HPCC++: Enhanced High Precision Congestion Control

draft-miao-ccwg-hpcc

draft-miao-ccwg-hpcc-info

Rui Miao, Surendra Anubolu, Rong Pan, Jeongkeun Lee, Barak Gafni, Jeff Tantsura, Allister Alemania, Yuval Shpigelman

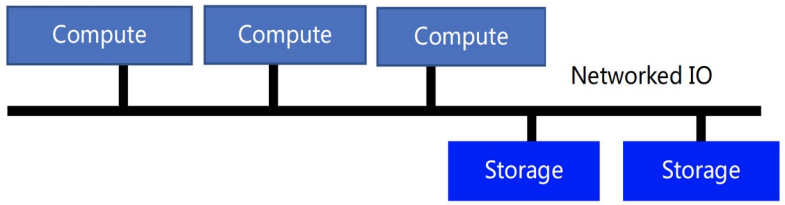
IETF-120 CCWG

Jul 2024

Cloud desires hyper-speed networking

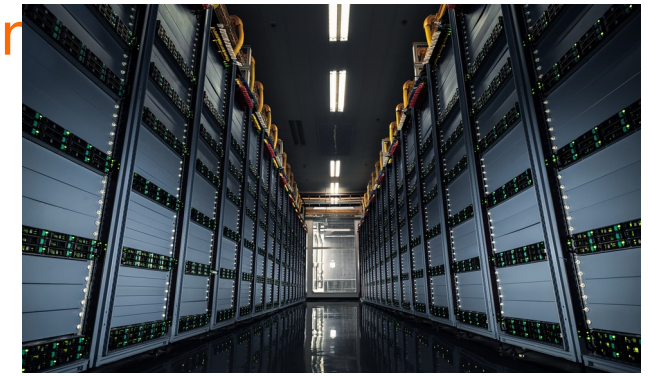
Today, clouds have ⇐ bigger data to compute & store
faster compute & storage devices
more types of compute and storage resources

High-performance storage



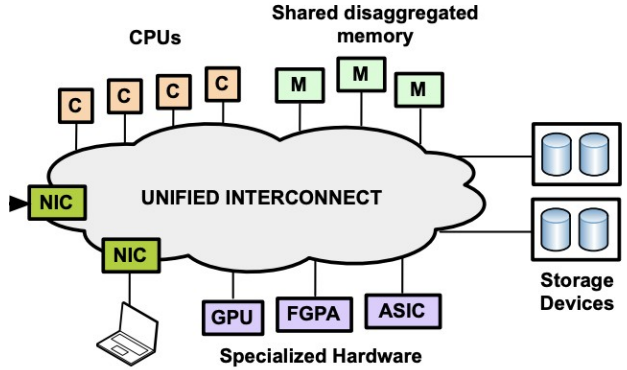
- Storage-compute separation is norm
- HDD→SSD→NVMe
- Higher-throughput, lower latency
- 1M IOPS / 50~100us

High-performance cor



- Distributed deep learning, HPC
- CPU→GPU, FPGA, ASIC
- Faster compute, lower latency
- E.g. latency <10us

