**IEEE 802.1** May 2022 Interim Session

# Source Flow Control Simulation Results Fairness and Performance

Jeremias Blendin
Contributors: Jeongkeun "JK" Lee, Yanfang Le, Pedro Yebenes Segura, Paul Congdon

# Agenda

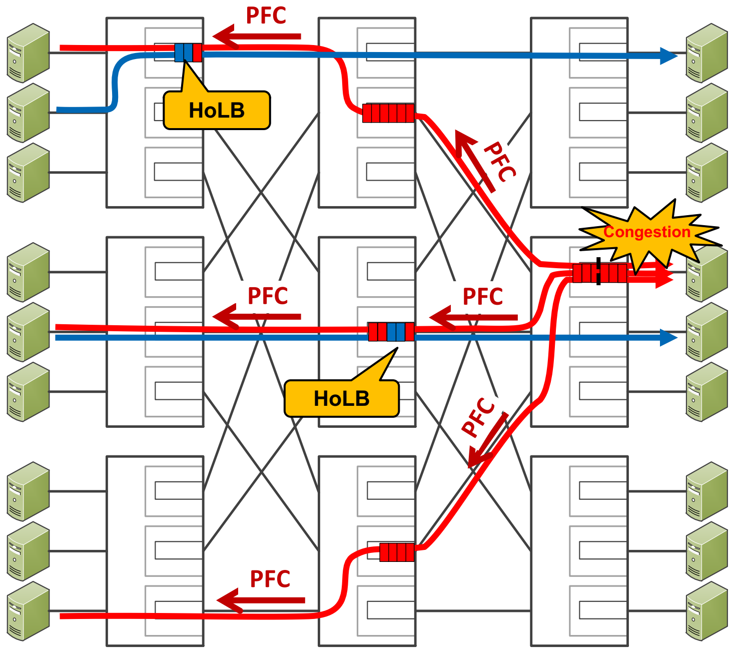
- SFC Introduction
- Simulation Overview
- SFC Benefits
- SFC Fairness

# Source Flow Control

- Background: Future 802.1 Congestion Management Tools
  - 802.1Qbb - Priority-based Flow Control (PFC)
    - Hop-by-hop flow control
  - P802.1Qcz - Congestion Isolation (CI)
    - Improve PFC by isolating congested queues and reduce hop-by-hop head-of-line blocking
- This talk: Source Flow Control (SFC)
  - Signal from switch directly to traffic source
  - Remove head-of-line blocking from network
  - SFC w/ Proxy design to accelerate deployment
  - Does not require complex buffer tuning

# Source Flow Control (SFC) High Level Concept

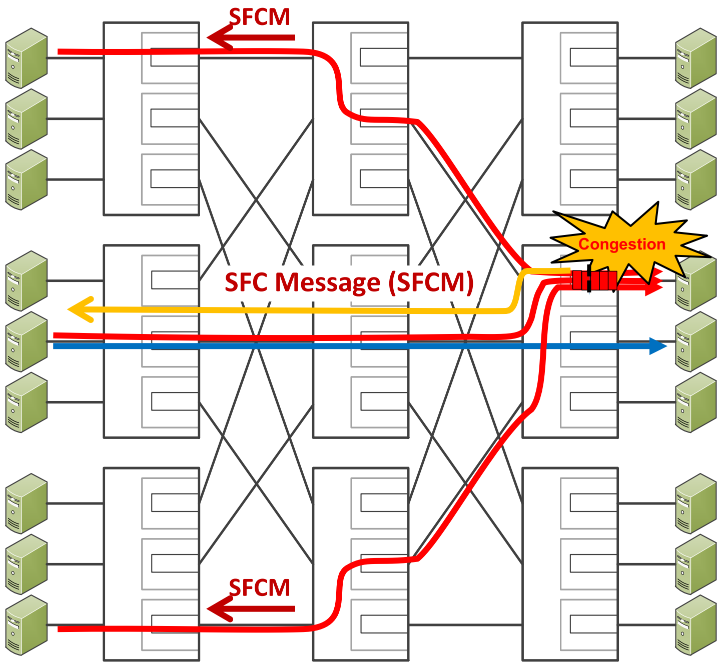
**Today: 802.1Qbb - Priority-based Flow Control (PFC)**



Congested Flow

Victim Flow

**Proposed: Source Flow Control**



- Operational concerns
  - Head-of-Line blocking
  - Congestion spreading
  - Buffer Bloat, increasing latency
  - Increased jitter reducing throughput
  - Deadlocks with some implementations

- Benefits over PFC
  - Use local switch telemetry to trigger back-to-sender signaling
  - Edge-to-edge FC signaling using **L3** message
  - Removes head-of-line blocking completely from the network
  - SFC signaling directly to transport protocol end-point (per-flow)
  - Works with many types of transport protocols (RoCEv2, TCP, UDP)

