UC San Diego

CSE 291: Operating Systems in Datacenters

Amy Ousterhout

Oct. 19, 2023

UC San Diego

Agenda for Today

- Announcements + reminders
- Background on congestion control in datacenters
- Homa discussion
- Swift discussion
- Homa vs. Swift discussion

Reminders

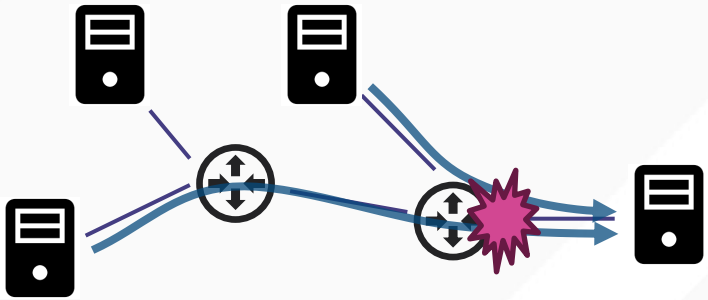
- Research project
  - Proposals due Monday 10/23 at 11:59 pm
  - Check ins next week – see Canvas to sign up
- Office hours today: 2:30-3:30 pm
- For Tuesday:
  - No need to review the “Killer Microseconds” paper
  - Do submit a review for Shenango

UC San Diego

Congestion Control in Datacenters

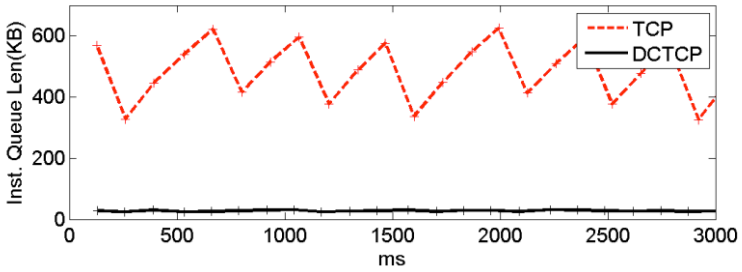
Network Congestion

- Traffic flows compete for use of the network
- Congestion control: determines which flow gets to send when and at what rate



TCP

- Congestion window (cwnd): number of bytes that can be outstanding at once
- TCP adjusts the cwnd based on additive increase/multiplicative decrease (AIMD)
  - No congestion: cwnd += 1
  - Congestion: cwnd = cwnd / 2
- Detect congestion when a packet is dropped

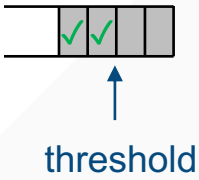


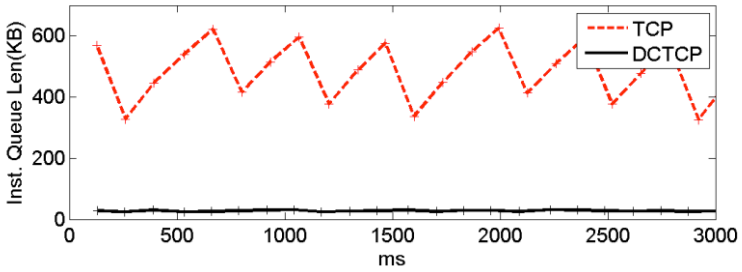
"saw-tooth" pattern

M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan. Data Center TCP (DCTCP). SIGCOMM '10.

DCTCP

- TCP does not work well in datacenters
  - Large “background” flows cause queueing in the network
  - Latency-sensitive “foreground” traffic suffers from high latency
- Particularly bad with partition/aggregate workloads
  - Applications need low tail latency (e.g., 99.9%)
- Goal: decrease the sending rate before the queues fill up
  - Mark packets when queueing exceeds a threshold





lots of queueing

little queueing

M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan. Data Center TCP (DCTCP). SIGCOMM ’10.

What is optimal?

- Goal: minimize the average time to send a message
- Optimal policy: shortest remaining processing time (SRPT)
  - Sends the message with the fewest bytes remaining first
- Challenges with SRPT
  - Need to know the message size
  - May starve long messages
  - Not necessarily optimal with multiple switches
- Many protocols approximate SRPT
  - pFabric, PIAS, pHost, Homa

Congestion Control in Datacenters

- Several different types
  - Endhost only: TCP, **Swift**
  - Switch feedback: ECN, DCTCP, XCP
  - Switch scheduling: pFabric, QJump, PDQ
  - Centralized: Fastpass
  - Receiver/credit-based: **Homa**, NDP, pHost, ExpressPass

UC San Diego

Homa Discussion
Swift Discussion