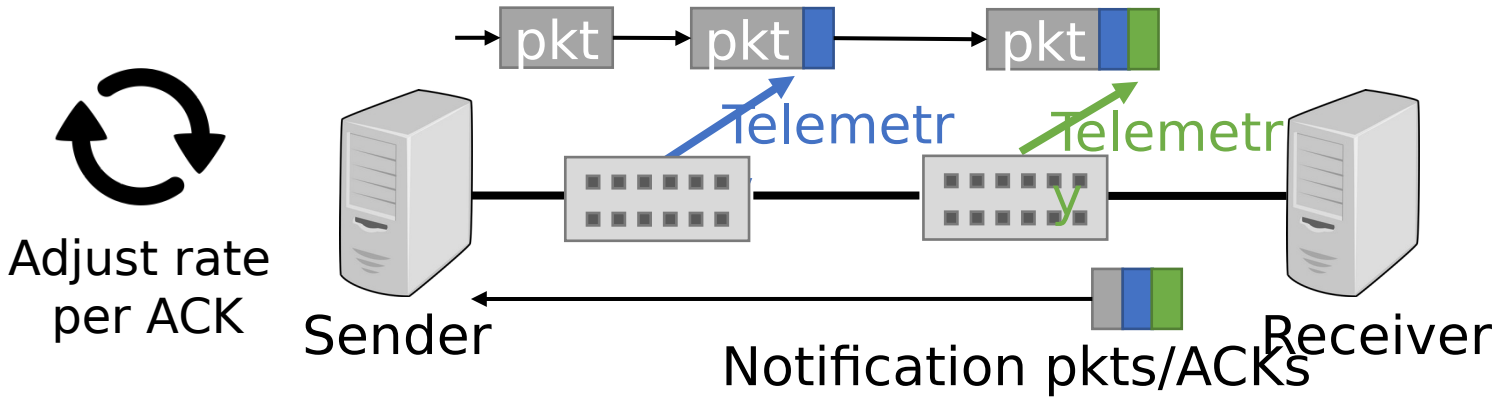
Resource disaggregation



- More network load
- Need ultra-lower latency: 3-5us, > 40Gbps (Gao Et.al. OSDI'16)

HPCC++: Enhanced High Precision Congestion Control

- New networking ASICs have in-band telemetry capabilities
- Packets collect telemetry on their route
- Can we use in-band telemetry as precise feedback for congestion control?



Support CSIG (Congestion SIGnaling)

- CSIG is a compact in-band telemetry [I-D.ietf-ravi-ippm-csig]
  - fixed-size aggregate metric computed over the hop devices
  - Compact (supported) and expanded (require further defined) formats with Type-Value format
  - u’ = u1 + u2, where u1 = min(ack.L[i].qlen,L[i].qlen) / (ack.L[i].B*T) u2 = txRate / ack.L[i].B

