IEEE 802.1 July 2022 Plenary Session

Source Flow Control Simulation Results

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

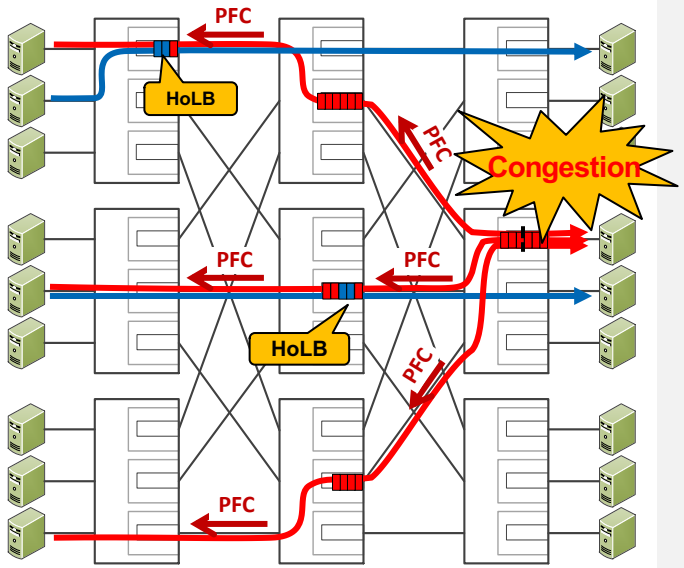
Agenda

- SFC Introduction
- Simulation Overview
- Comparing SFC with PFC
- Comparing Workloads
- SFC Parameter Sensitivity

SFC High Level Concept

- Source Flow Control
  - Signal from switch directly to traffic source: per-flow pausing
  - Removes head-of-line blocking from network
  - Simplify deployments compared to PFC
    - Does not require complex buffer tuning
    - Completely remove risk of deadlocks
- How is SFC triggered in our implementation?
  - Trigger on egress queue depth threshold
  - Sender pause time = expected drain time to target queue depth
  - Not necessarily part of the standard (see earlier discussion https://www.ieee802.org/1/files/public/docs2022/new-congdon-SFC-design-topics-0622-v01.pdf)

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



Proposed: Source Flow Control (SFC)

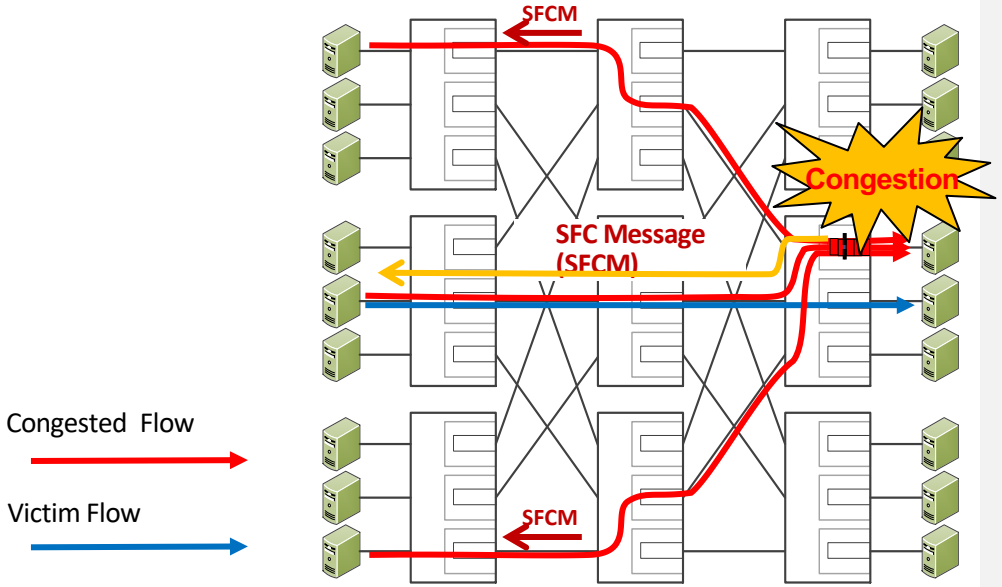


Figure source: https://mentor.ieee.org/802.1/dcn/21/1-21-0068-01-ICne.pdf

SFC w/ ToR Proxy

- SFC with ToR Proxy
  - Works with today's RDMA NICs
  - SFC proxy converts SFC message to PFC frame at sender ToR
  - Removes congestion from network
    - HolB possible at sender NICs but not in switches

