HPCC++: Enhanced High Precision Congestion Control

draft-pan-tsvwg-hpccplus-02

Rui Miao, Hongqiang Harry Liu, Rong Pan, Jeongkeun Lee, Changhoon Kim, Barak Gafni, Yuval Shpigelman

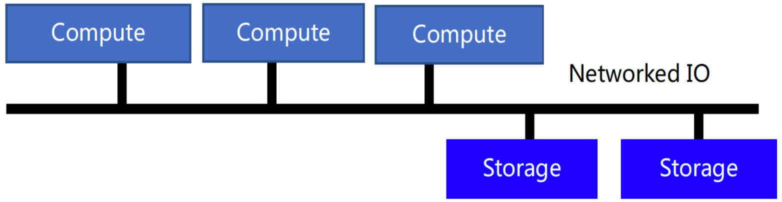
IETF-110 iccrg

March 2021

Cloud desires hyper-speed networking

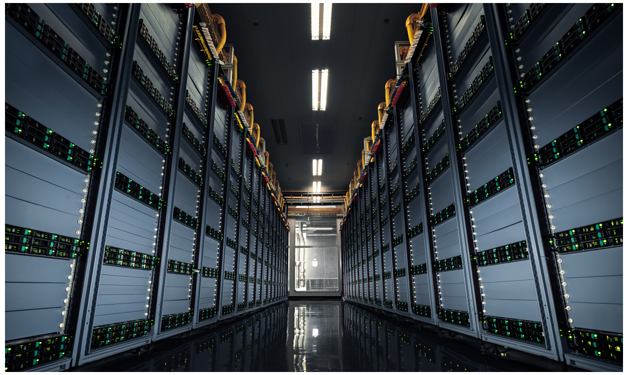
Today, clouds have
- bigger data to compute & store
- faster compute & storage devices
- more types of compute and storage resources

High-performance storage



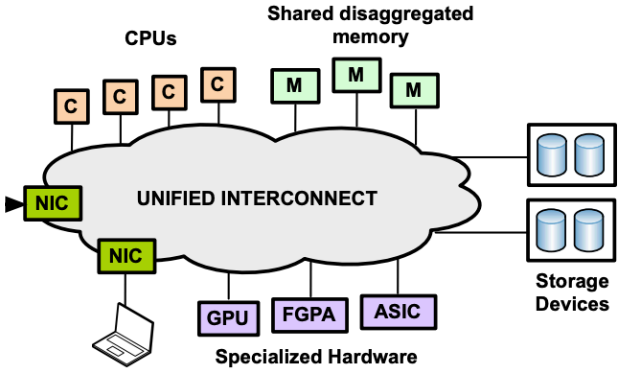
- Storage-compute separation is norm
- HDD→SSD→NVMe
- Higher-throughput, lower latency
- 1M IOPS / 50~100us

High-performance computation



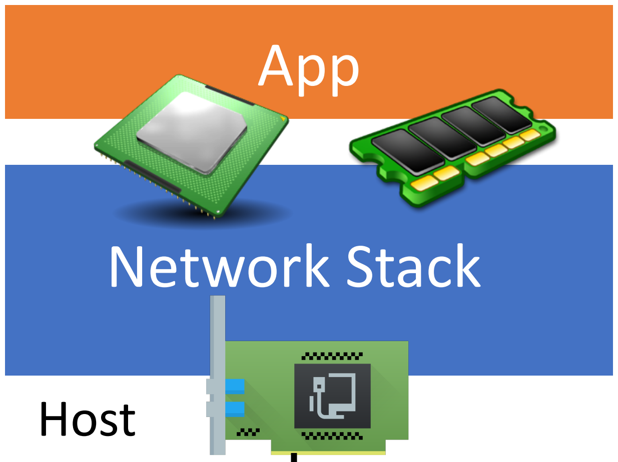
- Distributed deep learning, HPC
- CPU→GPU, FPGA, ASIC
- Faster compute, lower latency
- E.g. latency <10us