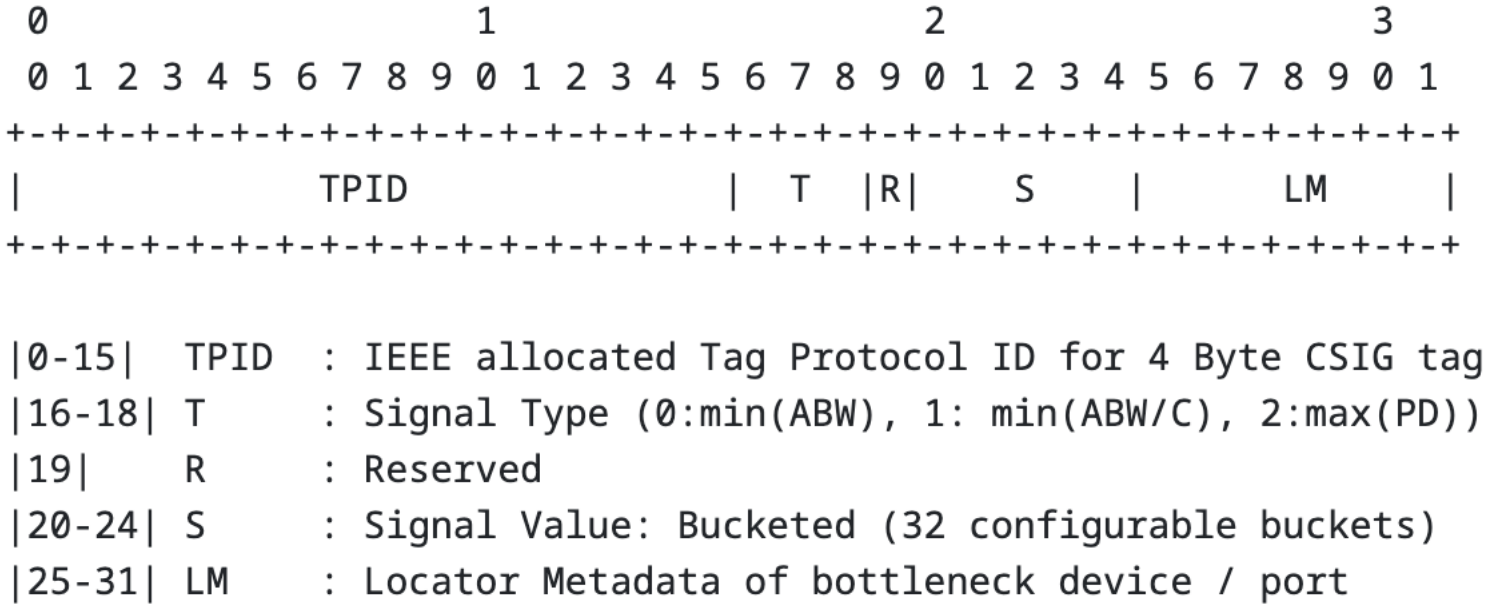


Figure 4: CSIG Compact Header Format

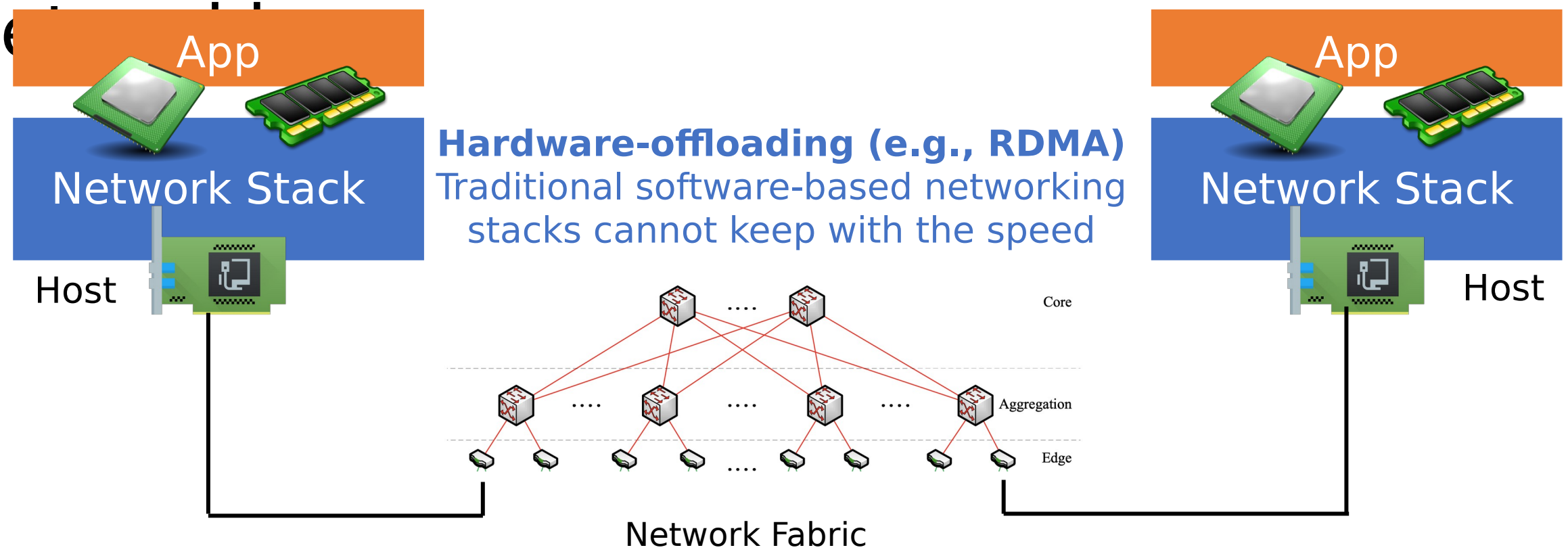
Support CSIG (Congestion SIGnaling)

- Handle path changes
  - Two consecutive packets are required for the new path
  - Max(qlen/B) can be interpreted as an `expected` sojourn time for the packet in the tail
- max(PD) is an approximate of max(qlen/B)
  - max(PD) is sojourn time of signal-carrying packet. i.e., packet at the tail
  - max(qlen/B) is the queue length when the packet is dequeuing.

Your Feedback is Appreciated!

BACK UP

to form hyper-speed ne



Hardware-offloading (e.g., RDMA)
Traditional software-based networking stacks cannot keep with the speed

Real-time Congestion control (CC)
Lots of data and communication => more pressure on the network

