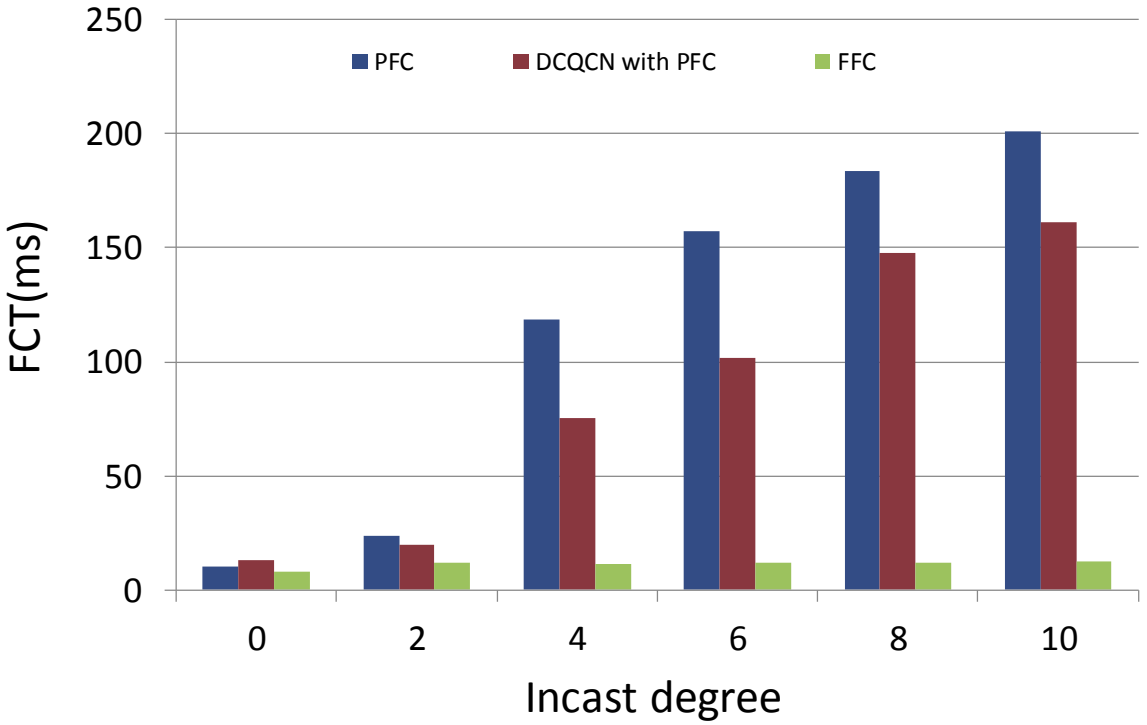
Incast degree and FCT

Median flow completion times of user flows (50% workload)



