A Flow Control Scheme based on Per Hop and Per Flow in Commodity Switches for Lossless Networks

陽明交大資工系 王協源教授

December 21, 2021

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 3, 2021, accepted November 16, 2021, date of publication November 19, 2021,
date of current version November 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129595

A Flow Control Scheme Based on Per Hop and Per Flow in Commodity Switches for Lossless Networks

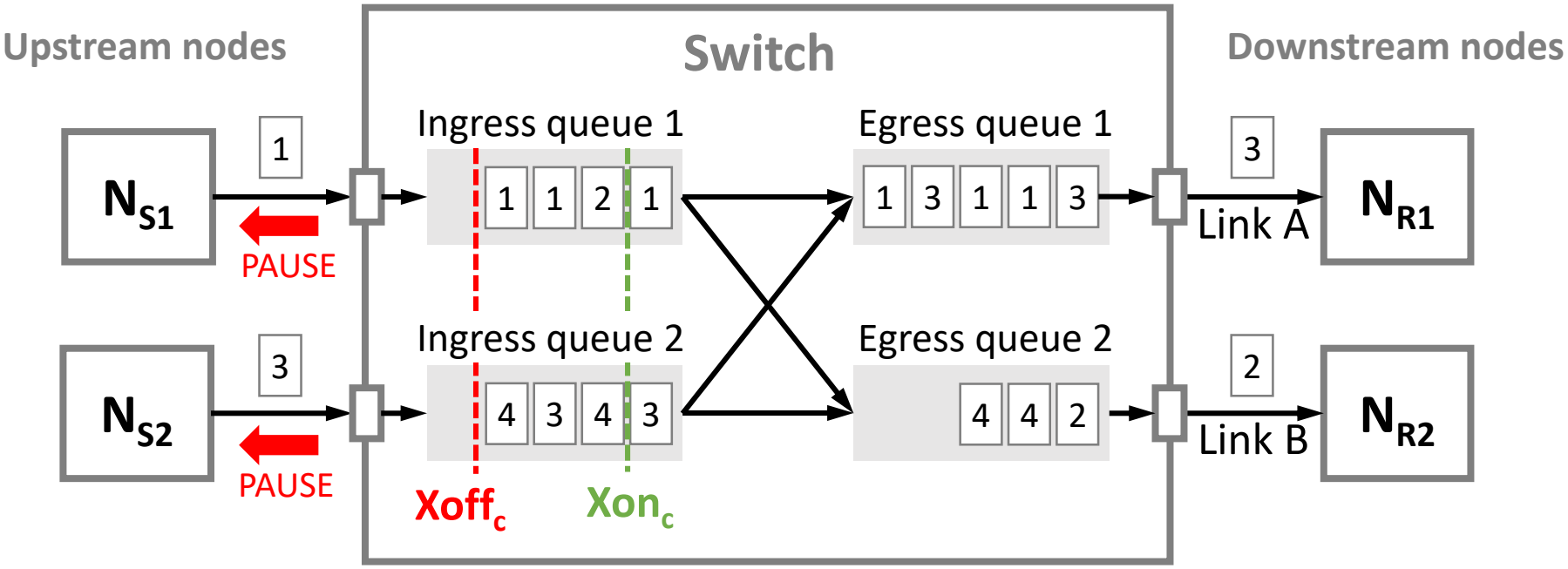
SHIE-YUAN WANG, (Senior Member, IEEE), YO-RU CHEN, HSIEN-CHUEH HSIEH, RUEI-SYUN LAI, AND YI-BING LIN, (Fellow, IEEE)
Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

Outline

- Motivation
- Design and Implementation
- Performance Evaluation
- Conclusion

Motivation

The Priority-based Flow Control (PFC) Design Used In A Switch



Performance Problems with PFC

- Congestion Spreading in PFC
  - May pause victim flows and spread the congestion
  - May decrease link utilization significantly
- Deadlock in PFC
  - May occur due to the Cyclic Buffer Dependency problem
  - May make an entire network enter a standstill situation
- Packet Loss in PFC
  - PFC is unable to avoid egress buffer overflow without a huge oversubscription ratio in buffer allocation
  - May reduce throughput and increase latency severely

PFFC Can Avoid Many Problems with PFC

- Control each flow individually so that the congestion will not be spread to other flows
- Each flow has its own buffer space, which eliminates the Cyclic Buffer Dependency and thus no deadlock will occur.