Figure source: https://mentor.ieee.org/802.1/dcn/21/1-21-0068-01-ICne.pdf

# Source Flow Control (SFC) w/ ToR Proxy

- SFC with ToR Proxy
  - SFC proxy converts SFC message to PFC frame at sender ToR
  - Works with today's RDMA NICs
  - Removes congestion from network switches
    - Only small chance for head-of-line blocking at sender NIC

Source Flow Control (w/ ToR Proxy)

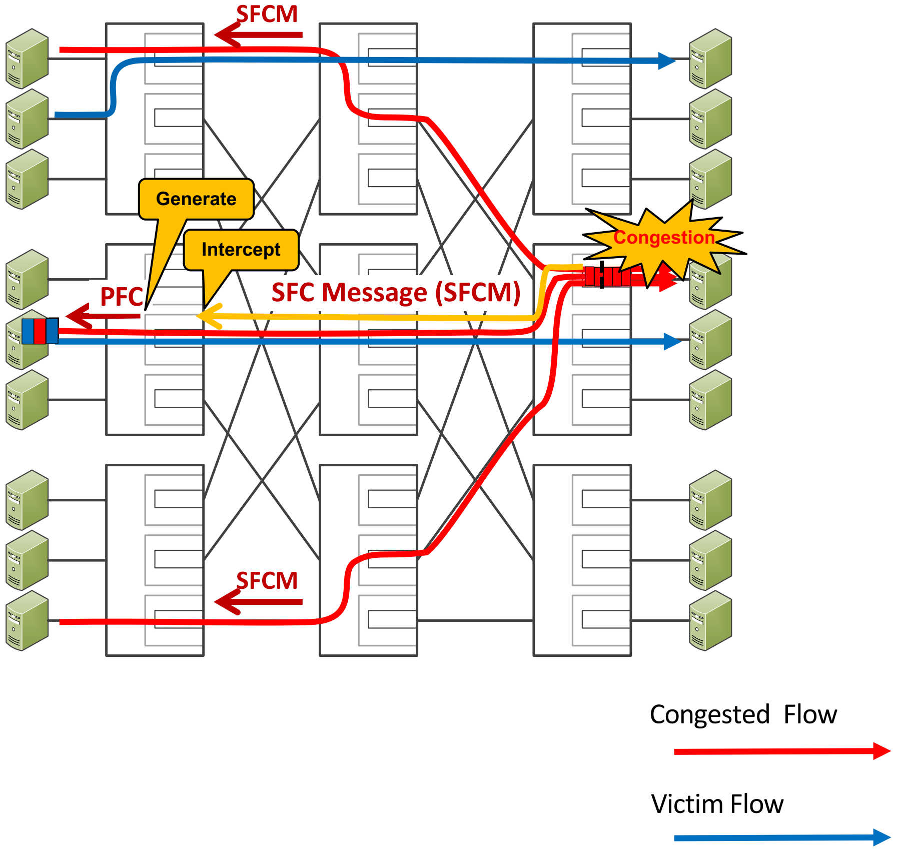


Figure source: https://mentor.ieee.org/802.1/dcn/21/1-21-0068-01-ICne.pdf

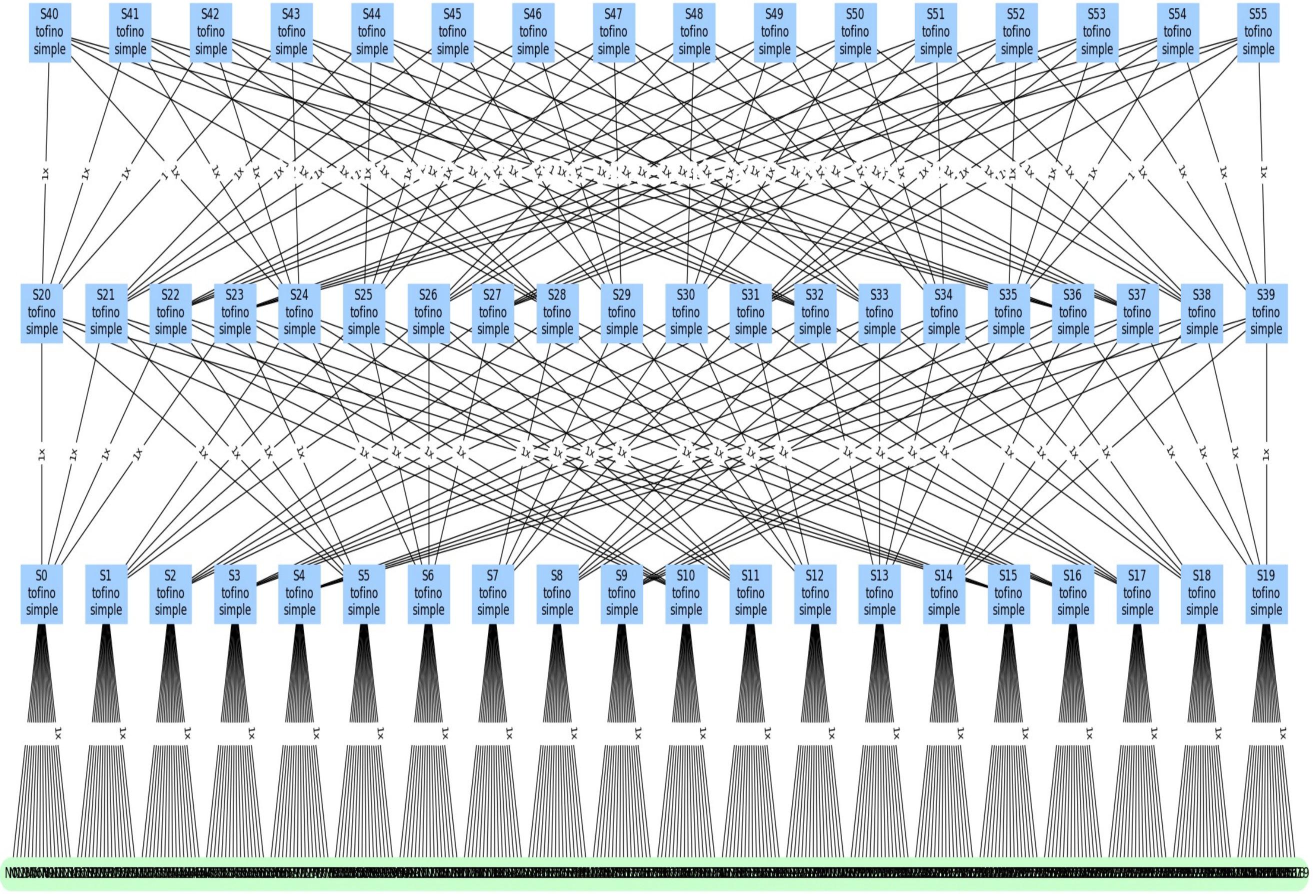
# Simulation Overview

# Simulation: Goals

- Show benefits of SFC
  - Increase application performance
  - Reduce of in-network buffering
  - Does not affect fairness
- Metrics
  - App perf: Flow Completion Time (FCT)
  - In-network buffering: Switch buffer occupancy
  - Fairness: Per-flow link capacity share

# Simulation Setup

- Simulation Software
  - Fabsim-X (Intel proprietary)
    - Musleh, Malek, et al. "Fabsim-X: A simulation framework for the analysis of large-scale topologies and congestion control protocols in data center networks." IEEE *MASCOTS* 2020.
    - Note: we observed comparable results on NS3
- Network topology
  - 3-tier fat-tree (100/400GbE)
  - 320 nodes, 56 switches
  - Full bisection bandwidth
- PFC, DCQCN, SFC parameters in backup slides
  - Note: PFC + DCQCN is sensitive to tuning (workload-specific)
  - PFC/DCQCN parameters are selected to ensure losslessness



