2021 OFA Virtual Workshop

FLATTEN THE CURVE: SOURCE FLOW CONTROL FOR SUB-RTT MANAGEMENT OF NETWORK TAIL LATENCY

Jeongkeun (JK) Lee, Principal Engineer

w/ Jeremias Blendin, Yanfang Le and Grzegorz Jereczek

Intel Barefoot

intel®

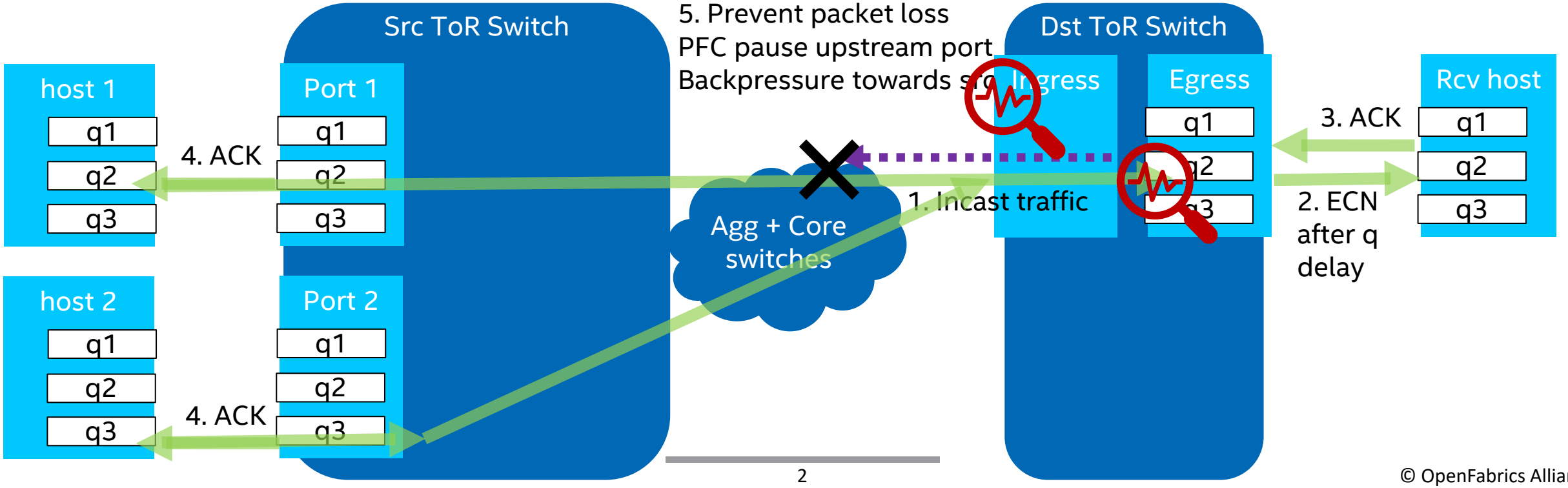
HIGH-PERFORMANCE NETWORKING TODAY

High performance Networking (HPN)

- Remote direct memory access traffic (RDMA)
- Protocols are implemented in NIC hardware
- Dedicated networks with low latency ~20us RTT
- Main performance metric: flow completion time

State of the art: RoCEv2

- Lossless Ethernet: PFC to prevent packet loss
- ECN-CNP congestion feedback from rcv to src
- Drawbacks: PFC storm, deadlock
- HoL blocking slows down the network fabric