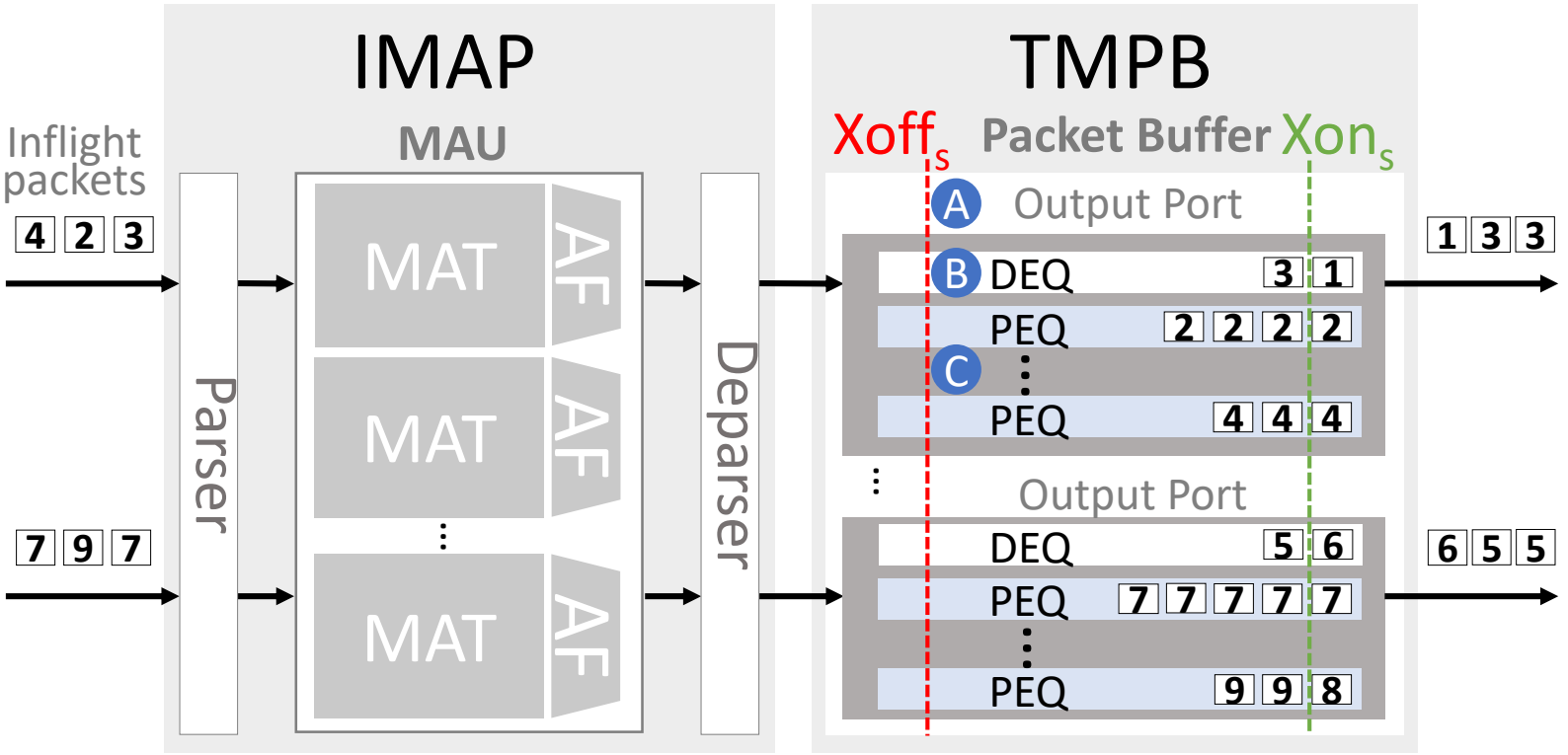
Our Contribution

- We design a novel per-hop per-flow flow control scheme named PFFC and successfully implement it in P4 hard-ware switches.
- Experimental results show that many serious problems with PFC such as congestion spreading, deadlock, packet loss, etc. are all gone in PFFC and the average flow completion time of mice flows in PFFC is shorter than that in PFC.
- Results has been published in IEEE Access, November 2021.

Design and Implementation

Per-flow Pause and Resume Designs

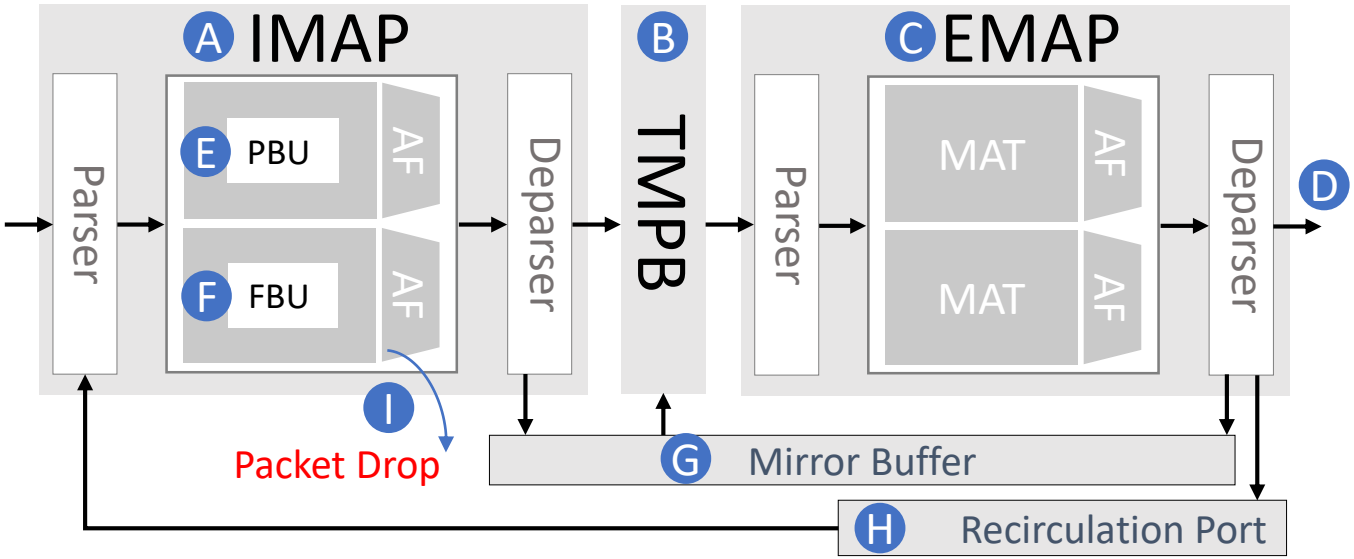
The Egress Queue Design in an Output Port (in PFFC)