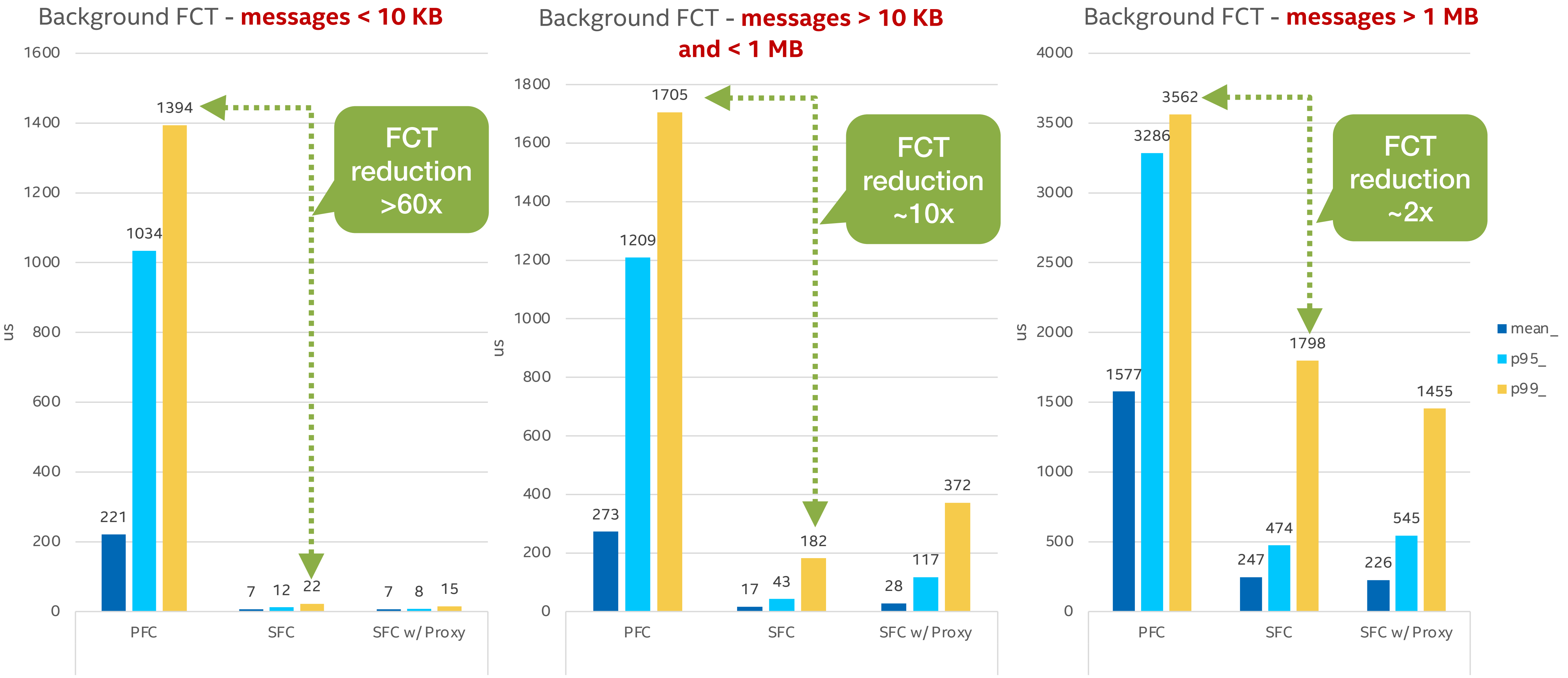
# Traffic Load Configuration

- Network protocol
  - RDMA with DCQCN
    - State-of-the-art flow control in modern RDMA NICs
    - Use variant with “initial window” mechanism
      - Source: Li, Yuliang, et al. "HPCC: High precision congestion control." *ACM SIGCOMM* 2019.
- Background traffic
  - 320 hosts
  - Traffic based on Google RPC workload
    - Source: Montazeri, Behnam, et al. "Homa: A receiver-driven low-latency transport protocol using network priorities." *ACM SIGCOMM* 2018.
