All You Need is Deep Buffer: TCP flows on Long-distance Networks

K. Koizumi, G. Honjo, J. Shitami, J. Tamatsukuri, H. Tezuka, M. Inaba, K. Hiraki

The University of Tokyo

[ImPACT Program] Planned Serendipity セレンディピティの計画的創出

Background

- Network bandwidth has been increasing and the use of 100 Gigabit Ethernet is starting to be widely spread in the backbone and high-performance cluster networks
- Many TCP congestion control algorithms have been proposed for fair use of network bandwidth among multiple TCP streams on long fat-pipe networks (LFN)
- **Fairness between the TCP streams is not maintained because congestion control at each node does not work**

Experimental Settings

- We analyze the performance of multiple TCP streams with different capacities and congestion control algorithms in laboratory networks
- We evaluate TCP congestion control algorithms with the network simulator ns-3 and real network testbeds with up to 32x 1 GbE end nodes, 2x 10 GbE end nodes, ANUE Network Delay Emulator, and 1/10 GbE network switches
- The bandwidth of inbound traffic is over 16 Gbps and that of a bottleneck link is 10 Gbps

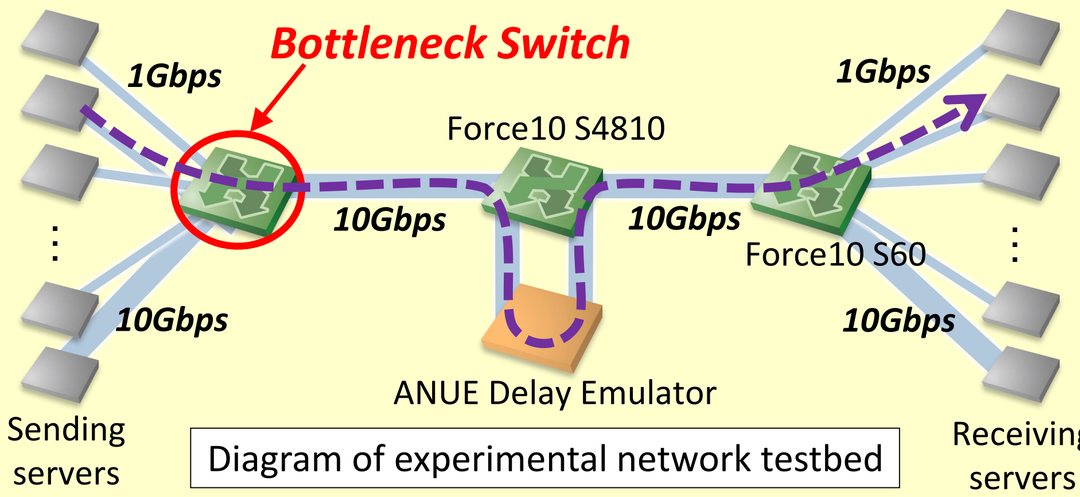


Diagram of experimental network testbed

Server & network specifications

CPU	Intel Core i7 3770K (Ivy Bridge) 3.50 GHz 4 cores / 8 threads
Mem	DDR3 32GB
NIC	Chelsio S310E-SR (10 GbE) T590-LP-CR (40 GbE)
OS	CentOS 7.2.1511 Linux Kernel 3.10.0-327.el7.x86_64
NIC	Intel Gigabit CT Desktop Adapter (1G) Chelsio T310 Single Port Adapter (10G) Realtek 8111F (control)
App	iperf 3.0.11
10 GbE NDE	ANUE Network Emulator (RTT=100 or 200 ms)

Buffer size of switches

Force10 S60	Foundry FESX424	Brocade ICX6650
1.25 GB	32 MB	8 MB

Discussion

Deep buffers

- In **FESX424** and **ICX6650**:
  - Close-to-wire-speed UDP traffics are lost because of packet burst
  - Many packet losses lead to lowering the total throughput
- In **S60** (it has the largest buffer):
  - S60 buffer mitigates the bursts
  - S60 achieves wire speed and fair data transfer in many TCP algorithms and both RTTs; deep buffer is all you need
- One 10 Gbps stream is delivered equally with 1 Gbps streams with S60 & ICX6650, however, with FESX424, throughput of 10G is unfairly high