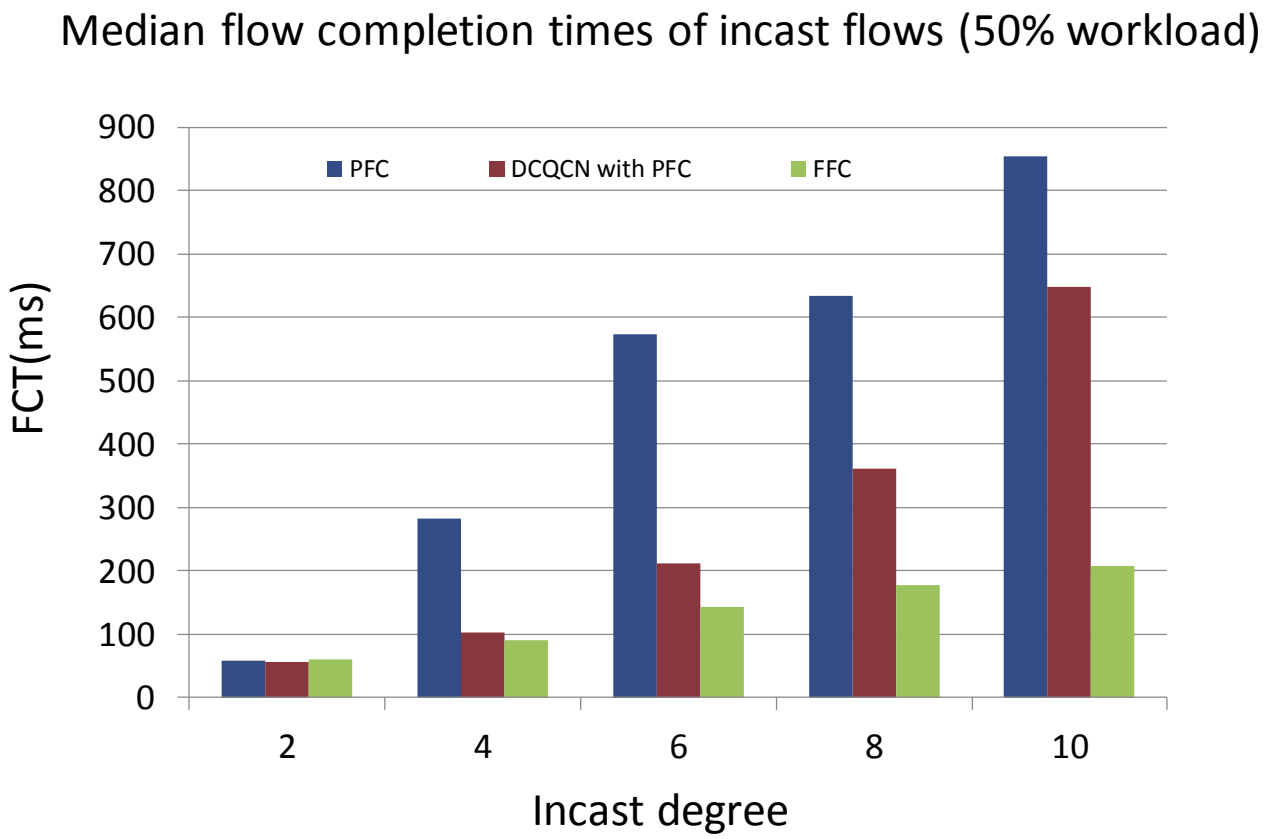
With incast degree increasing, FCT for FFC remains almost unchanged. With an incast degree of 10:1, FFC is nearly 10X better than DCQCN with PFC.

99th percentile flow completion times of user flows (50% workload)



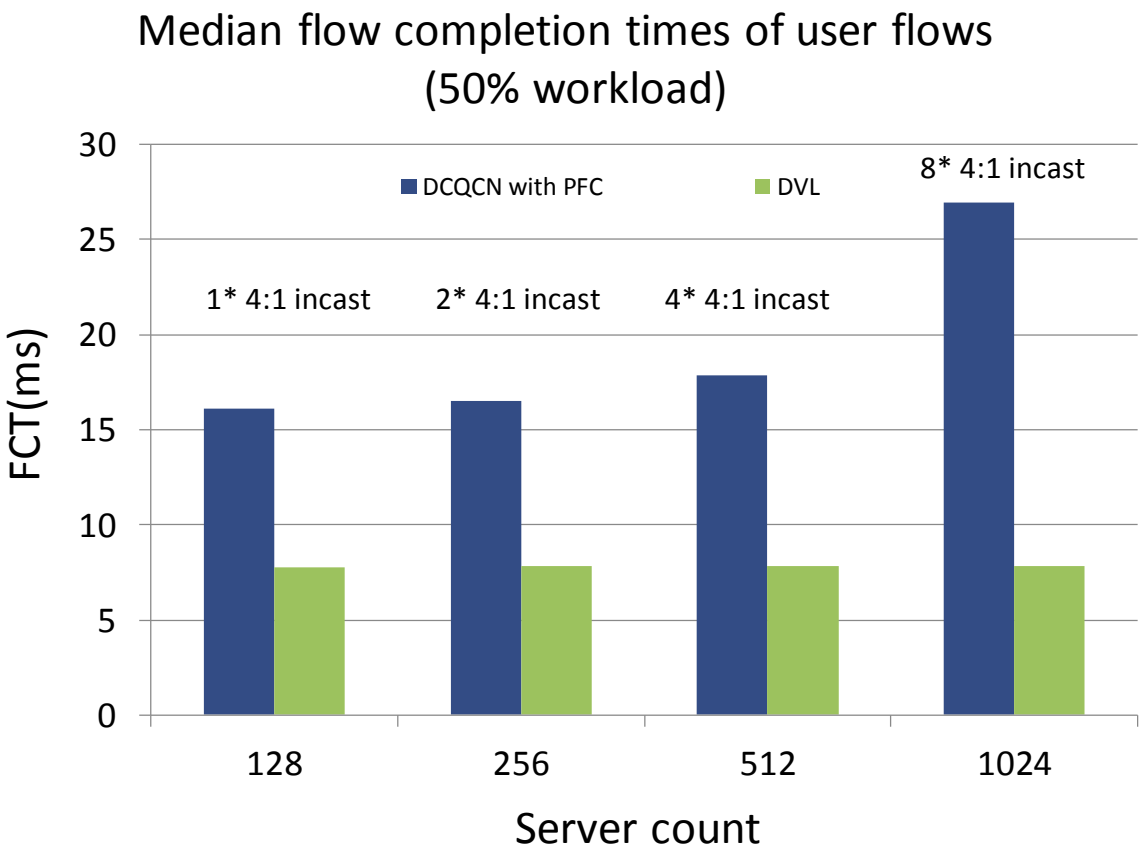
The tail FCT for FFC also remains almost unchanged, and the improvement is more significant than median FCT.