MAU: Match-Action Unit
DEQ: Default Egress Queue
PEQ: Pause Egress Queue

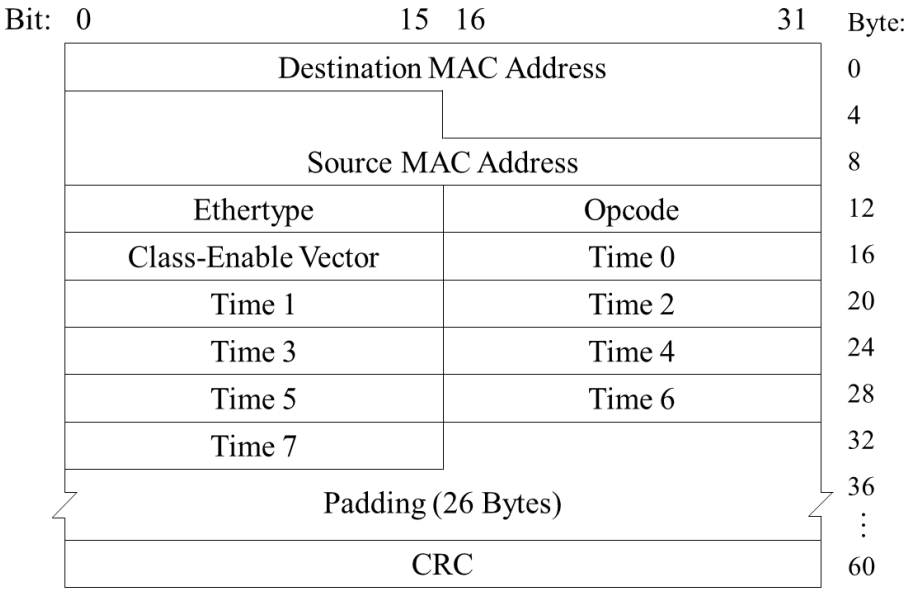
Per-flow Buffer Usage Accounting Designs