TCP algorithms

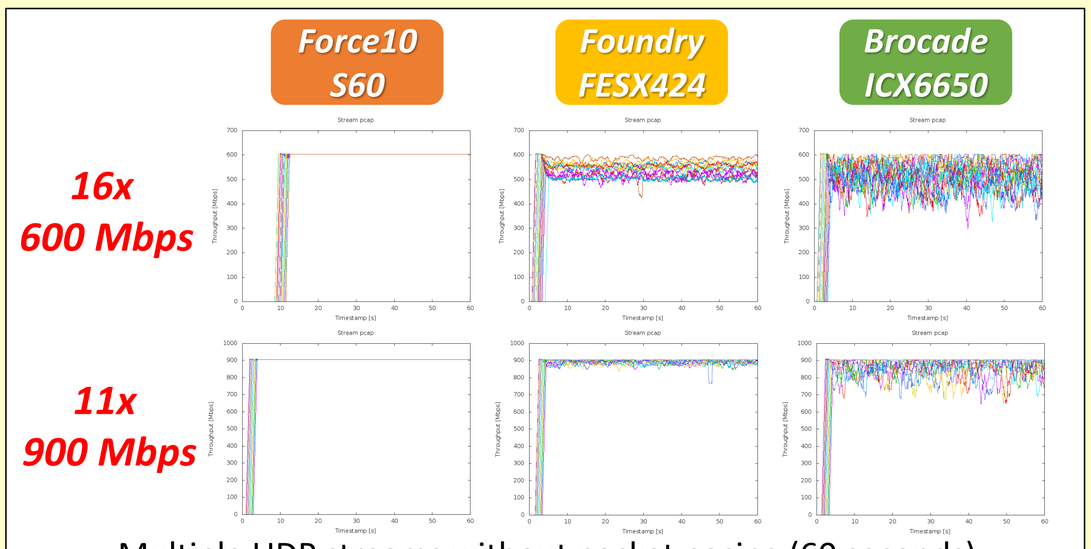
- **CUBIC** increases throughputs after congestions; however, the streams are competing for bandwidth and biased
- The throughput of each **H-TCP** stream converges at an appropriate equally divided rate
- **H-TCP** also shows better performance in the total throughput and fairness index

革新的研究開発推進プログラム ImPACT Impulsing Paradigm Change through Disruptive Technologies Program

This research was partially funded by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan)

Experimental Results

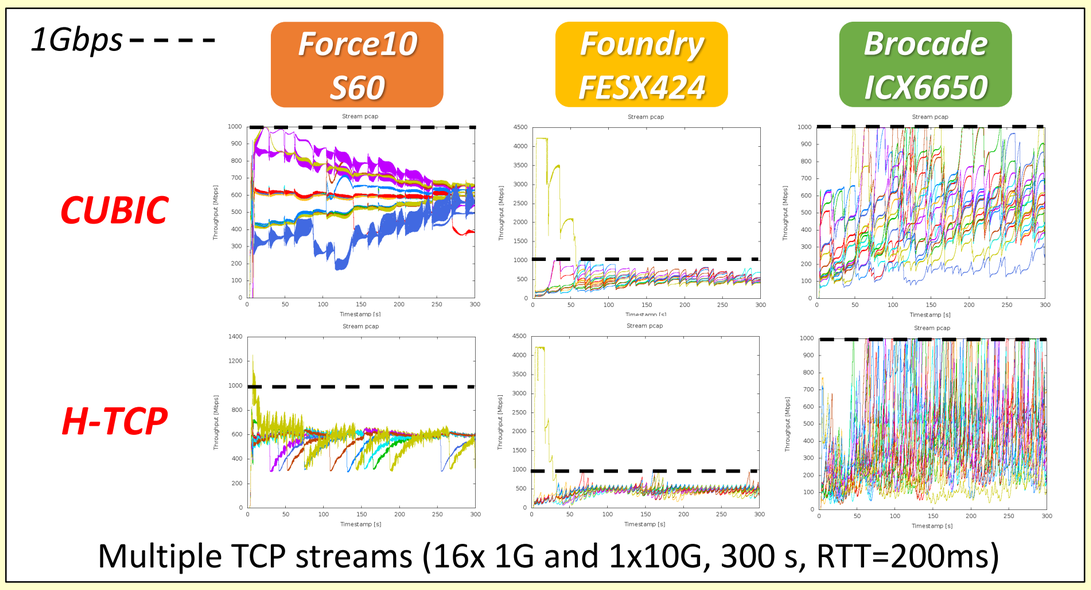
UDP throughput performance



Multiple UDP streams without packet pacing (60 seconds)