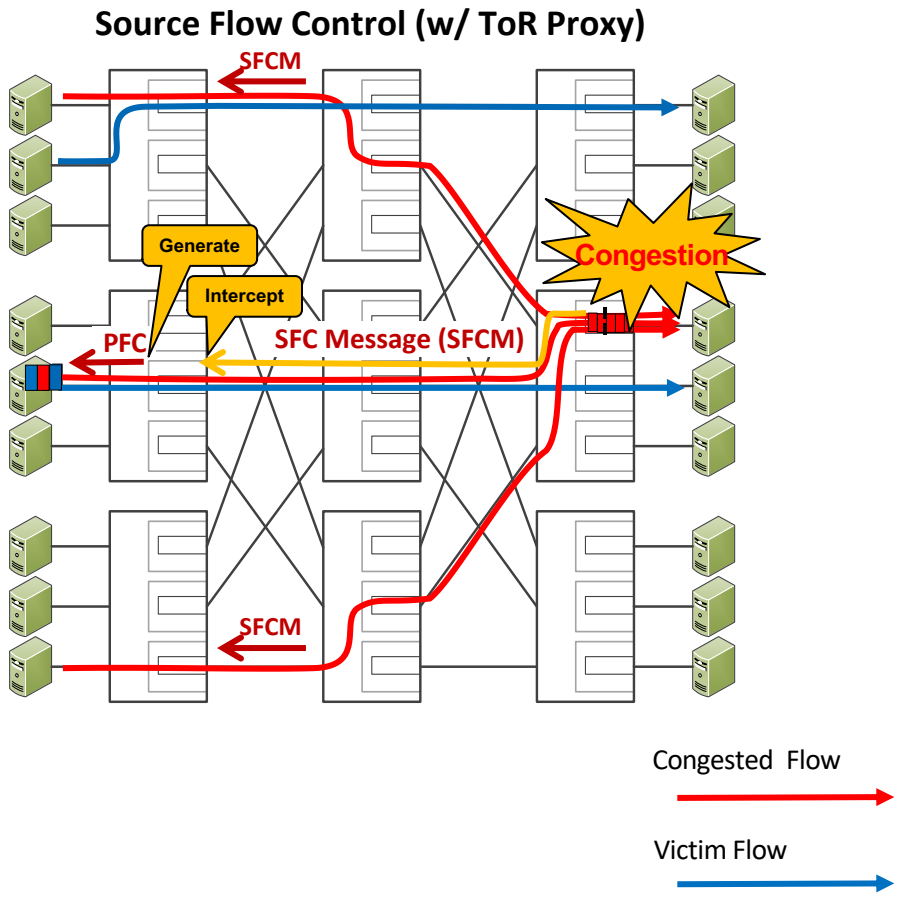


Figure source: https://mentor.ieee.org/802.1/dcn/21/1-21-0068-01-ICne.pdf

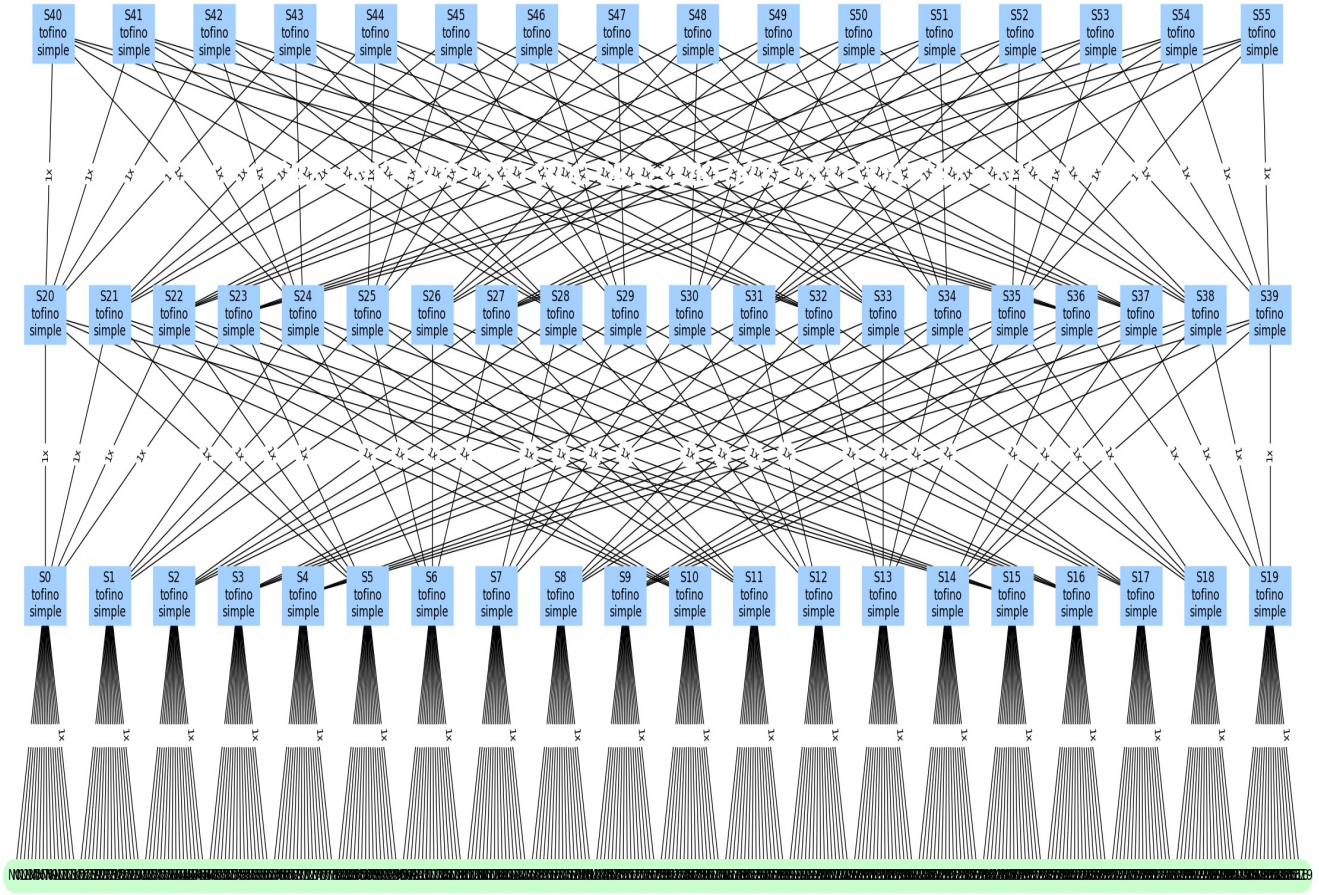
Simulation Overview

Simulation: Goals

- Show behavior of SFC
  - Application performance
- Metrics
  - Application performance: Flow Completion Time (FCT) slowdown
    - Compare the measured FCT with FCT of flow that is transmitted in unloaded network
- Methodology
  - Evaluate flow control mechanisms by measuring effects of incast traffic on non-incast traffic (background traffic) in the network
  - Transport protocol: RDMA with DCQCN
    - State-of-the-art flow control in modern RDMA NICs
    - Use variant with “initial window” mechanism (DCQCN+W) introduced in HPCC paper
  - Follow state-of-the-art approach:
    Li, Yuliang, et al. "HPCC: High precision congestion control." *ACM SIGCOMM* 2019.