The Per-flow Buffer Usage Accounting Design in PFFC

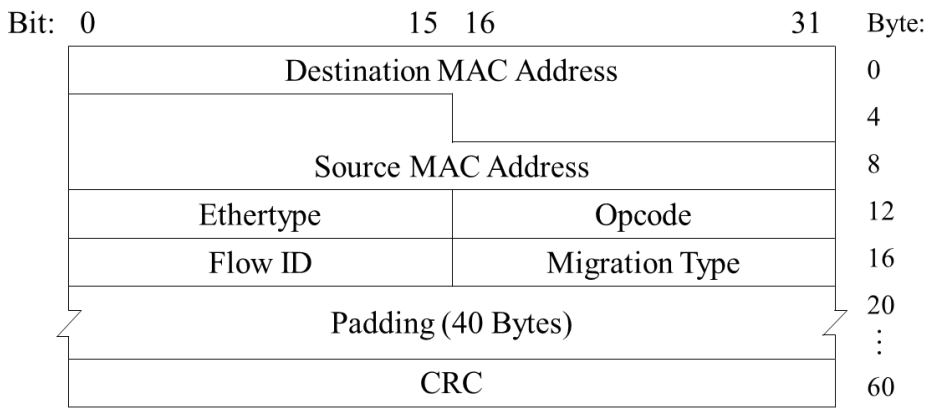


Per-flow Pause/Resume Frames Generation and Processing Designs

The Control Frames Used in PFFC



The format of the PFC control frame



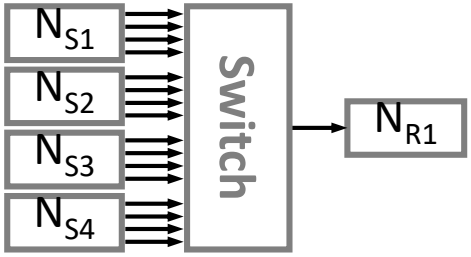
The format of the PFFC control frame

Performance Evaluation

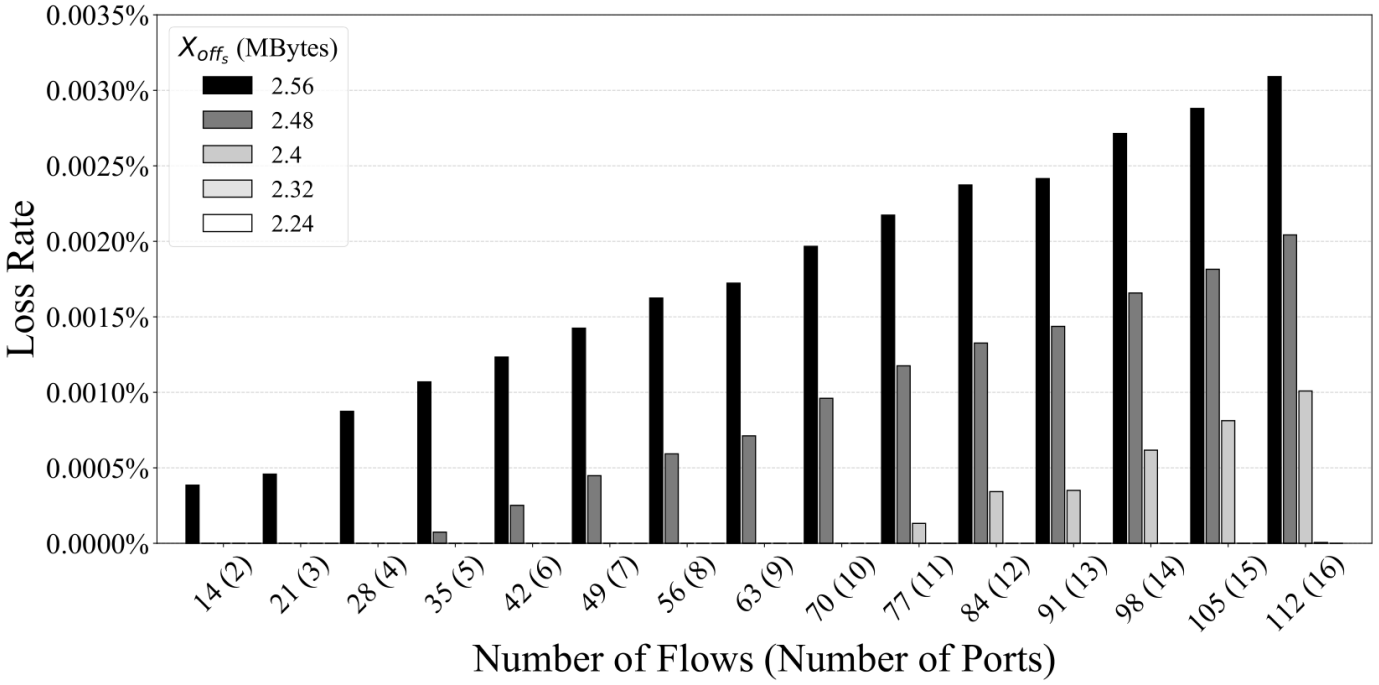
Experimental Setup

- Five hosts
  - Each has a 12-core Intel 3.2 GHz i7-8700 CPU and 16 GB RAM with Intel X710 network interface cards.
- Four P4 hardware switches
  - Breakout 40G QSFP+ ports to four 10GbE ports and connect each 10GbE port to a host.
  - 8 ingress queues for every 10GbE port
- Use iperf version 2.0.13 to generate UDP and TCP traffic.