Resource disaggregation

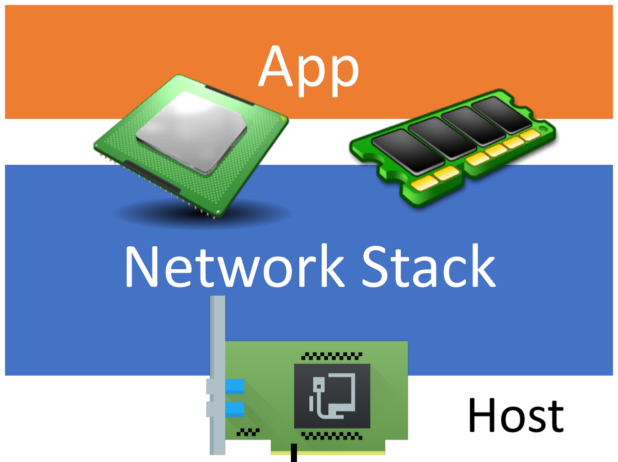


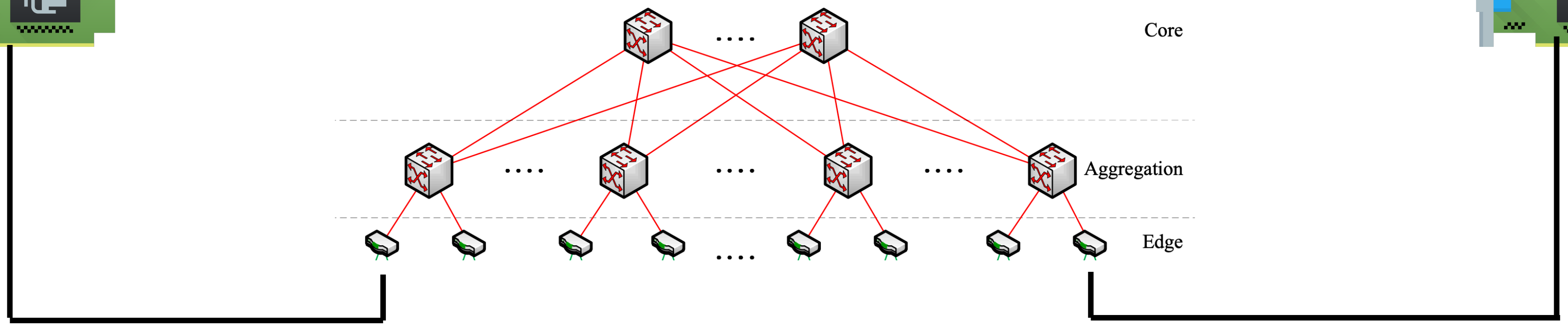
- More network load
- Need ultra-lower latency: 3-5us, > 40Gbps (Gao Et.al. OSDI'16)

Hyper-speed network chips != hyper-speed networking



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





Network Fabric

Congestion control (CC)
Since, end hosts are aggressive, network is more vulnerable to congestion & packet loss

Realistic challenges in current CC in RDMA networks

Challenges in current CC

- Operation challenge-1: PFC storm & deadlock
  - **Running lossy networks is desired, but there is a convergence challenge!!!**

Challenge-1: Slow Convergence

- Operation challenge-2: running multiple applications
  - **QoS queues are scarce resources!!!**