Simulation Setup

- Simulation Software
  - NS3: Based on the open-source code published for the HPCC paper
    https://github.com/alibaba-edu/High-Precision-Congestion-Control
- Network topology (Follow HPCC)
  - 3-tier fat-tree (100/400GbE)
  - 320 nodes, 56 switches
  - Full bisection bandwidth
  - 12us round-trip time (RTT)
- PFC, DCQCN, SFC parameters in backup slides
  - Note: PFC + DCQCN is sensitive to tuning (workload-specific)



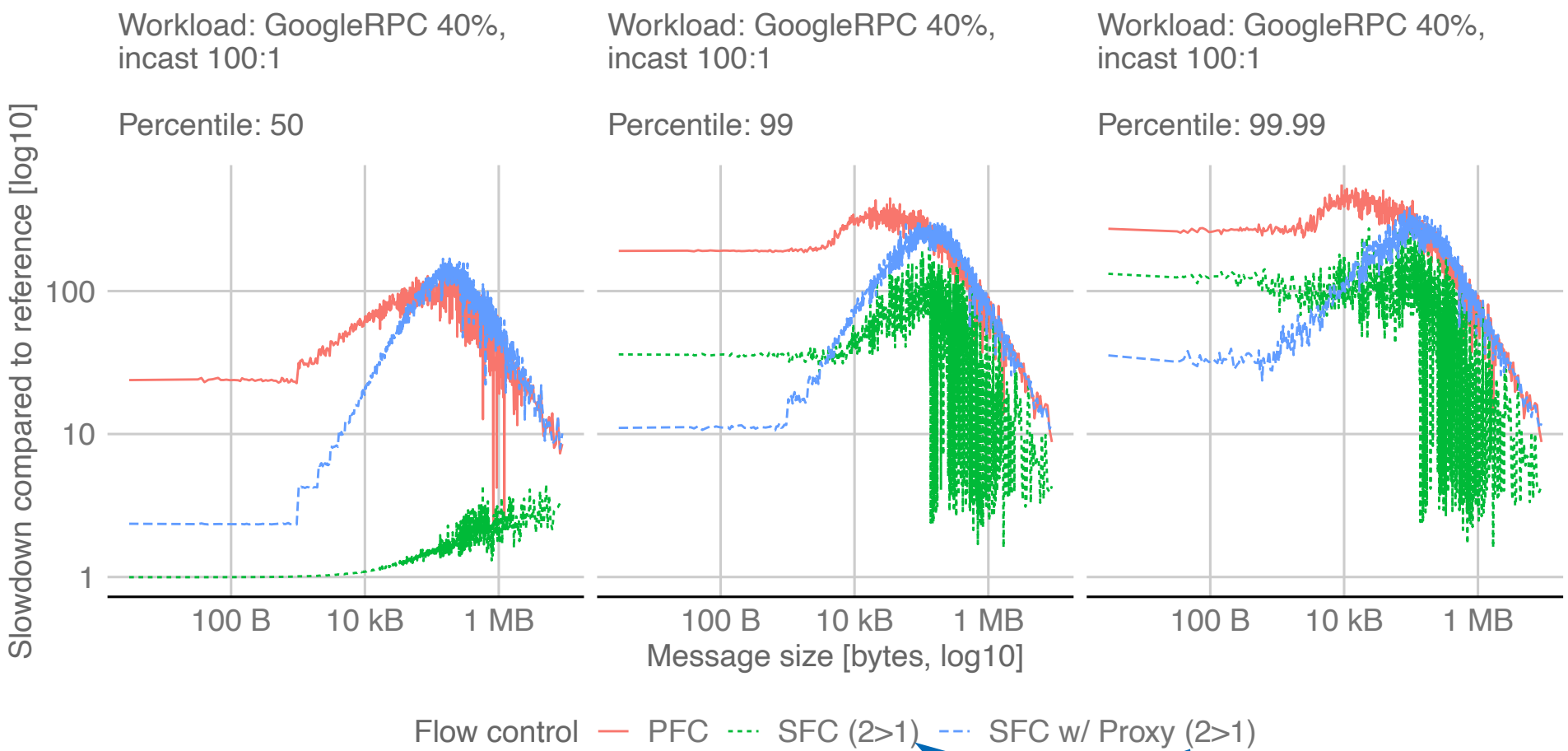
Comparing SFC with PFC

Traffic Load Configuration

- Background traffic
  - “normal” application workload on the network
  - 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 40%
  - Trace is 16x longer than the one presented at the IEEE 802.1 May 2022 Interim meeting
- In-cast traffic
  - 100:1 incast
  - Message size 256 KB
  - The incast traffic load is 10% of the network capacity