$Xoff_s$ Threshold Exploration

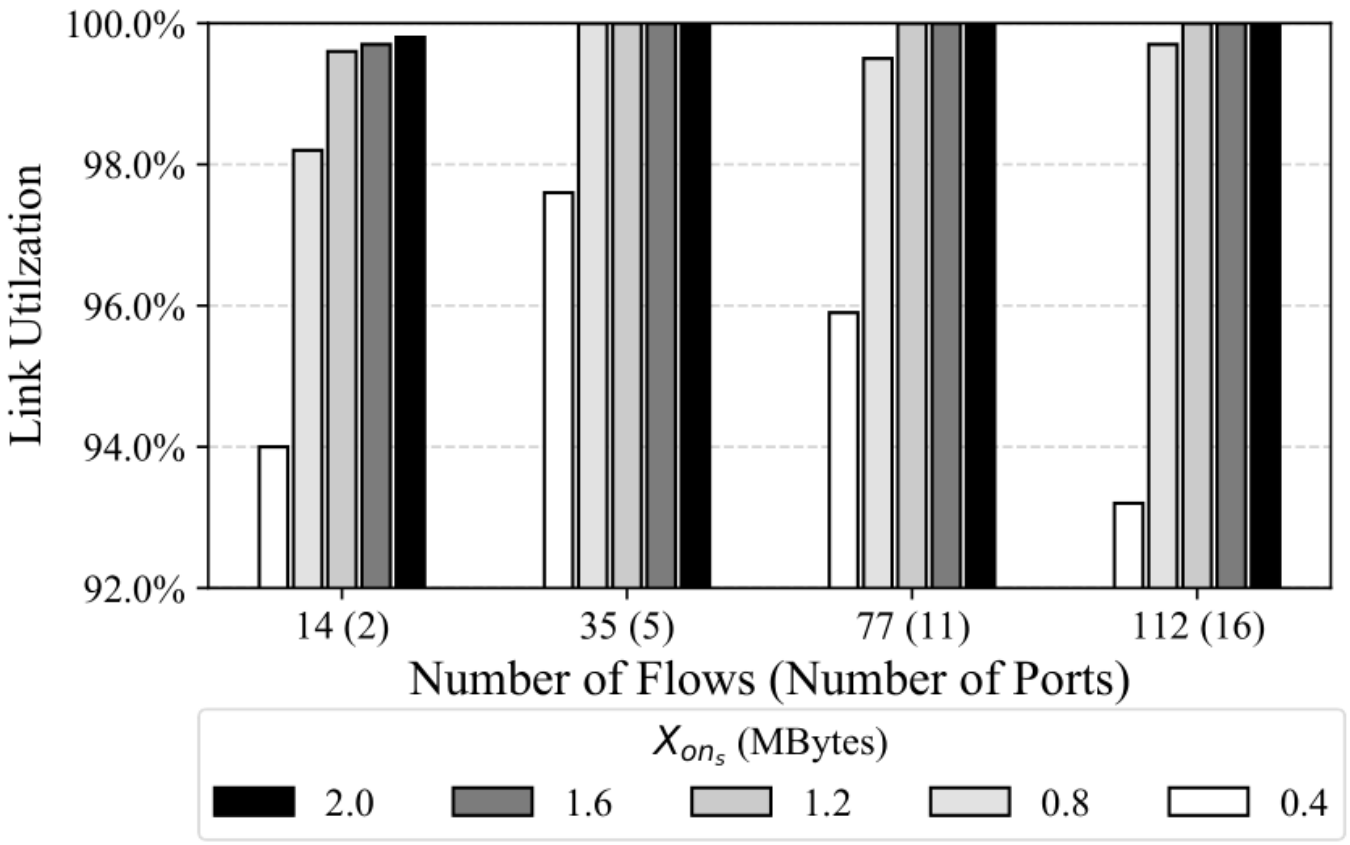


The used network topology



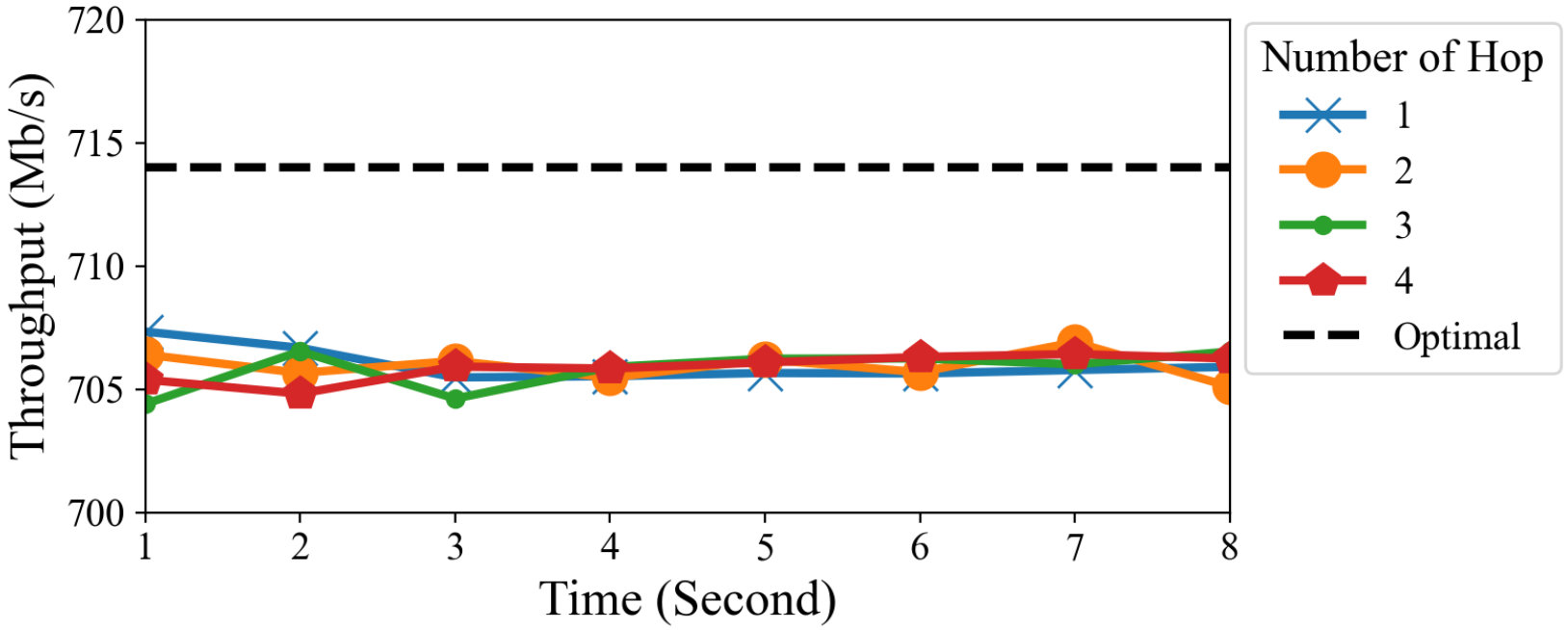
The Packet Loss Rate vs. Number Of Flows

Xon_s Threshold Exploration



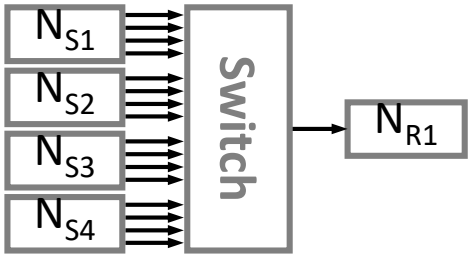
PFFC Is Scalable on Multi-Hop Networks

Multi-hop topologies: the number of hops (N) is ranged from 1 to 4.