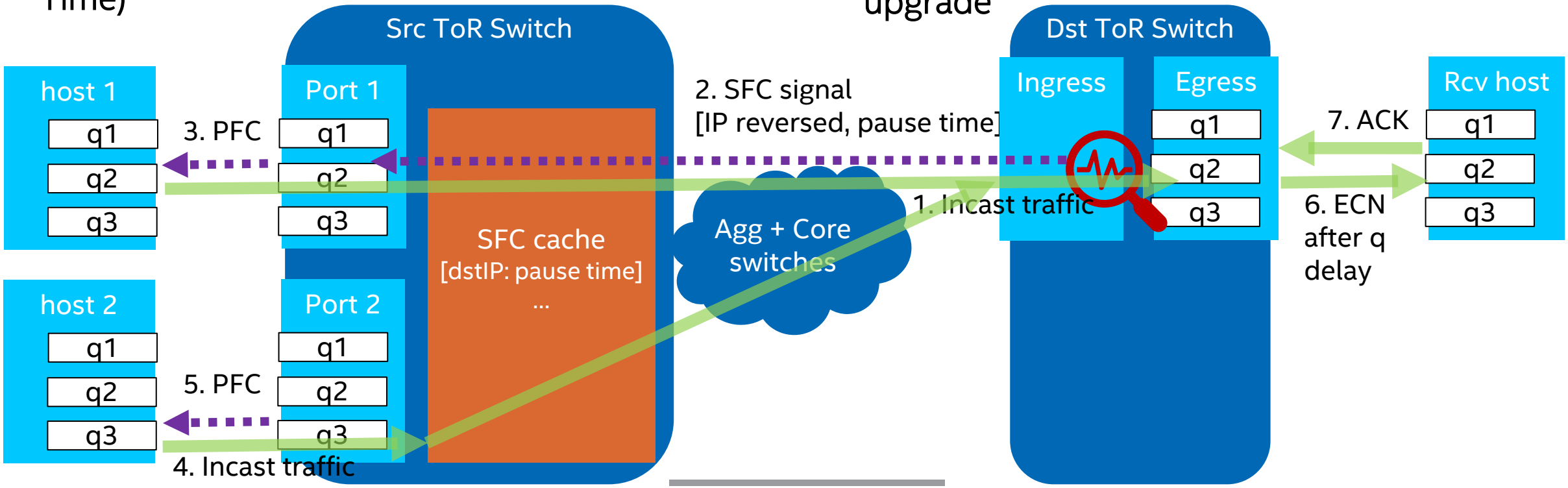
SFC (SOURCE FLOW CONTROL) IN 1-SLIDE

What is SFC?

- Edge-to-Edge signaling of congestion
- Flow control that instantly 'flattens the curve'
- Signaling + flow ctrl all in sub-RTT
- 2~10x reduction of tail FCT (Flow Completion Time)

SFC is not

- ~~lossless network~~ vs minimal switch buffering
- ~~e2e congestion ctrl~~ vs NIC flow ctrl
- ~~pausing switches~~ → minimal PFC side effects
- ~~need greenfield deployment~~ → ToR-only upgrade



FAQS

Why not E2E congestion control?

- Faster link speed ➔ shorter RTTs to finish a message ➔ need sub-RTT reaction
- E2E CC relies on forward signal, packets carrying the signals delayed by the congestion
- Cannot react to incast, sudden congestion
  - Swift (Google CC) reports large tail latency (up to 20x RTT @ 99.9th) due to incast or higher QoS traffic

Why not just 'backward' CNP from switches?

- CNP cuts rate by half ➔ take multiple RTTs to flatten down the curve of incast buildup
- CNP reaction by sender NIC on TX wire can be slow, up to 20us
- Note) PFC reaction time: max **614.4ns by IEEE 802.1Qbb**

What if (rare) congestion queue drops?

- Simple switch solution: prevent RTOs by using higher drop threshold on RDMA 'last' pkts

FAQS

New parameters? Yes, but simple config

- E.g., SFC trigger threshold = SFC drain target = ECN threshold

Edge link (NIC-ToR) HoL blocking?

- Yes, but can be minimized by using multiple HW queues for one Traffic Class
- Possible by SW (NIC driver) change

Can it handle Rx NIC congestion?