Flow Completion Time Results

- Key findings
  - Significant improvements of SFC and SFC w/ Proxy for small messages
  - SFC outperforms PFC in all percentiles
  - SFC w/ Proxy performs SFC for small message sizes and on par with PFC for large messages



Trigger >Target in BDP

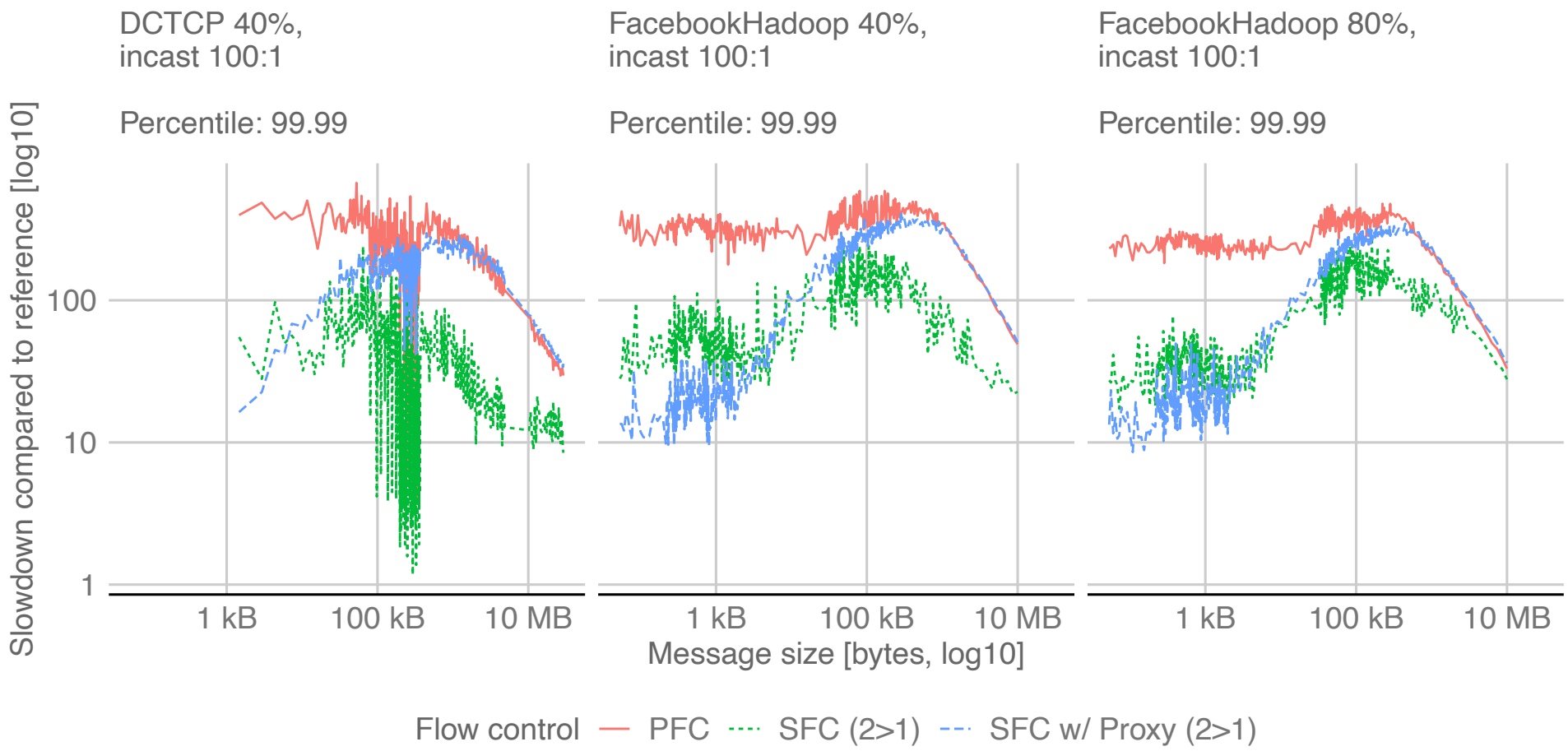
Comparing Workloads

Investigating a wider range of Workloads

- Background traffic
  - 320 hosts
  - Workloads
    - DCTCP
      Mohammad Alizadeh, Albert Greenberg, David A. Maltz, Jitendra Padhye, Parveen Patel, Balaji Prabhakar, Sudipta Sengupta, and Murari Sridharan. Data center tcp (dctcp). In Proceedings of the ACM SIGCOMM 2010 Conference, SIGCOMM '10, 2010.
      - Load factor: 40%
    - FacebookHadoop
      ArjunRoy,HongyiZeng,JasmeetBagga,GeorgePorter, and Alex C. Snoeren. Inside the social network's (dat- acenter) network. In Proceedings of the 2015 ACM Conference on Special Interest Group on Data Commu- nication, SIGCOMM '15, 2015.
      - Load factors: 40%, 80%
- In-cast traffic
  - 100:1 incast
  - Message size 256 KB
  - The incast traffic load is 10% of the network capacity

Flow Completion Time

- Key findings
  - Significant improvements of SFC and SFC w/ Proxy for small messages
  - SFC outperforms PFC significantly in all percentiles and all message sizes
  - The difference between SFC and SFC w/ Proxy for small messages sizes is less pronounced than in the GoogleRPC workload