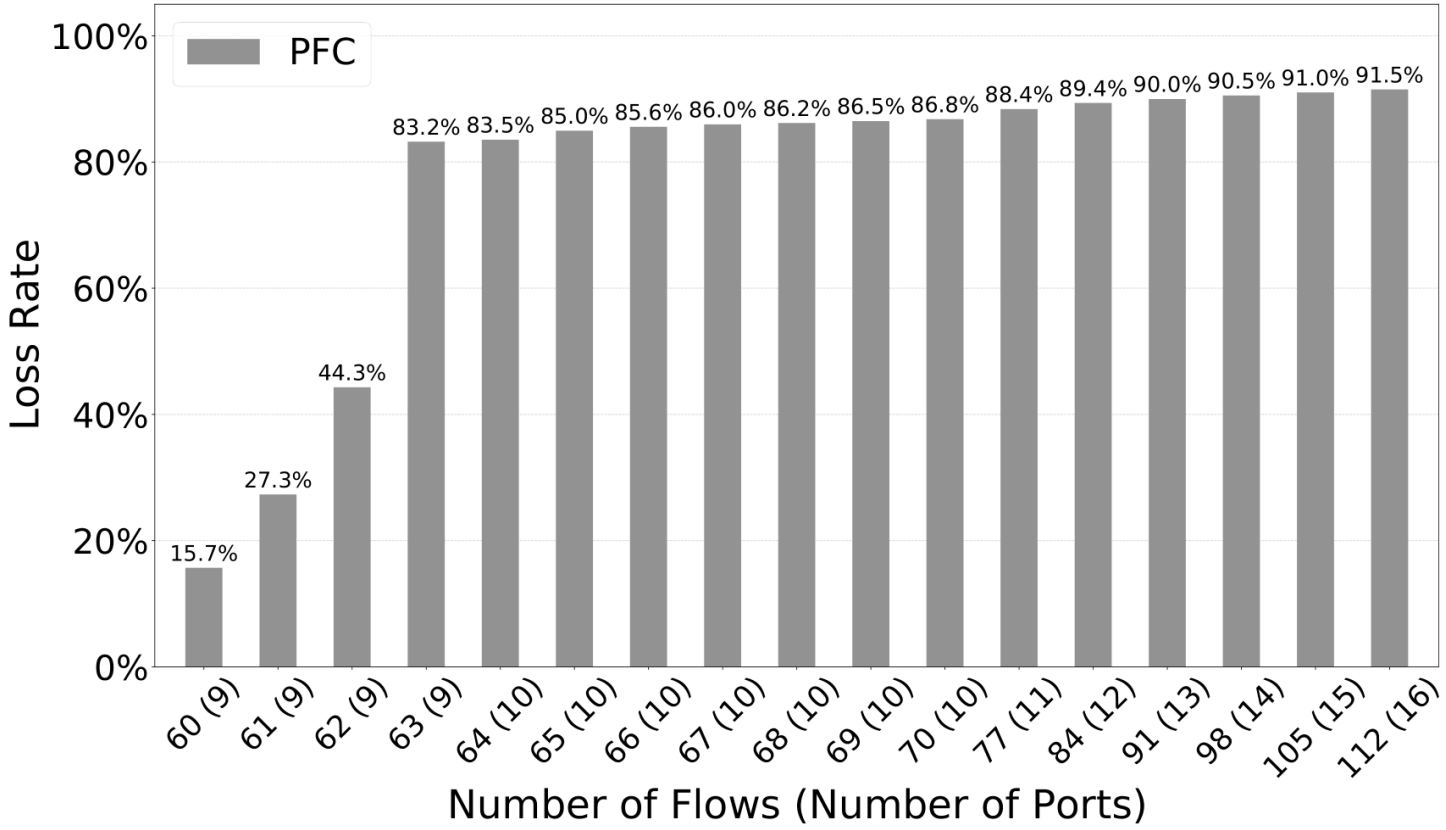


The average throughput of TCP flows on multi-hop topologies

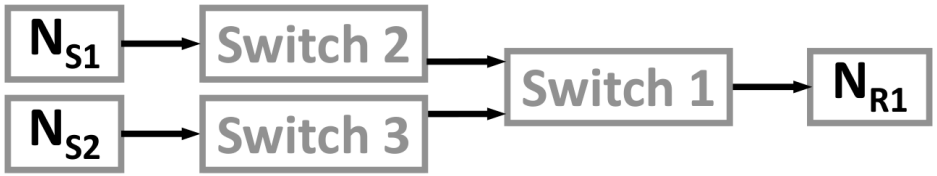
Packets May be Lost in PFC



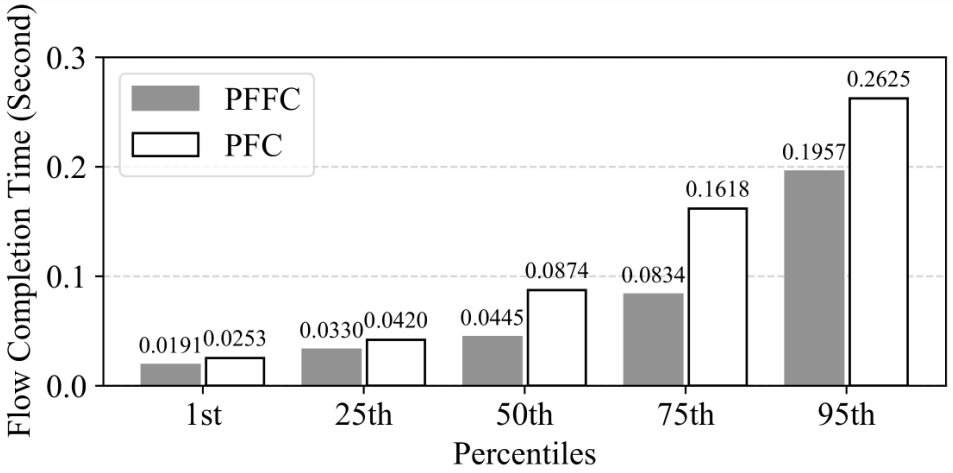
The used network topology



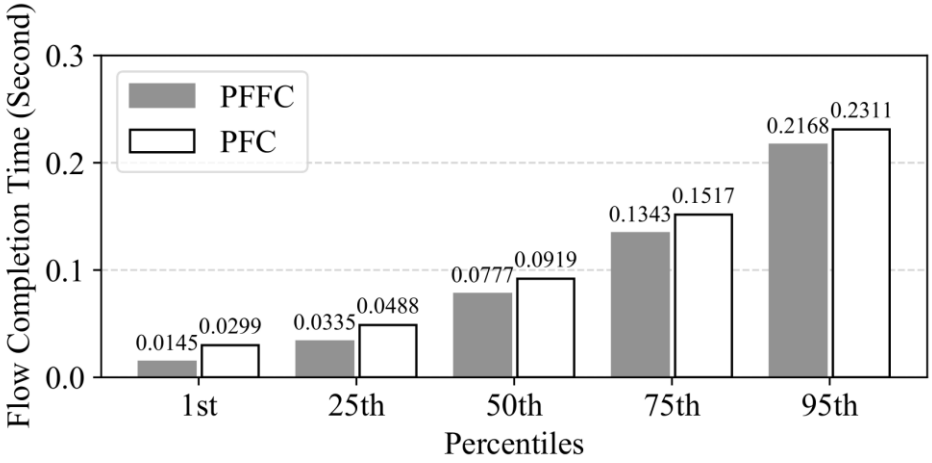
Flow Completion Time of Mice Flows of PFFC is shorter than that of PFC