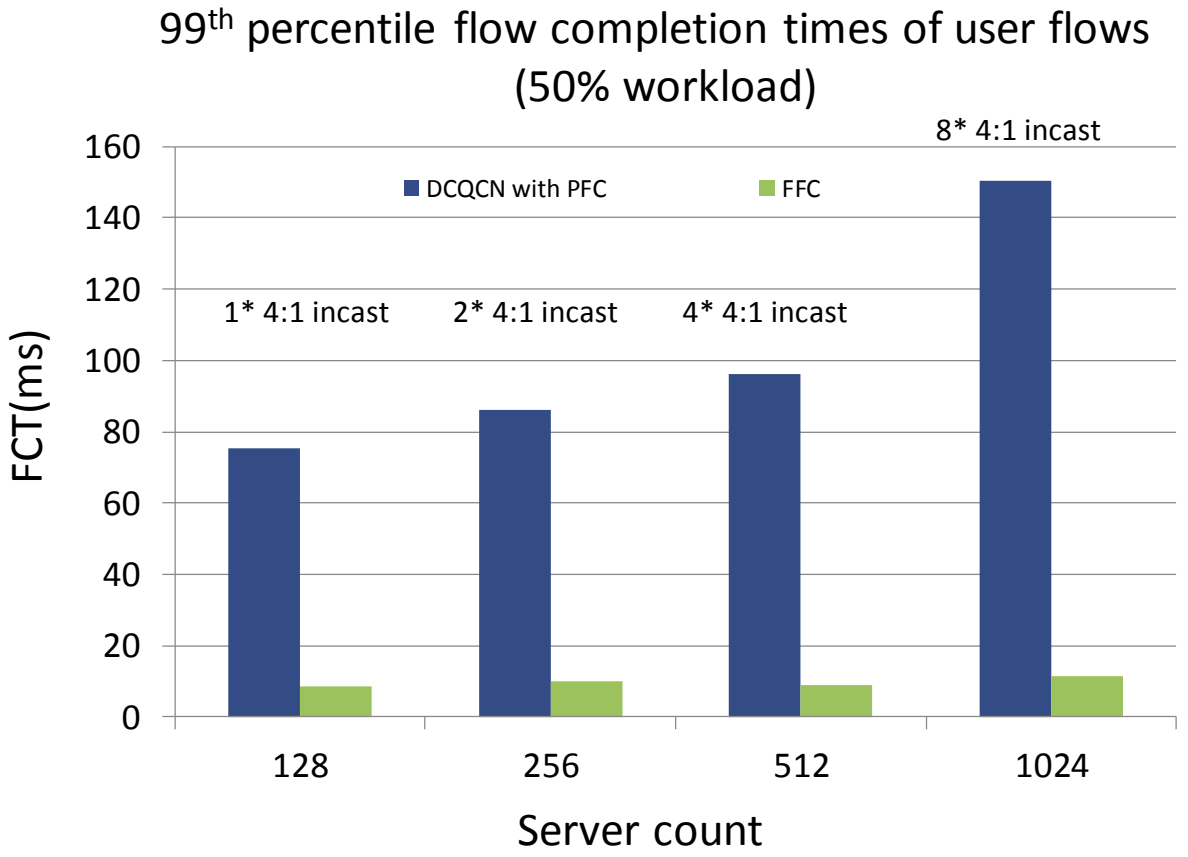
Incast degree and FCT



For incast flows, the FCT for FFC increases linearly, because without PFC no incast flow becomes an innocent flow.

Scale and FCT





When we scale out the network, we increase the count of incast proportionally, the FCT for DCQCN with PFC increases gradually, but FFC remains almost unchanged.

Same phenomenon appears in tail FCT. It implies that DCQCN with PFC has a scale bottleneck. When the FCT arrives at the limitation of application, there will be the upper limit of scale. While the FCT for FFC is not affected by the scale.

Thank you