  - Load factor 50%

- In-cast traffic
  - 120:1 incast
  - Message size 256 KB
  - The incast traffic load is 8% of the network capacity
    - Similar approach to: Li, Yuliang, et al. "HPCC: High precision congestion control." *ACM SIGCOMM* 2019.

# SFC Benefits

# Results: Background Traffic Performance



Background FCT - **messages < 10 KB**

FCT reduction >60x

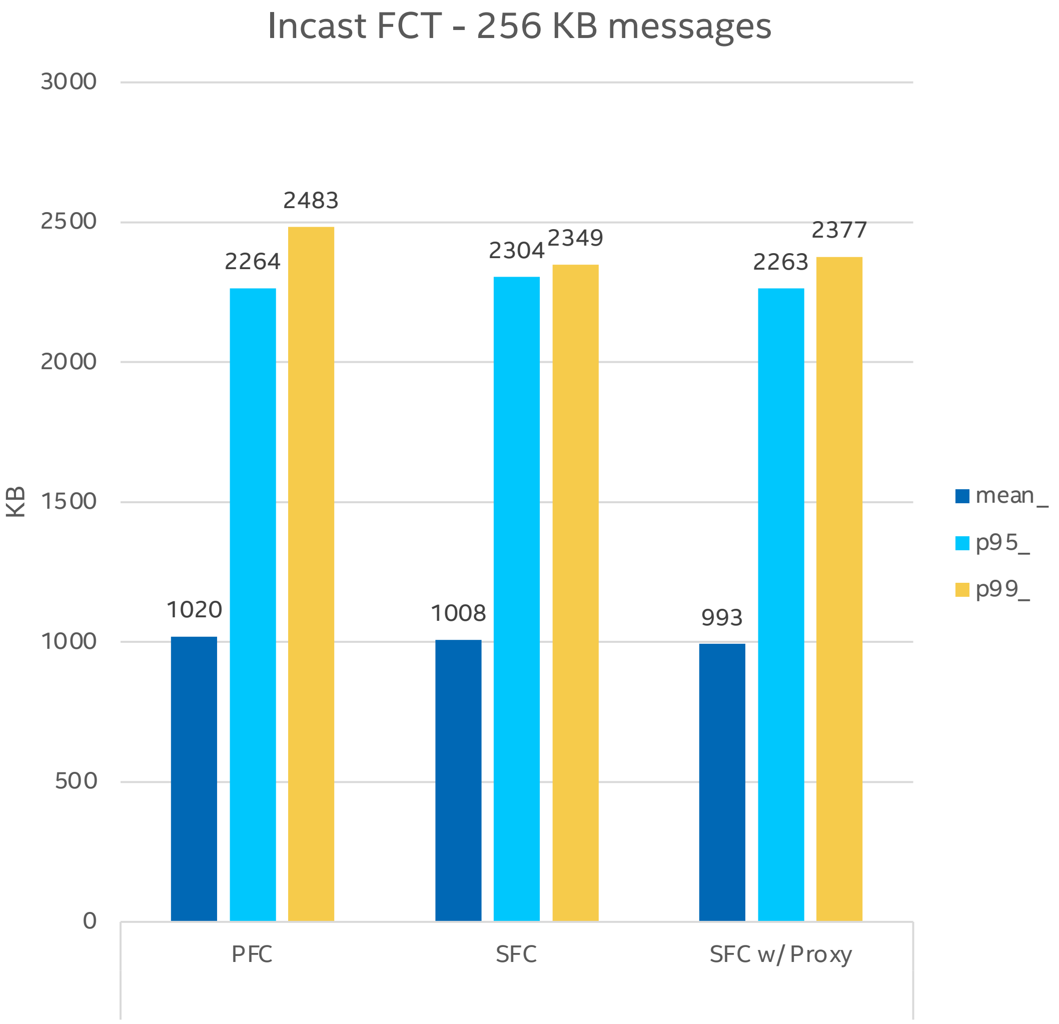
Background FCT - **messages > 10 KB and < 1 MB**