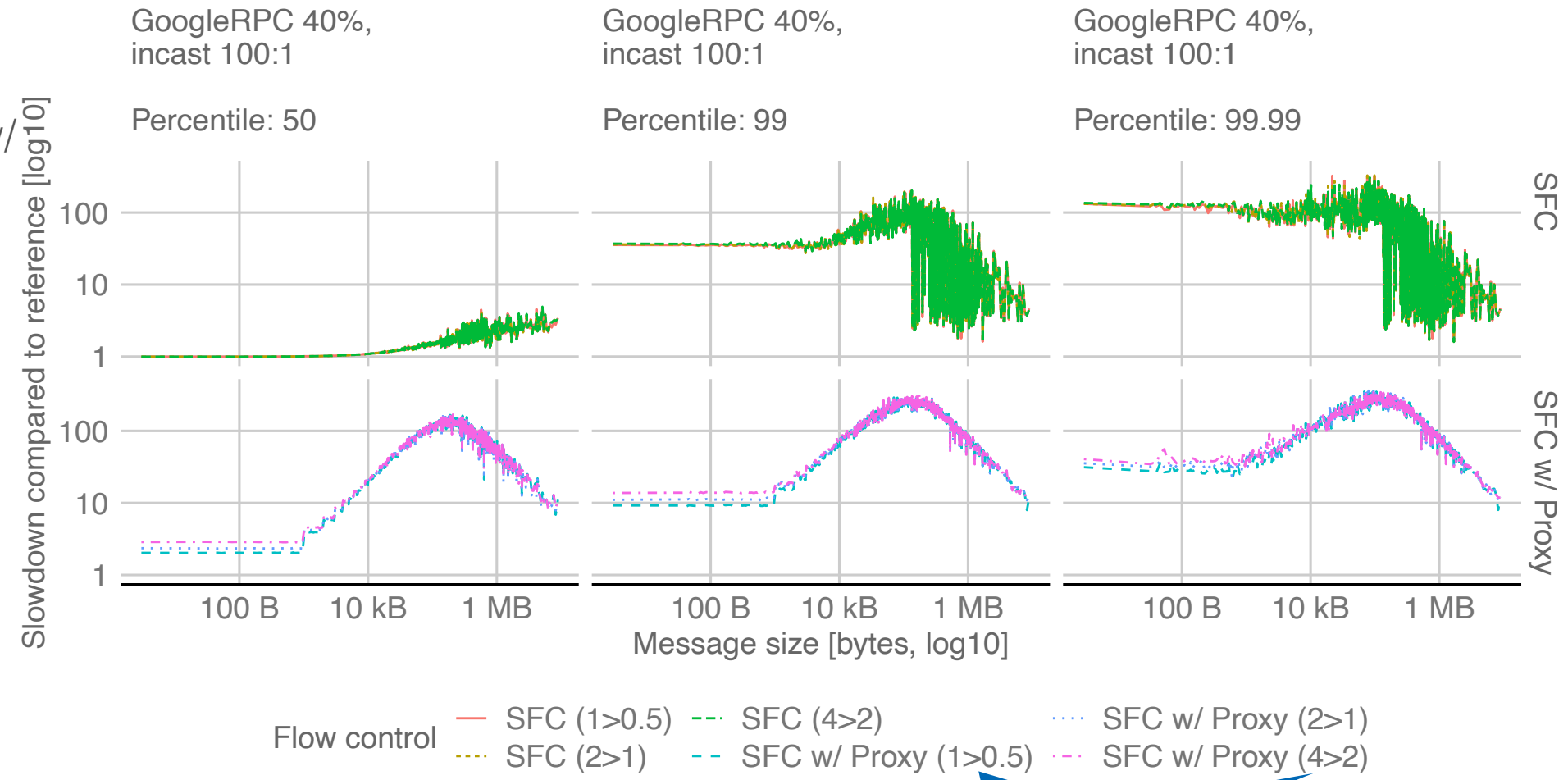
SFC Parameter Sensitivity

Investigating SFC sensitivity to trigger and target parameters

- Background traffic
  - 320 hosts
  - Workload: Google RPC
  - Load factor 40%
- In-cast traffic
  - 100:1 incast
  - Message size 256 KB
  - The incast traffic load is 10% of the network capacity
- SFC Parameters
  - Trigger threshold and target queue depth
  - BDP: 12us RTT * 100GbE = 150KB
  - Combinations: Trigger>Target
    - 1BDP > 0.5BDP
    - 2BDP > 1BDP
    - 4BDP > 2BDP

SFC Parameter Results

- Key findings
  - SFC and SFC w/ Proxy behave consistently with different parameter settings
  - Smooth response in FCT outcome for different parameter settings