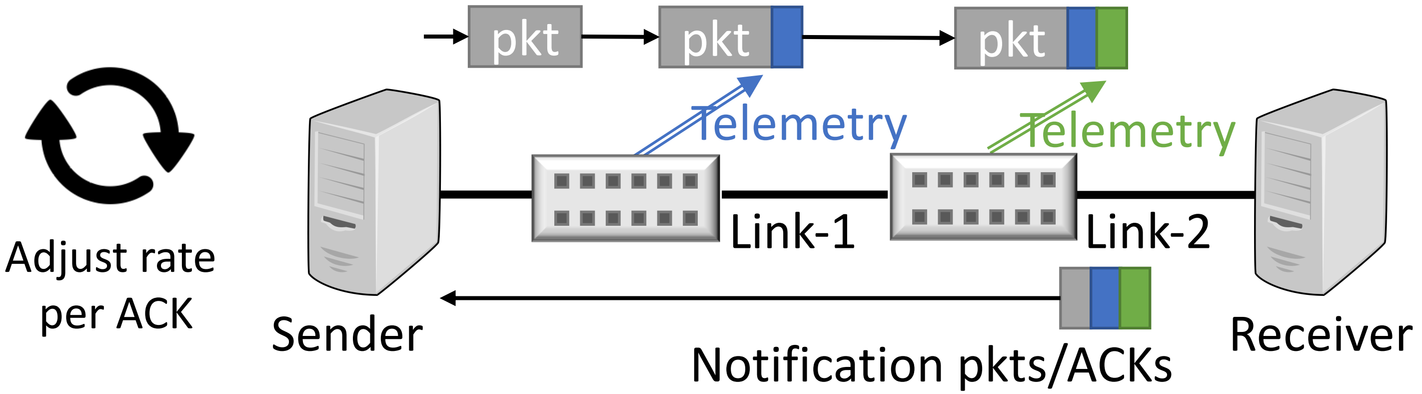
Challenge-2: Standing queue

- Operation challenge-3: complex parameter tuning
  - **DCQCN has at least 15 parameters to tune!!!**

Challenge-3: Heuristics in CC

HPCC++: Enhanced High Precision Congestion Control (SIGCOMM'19)

- New commodity ASICs have in-band telemetry ability
- Use in-band telemetry as precise feedback for congestion control



In-band telemetry vs ECN

- ECN = Explicit Congestion Notification, single-bit notification

- in-band telemetry provides fine-grained network load information
  - e.g., queue length, transmitted bytes, timestamp, link capacity, etc.
  - Quickly converge to proper rate to highly utilize bandwidth while avoiding congestion
  - Consistently maintain a close-to-zero queue for low latency

- Overhead of in-band telemetry (5-hop switches in DC paths)
  - Per-packet telemetry (w/ compact): up to 42B or 4.2% in a 1KB packet
  - Per-RTT probing packet (w/ IFA1.0): up to 200B or 2.5% for each 8KB data
  - In DC, bandwidth is generally abundant, but the latency is much more important