- Yes, by considering NIC-to-ToR PFC (Xon/Xoff) state in SFC trigger condition

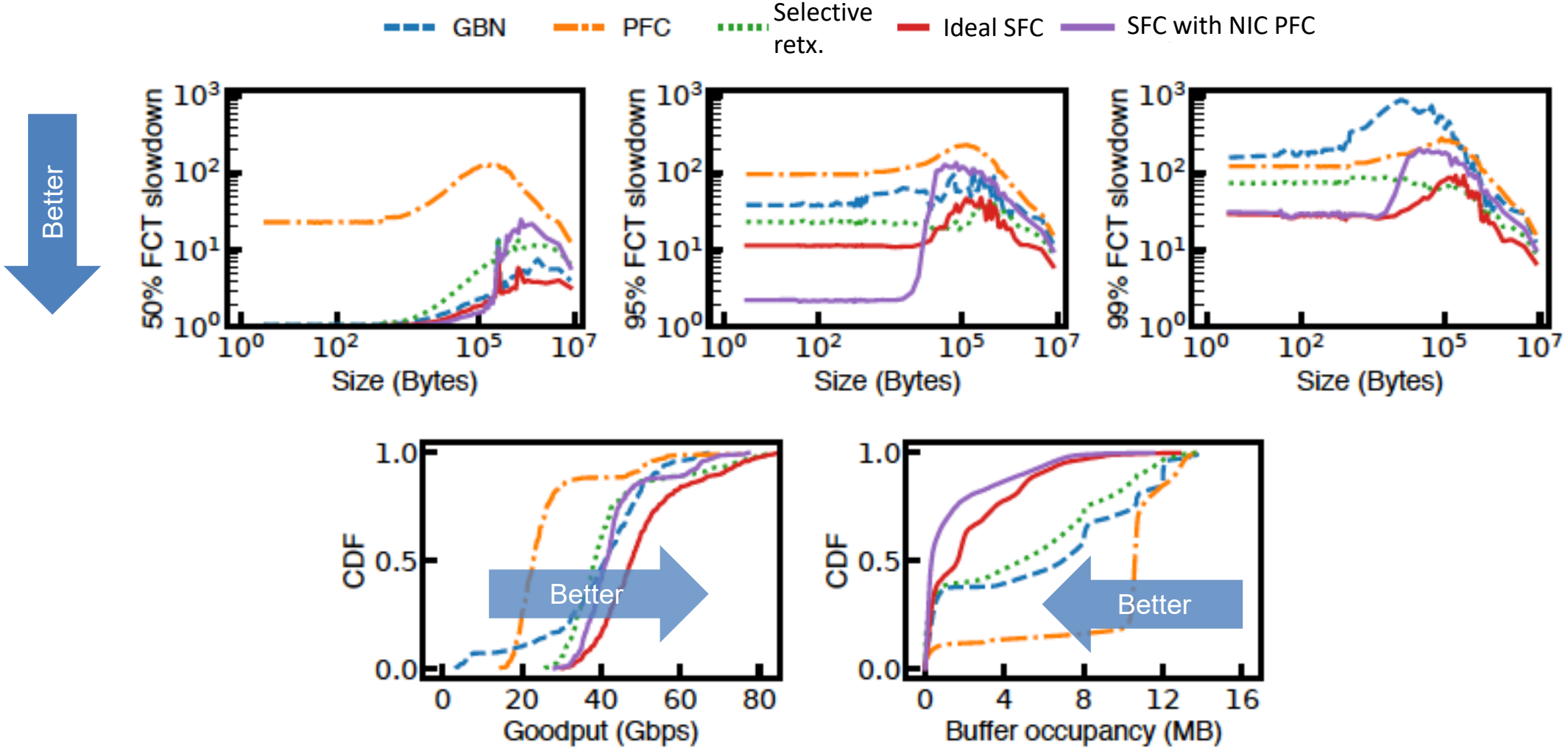
Is SFC only for incast?

- No, it reacts to queueing due to any case of “arrival rate >> departure rate”
  - Incast: arrival rate ↑
  - Higher QoS traffic: departure rate ↓

EVALUATION HIGHLIGHTS

- Eval setup: 14-node system, and 320-node simulation
- Work with real applications? Yes, SFC performed the best in VGG16 training
- Avoid HoL blocking? Yes, yielding small latency, high goodput
- Compare to selective retx (IRN) @ NICs
  - SFC >= IRN, as SFC avoids drops
- ToR-only deployment performs close to SFC @ every switch
- Robust over longer RTTs? Yes, thanks to SFC caching at src ToRs

RPC WORKLOAD, 50% BACKGROUND + 8% INCAST



2021 OFA Virtual Workshop

THANK YOU

JK Lee

Intel Barefoot