The network topology used for evaluating mice flows mixed with elephant flows

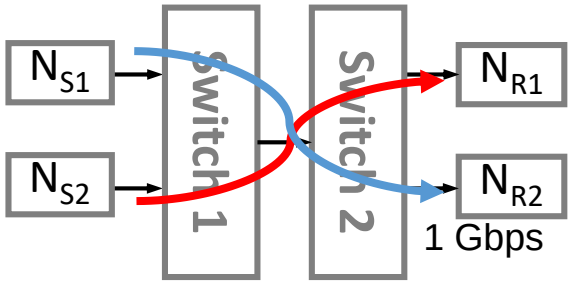


The flow completion time of TCP mice flows mixed with UDP elephant flows



The flow completion time of TCP mice flows mixed with TCP elephant flows

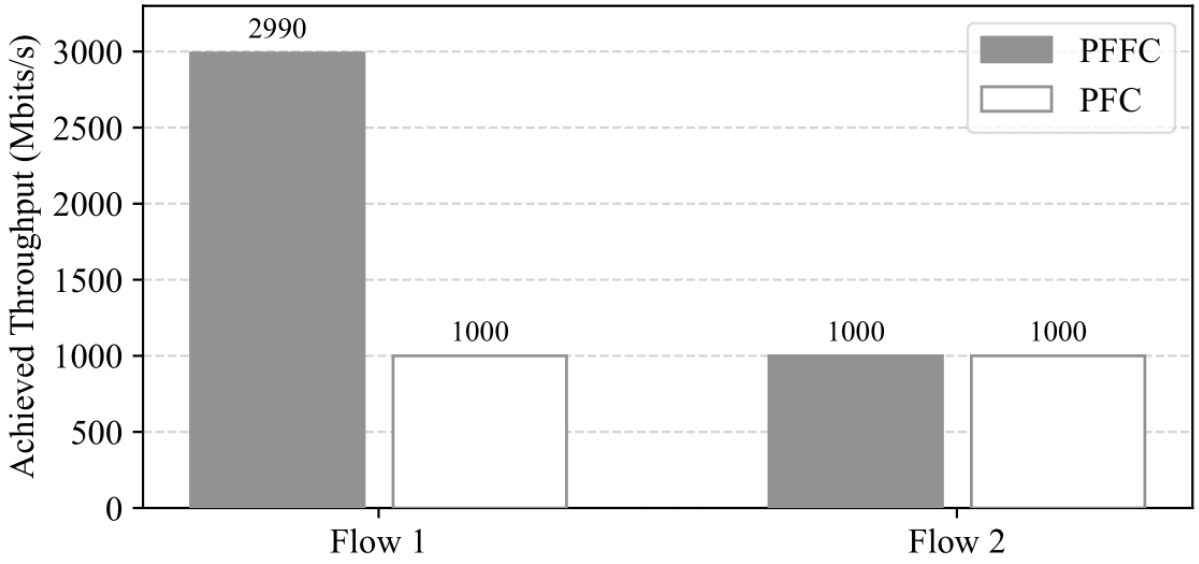
In PFC, The Throughput of Flow 2 Is Unnecessarily Affected by Flow 1