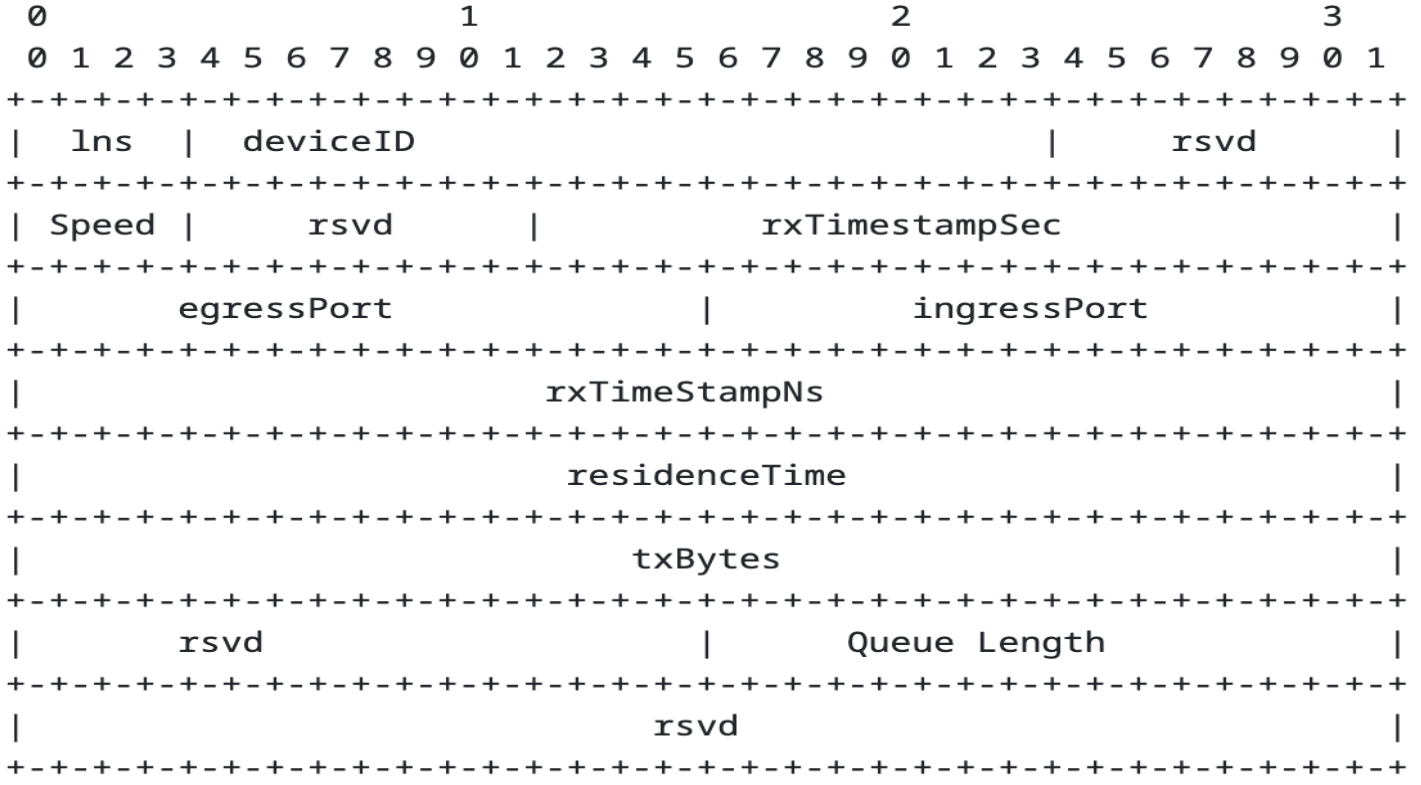
HPCC++ provides ideal performance

g in-band telemetry as the precise feedback

- Fast convergence
  - Sender knows the precise rate to adjust to
- Near-zero queue
  - Feedback does not only rely on queue
- Few parameters
  - Rich and precise feedback, reduces heuristics which requires more parameters