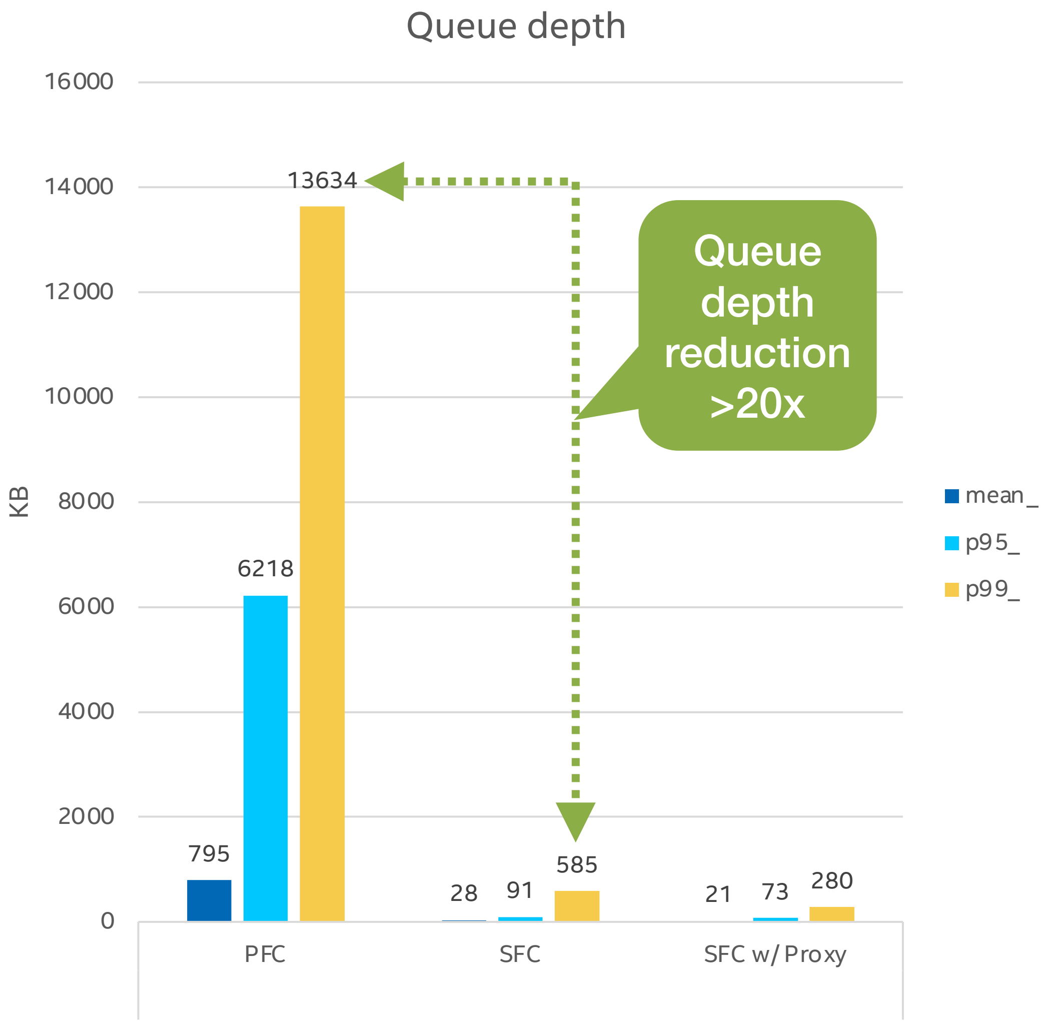
FCT reduction ~10x

Background FCT - **messages > 1 MB**

FCT reduction ~2x

# Results: Incast Performance and Queue Depth



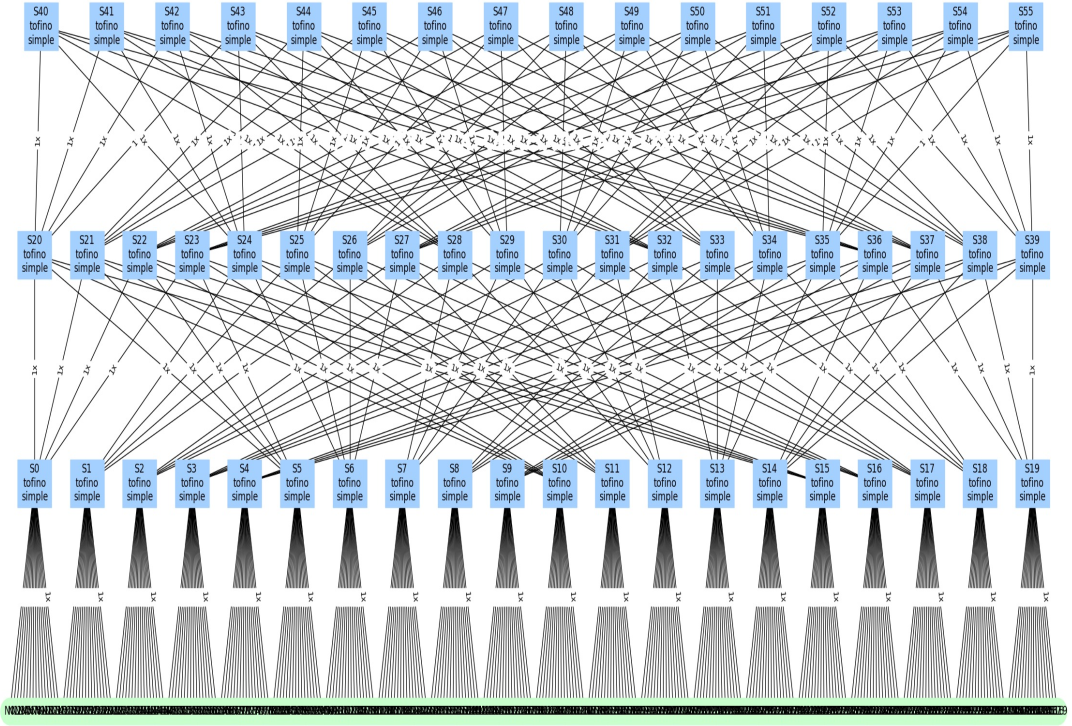


Queue depth reduction >20x

# Impact on Fairness

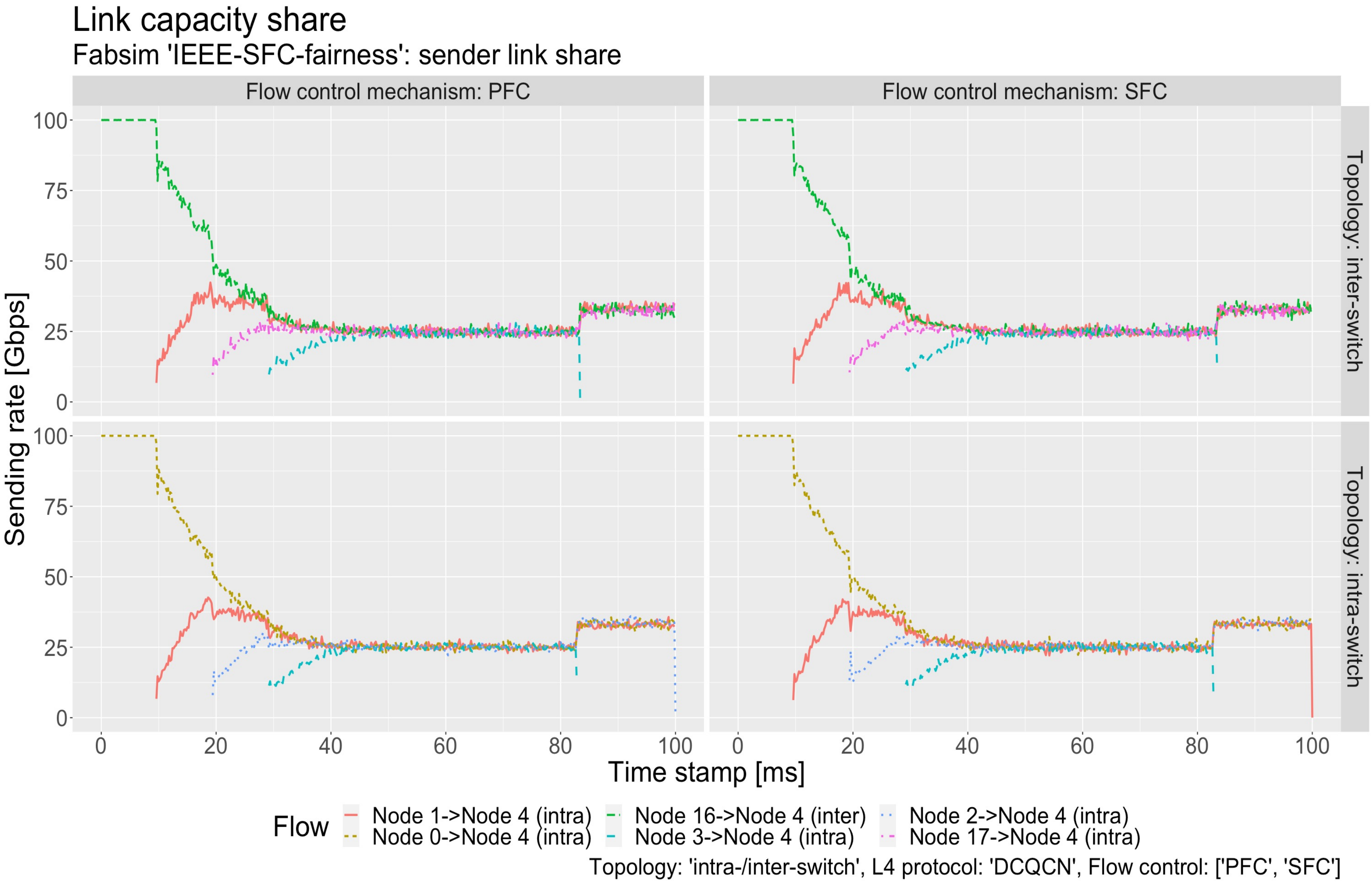
# Impact on Fairness: Overview

- Goal
  - Show that SFC's effect on fairness
- Approach
  - Use PFC as baseline
  - Staggered start/stop to observe flow-addition and flow-removal behavior on congested link
    - As suggested in: Perry, Jonathan, et al. "Fastpass: A centralized" zero-queue" datacenter network." ACM SIGCOMM 2014.
  - Use long-running, similar-sized flows
- When flow sizes are mixed, fairness index is not clearly defined
  - FCT partially reflects fairness, as worsend fairness will push up tail latency
- Scenarios
  - 4:1 incast, 100GbE
  - Intra-switch: 4 senders, 1 receivers connected to same switch
  - Inter-switch: Intra scenario with 2 senders connected to remote ToR switch