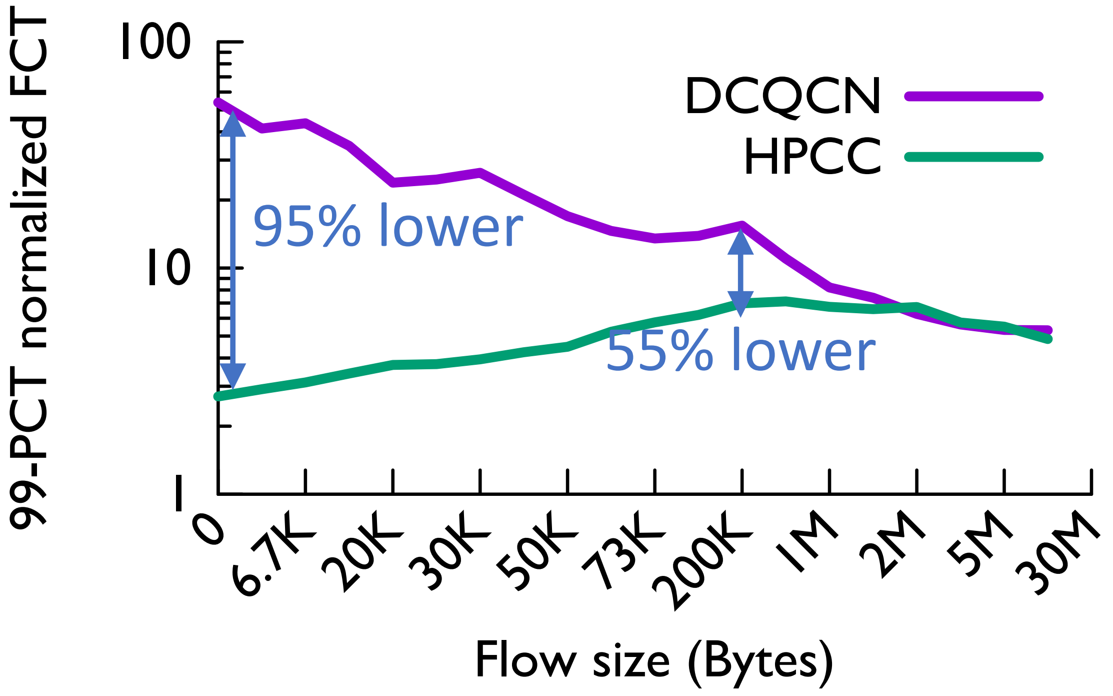
HPCC solves the 3 problems

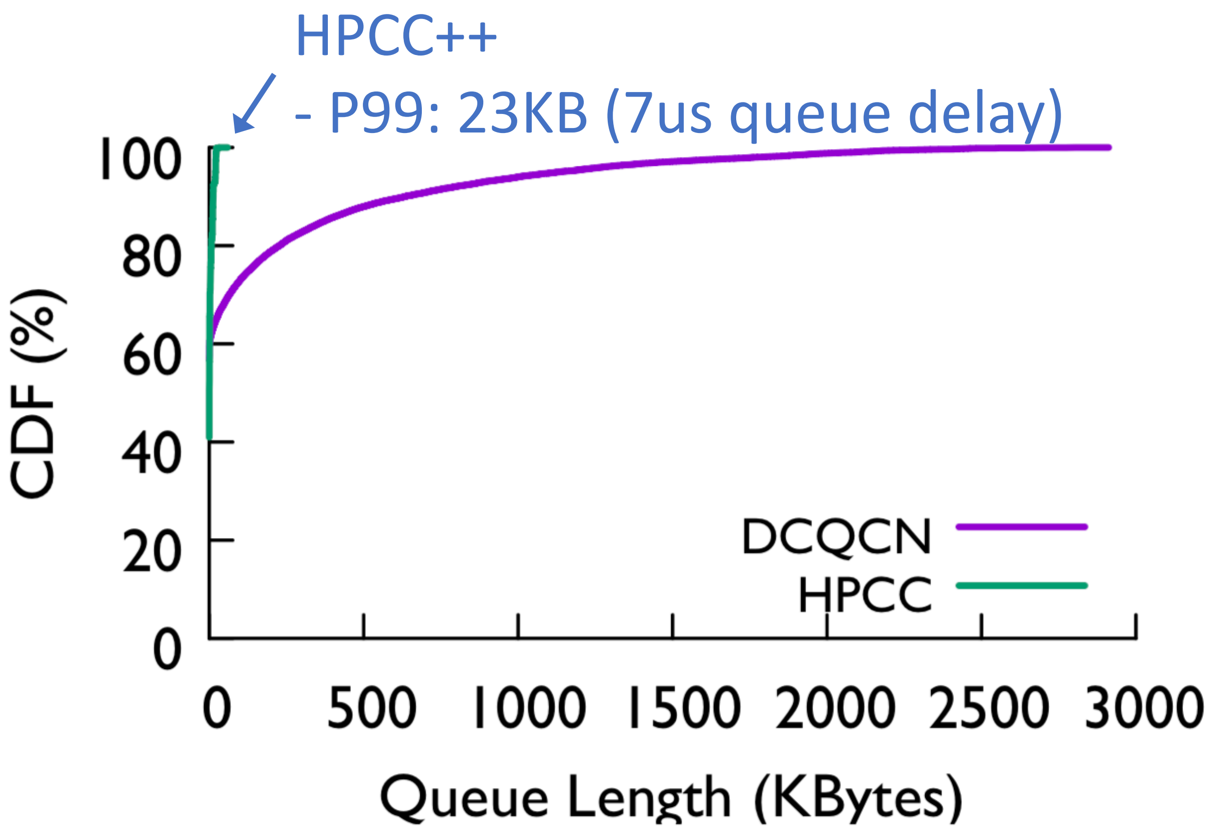
Using in-band telemetry as the precise feedback

- Fast convergence
  - ➢ Sender knows the precise rate to adjust to, on every ACK
- Near-zero queue
  - ➢ Feedback does not only rely on queue
- Few parameters
  - ➢ Precise feedback, so no need for heuristics which requires many parameters

HPCC++ achieves lower FCT and near-zero queue

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





Thank You