The used network topology

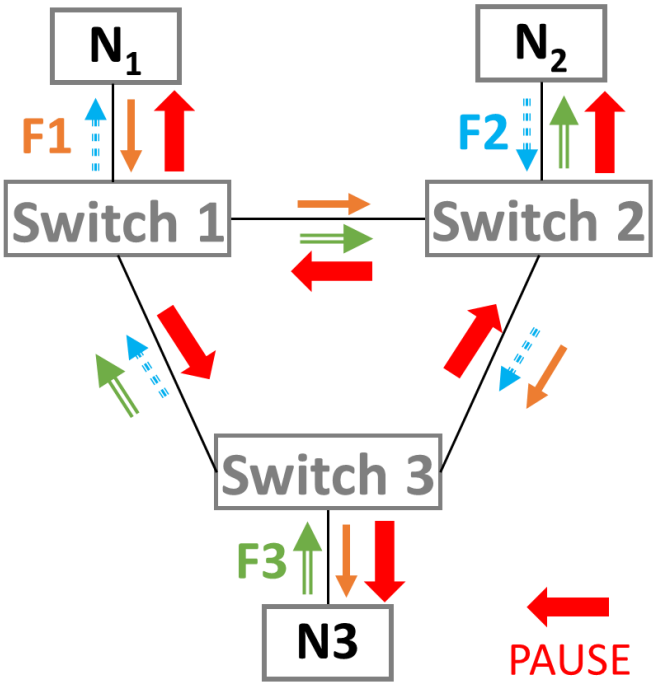
The bandwidth of all other links is 10 Gbps.

The sending rate of the two flows is 3 Gbps.

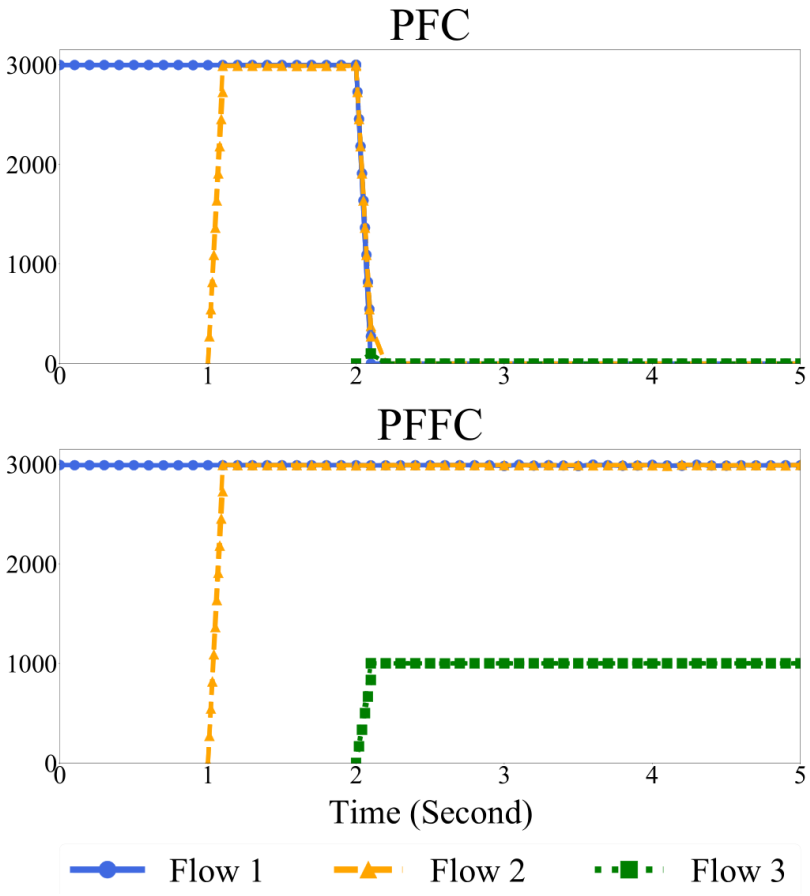


The achieved throughputs of two flows in PFC and PFFC

Deadlocks May Occur in PFC But Not in PFFC

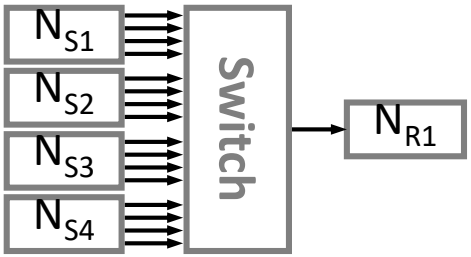


The network topology used for showing PFC deadlock problems



The interaction among three flows in PFC and PFFC

Bandwidth Overhead of PFFC Is Only Slightly Higher than That of PFC



The used network topology

TABLE 3. Bandwidth overhead of PFC and PFFC.

	Frame count	Mbyte count	Overhead over link (%)
PFC	156,461	100	0.40
PFFC	211,010	135	0.54

Conclusion

Conclusion

- We have designed a novel per-hop per-flow flow control scheme named PFFC and successfully implemented it in P4 hardware switches.
- Experimental results show that many serious problems with PFC such as congestion spreading, deadlock, packet loss, etc. are all gone in PFFC and the average flow completion time of mice flows in PFFC is shorter than that in PFC.
- Besides, the control frame overhead of PFFC is only slightly higher than that of PFC.

Thank you for your attention

Q&A