- S60 mitigates the packet bursts and achieves wire-speed UDP tranfer

TCP throughput performance

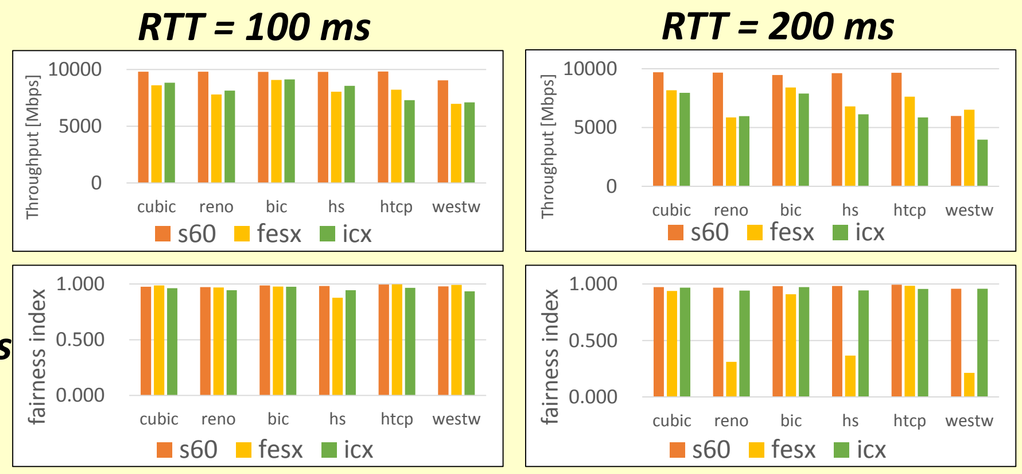


Multiple TCP streams (16x 1G and 1x10G, 300 s, RTT=200ms)

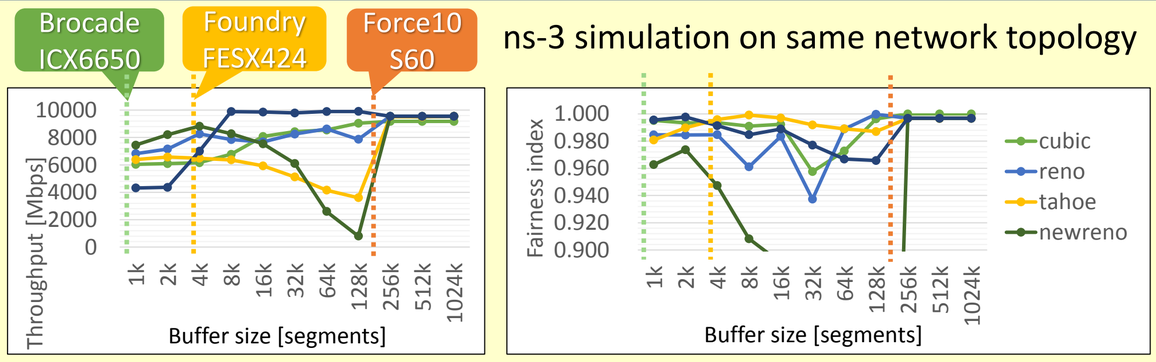
- CUBIC flows are competing for bandwidth, and H-TCP flows converge at an appropriate throughput

Average performance of multiple TCP streams

- We evaluate the throughput and fairness of three bottleneck switches



ns-3 Simulation



Throughput

Fairness index

Concluding Remarks

- We compare TCP algorithms and 3 bottleneck switches with packet buffer sizes of 1.25 GB, 32 MB, and 8 MB
- A deep buffer switch achieves the wire-speed data transfer rate with multiple TCP streams regardless of TCP algorithm
- Results about the buffer size match the ns-3 simulation results
- We clarified differences of multi-flow behavior in TCP algorithms