# SFC Fairness Comparison with PFC

- SFC does not hurt fairness provided by DCQCN



# Conclusion

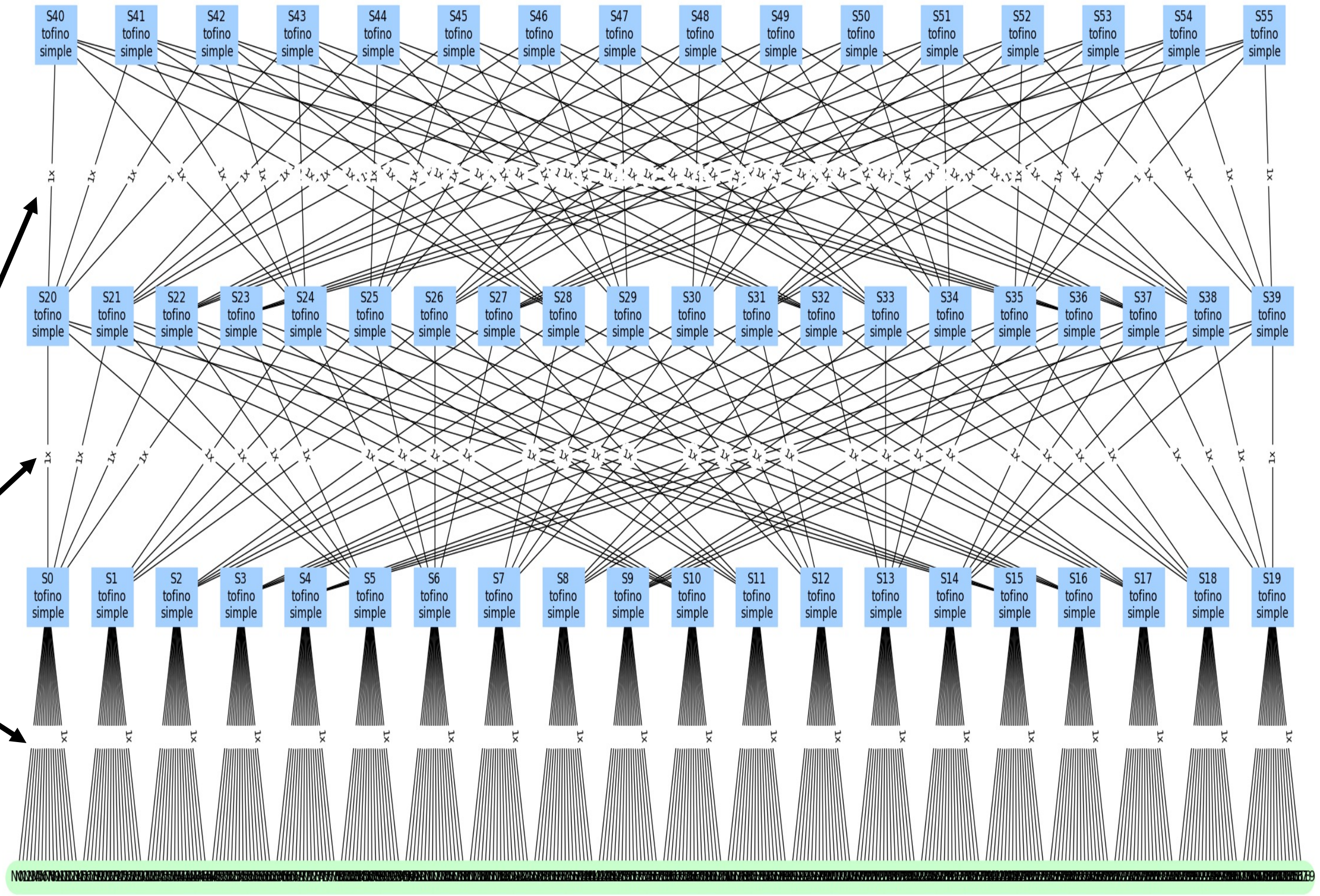
- Benefits compared to PFC
  - SFC can significantly reduce FCT for background traffic
  - SFC can significantly reduce queueing in the network
- No negative side-effects
  - SFC does not negatively affect the incast FCT
  - SFC does not negatively affect fairness
- Future work
  - Bigger scale and higher RTT simulations
  - Systems results

intel®

# Benefits Simulation Settings

# System Configuration in FabSim-X

- Network
  - 3-tier fat-tree
    - 320 nodes to 20 ToR sw (1 link)
    - 20 ToR sw to 20 agg sw (1 links)
    - 20 agg sw to 16 core sw (1 links)
  - Switch radix: 20
  - Link delay: 0.5 us
  - Link speed SW to SW: 400 Gbps
  - Link speed NIC to SW: 100 Gbps
  - MTU: 1024 B
  - RTT  12 us



# Switch Configuration Parameters

- SFC: Shared buffer
  - Ingress pool size: disable accounting (200 MB)
  - Egress pool size: 16 MB
  - Ingress guaranteed per port: 50 KB
  - Egress guaranteed per port: 50 KB
  - Sharing mechanism: dynamic thresholding
    - Ingress coefficient: 1
    - Egress coefficient: 1
  - No ingress drops (set it to a very high value), but we can have egress drops

- SFC configuration
  - Trigger threshold: 100 KB (2/3 BDP)
  - Target threshold: 50 KB (1/3 BDP)
  - Suppression period: 6 us (1/2 RTT)
  - Destination cache: ToR
- PFC: Static Ingress Buffer
  - Total buffer 16 MB / 20 ports = ~800 KB
  - PFC threshold 650 KB
  - No egress drops

# Congestion control configurations

- DCQCN
  - Fast recovery steps: 1
  - Gain: 0.0009813
  - Byte counter: 2097152
  - Timer: 4 us
  - Alpha timer: 5 us
  - AI: 0.0125
  - Hyper AI: 0.025
  - CNP period: 4 us
  - Window: 15us
  - ECN threshold: 50 KB (1/3 BDP)

# Fairness Simulation Settings

# Switch Configuration Parameters

- Shared buffer: Both
  - Egress pool size: 16 MB
  - Ingress guaranteed per port: 50 KB
  - Egress guaranteed per port: 50 KB
  - Sharing mechanism: dynamic thresholding
    - Ingress coefficient: 0.5
    - Egress coefficient: 2.0
- Shared buffer: SFC
  - Ingress pool size: disable (200 MB)
- Shared buffer: PFC
  - Ingress pool size: 9 MB

- SFC configuration
  - Trigger threshold: 100 KB
  - Target threshold: 64 KB
  - Suppression period: 6 us
  - Destination cache: Only in ToR
- Other settings are the same as the Benefits Simulation Setup