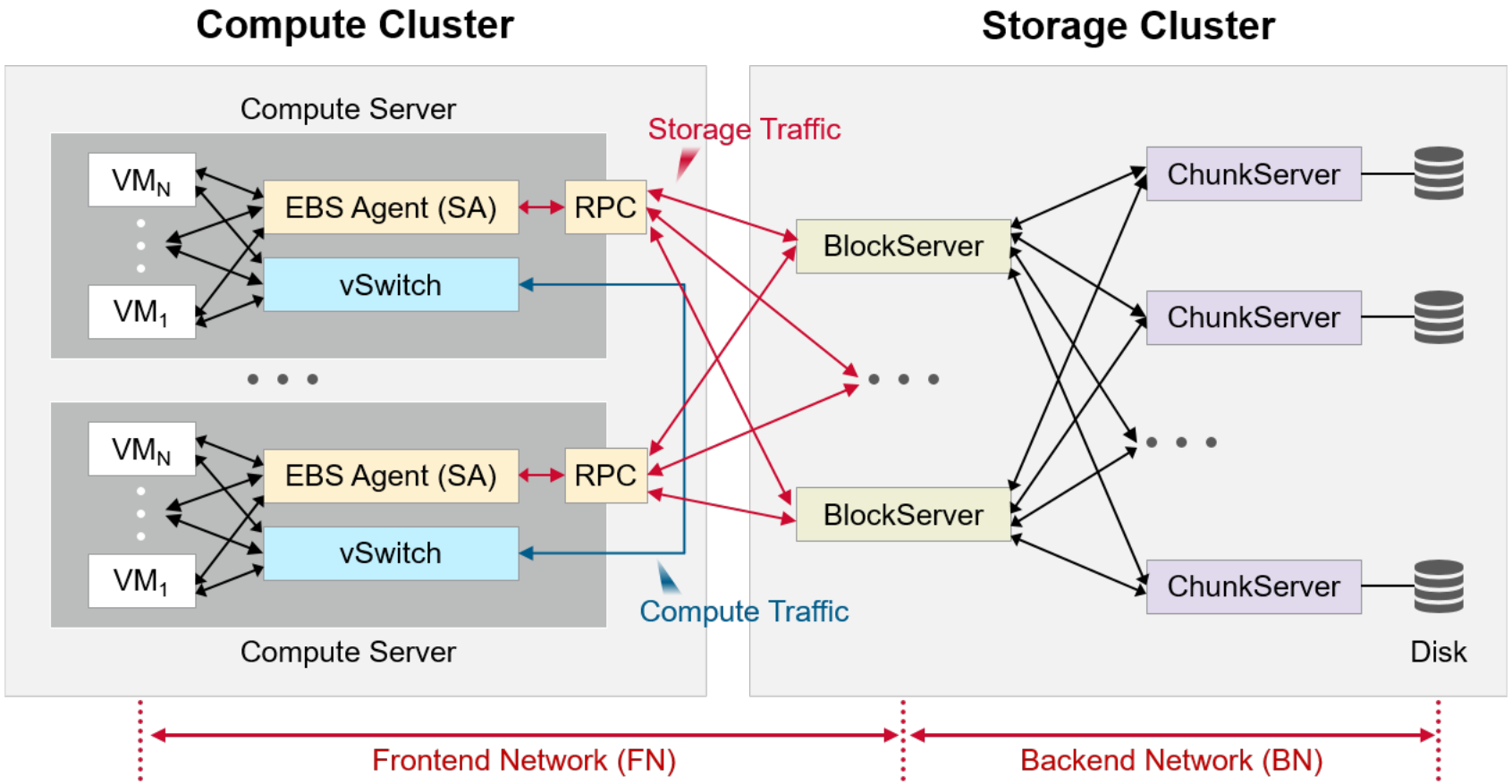
Our proposals

- draft-miao-ccwg-hpcc
  - Defines the algorithm using telemetry information, including queue length, transmitted bytes, timestamp, link capacity, etc.
- draft-miao-ccwg-hpcc-info
  - Provides environment-dependent packet formats of telemetry encodings, including IFA2.0, IOAM, P4



Deployment experience

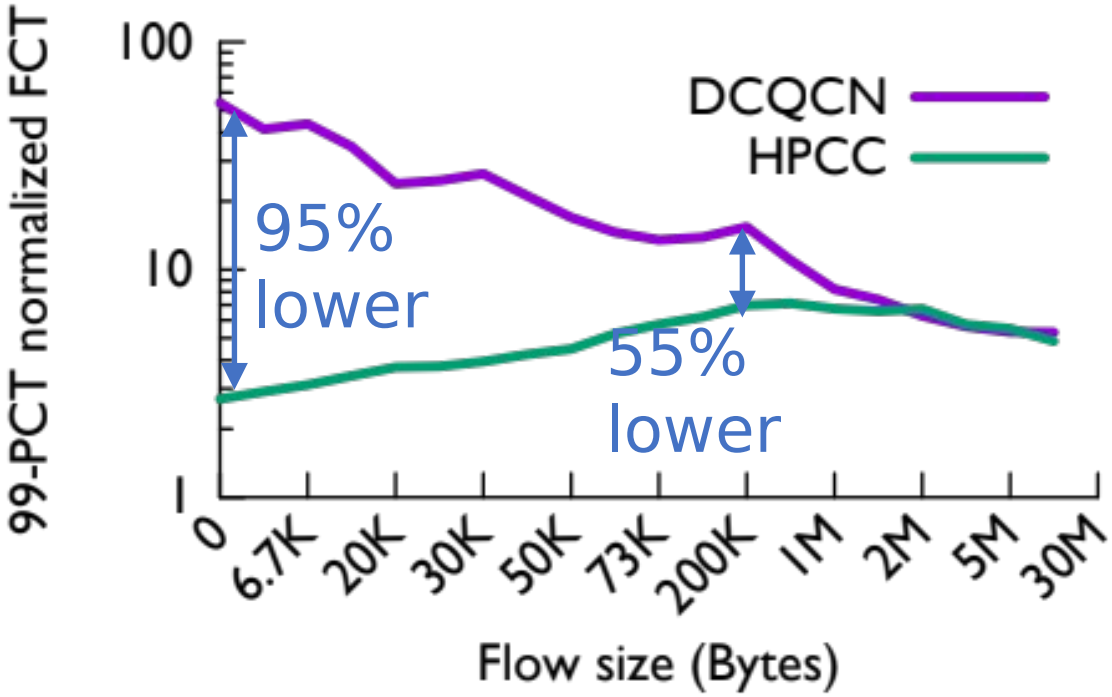
- Deployed widely for storage, AI training, and database applications in one of the major cloud providers
- Achieved significant boosts in throughput and latency