Conclusion

Conclusion

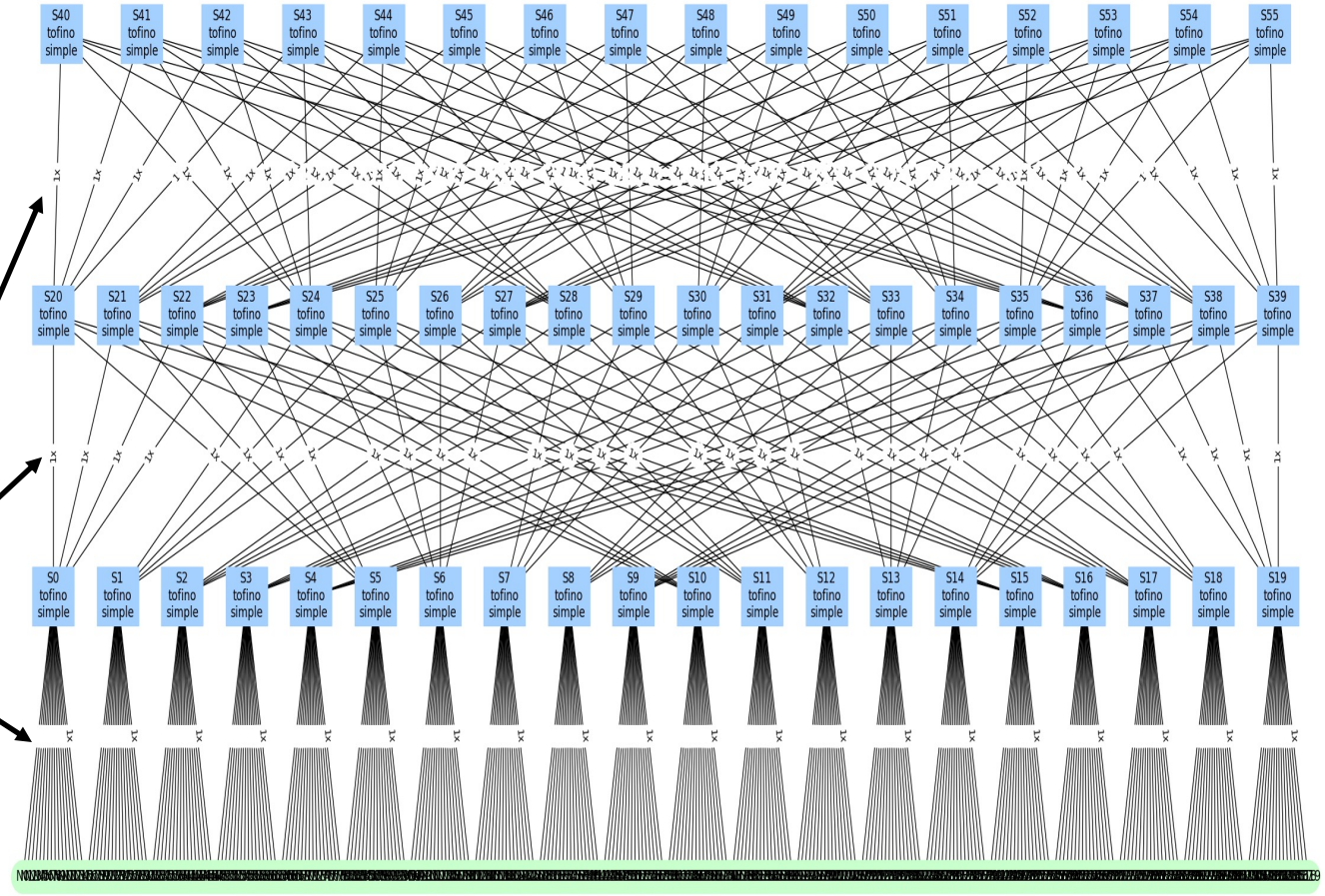
- SFC and SFC w/ Proxy performance
  - SFC can significantly reduce FCT for background traffic
  - Not surprisingly, workloads have a significant impact on performance improvements
  - FCT performance order is consistent: SFC > SFC w/ Proxy > PFC
    - SFC w/ Proxy: brown field approach that improves PFC performance without its operational downsides
  - Both SFC variants behaved consistently and good-natured to parameter changes
- Future work
  - 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: 4 KB
  - Egress guaranteed per port: 4 KB
  - Sharing mechanism: dynamic thresholding
    - Ingress coefficient: 2
    - Egress coefficient: 2
  - No ingress drops (set it to a very high value), but we can have egress drops
- SFC configuration
  - Suppression period: 6 us (1/2 RTT)
  - Destination cache: ToR
- PFC: Dynamically shared Ingress Buffer (HPCC paper)
  - Buffer: 13 MB pool + 3MB headroom
  - Dynamic sharing coefficient: 0.125
  - No egress drops

Congestion control configurations

- DCQCN
  - Using parameters as defined in: Li, Yuliang, et al. "HPCC: High precision congestion control." *ACM SIGCOMM* 2019.
  - Fast recovery steps: 1
  - Gain: 0.00390625
  - Rate increase timer: 900
  - Alpha timer: 1 us
  - CNP period: 4 us
  - AI: 0.05 Gbps
  - Hyper AI: 0.1 Gbps
  - Window: 15us
  - ECN threshold: 150 KB (1 BDP)