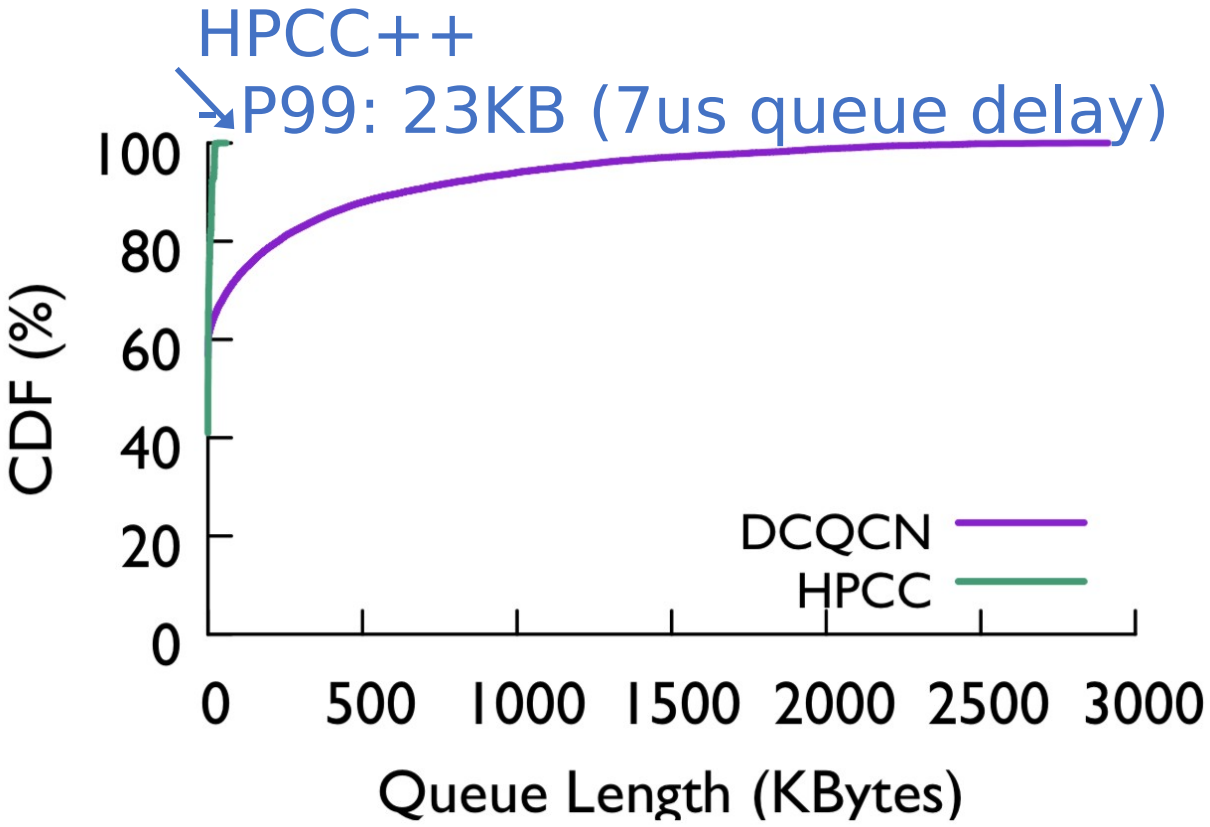


ce: https://www.broadcom.com/blog/high-precision-congestion-control

HPCC++ achieves lower FCT and near-zero queue

- In testbed, vs. DCQCN (hardware-based, widely used in industry)
  - Web search traffic at 50% load
- Extensive tests compared with other CC in simulation. HPCC performs better





Thank You