Recitation 11: DCTCP

MIT - 6.033
Spring 2022
Henry Corrigan-Gibbs

Plan

* Recitation Qs
* Background on DCs
* Queue game
* DCTCP

Logistics

* DPR due 3/18
* Volunteers?

Rectation Qs

1. What is the goal of DCTCP?
   * Better net utilization in DC

2. How does DCTCP differ from TCP?
   * Management of buffers

3. Why does DCTCP differ from TCP?
   * Different setting
     ↳ Control endpoints
     ↳ Very low RTT

Why is this a great paper?

- Simple, important (?) problem
- Elegant solution
  - ↳ No new hardware
  - ↳ Simple changes to endpoints
- Works in practice (e.g. Cisco)

Types of Flows in Data Center

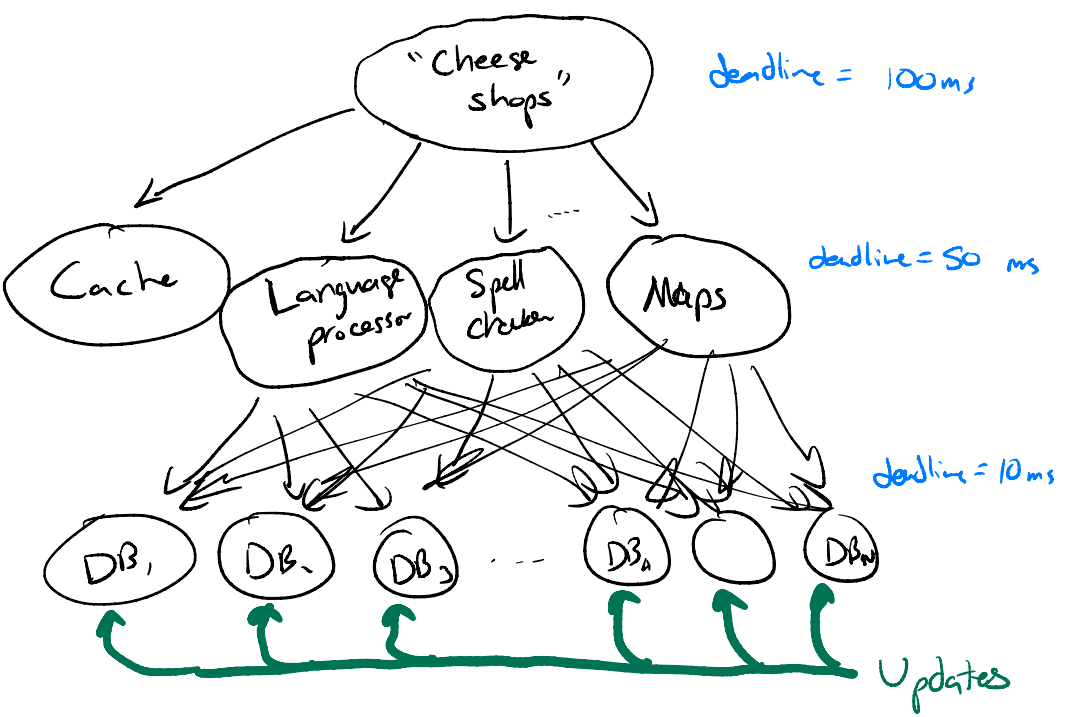
- Query traffic ~2KB
- Short message traffic ~1 MB

} Small, latency sensitive

- Background traffic ~50 MB

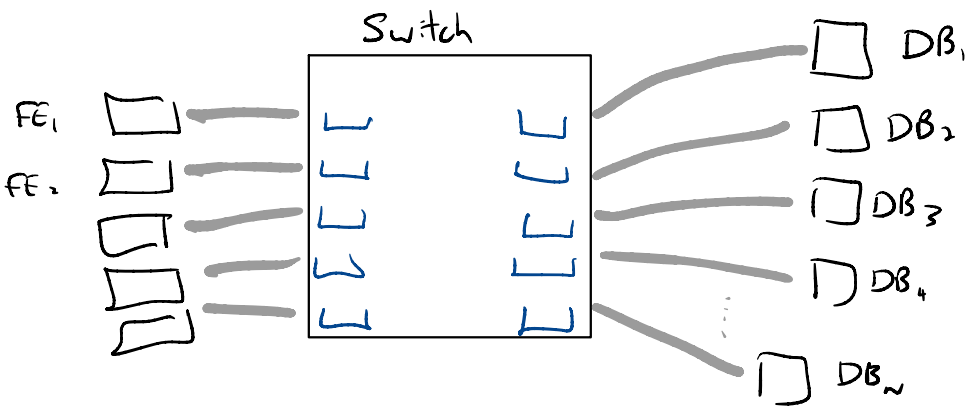
} big throughput is important

E.g. Search



Queues

Why do they exist?



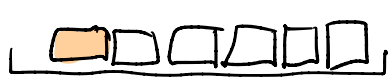
- If 10 FEs want to send to same DB at same time in-rate >> out rate → packets need to go somewhere
- Memory is $$$ → All ports share same queue buffer

→ What happens when queue fills up?

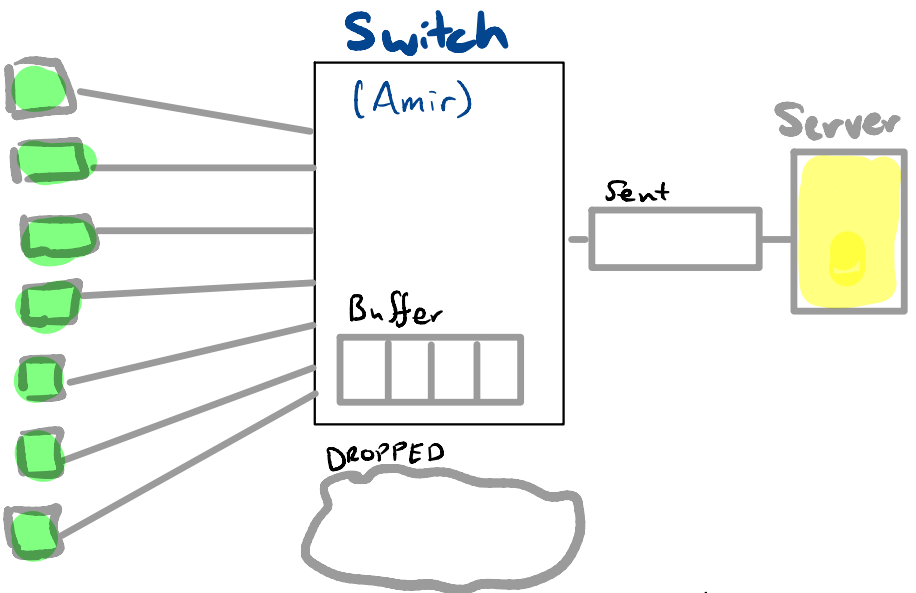
↳ Dropped packets = worse search results (connection times out)

→ Even before queue fills up, can be problematic

↳ Latency grows!



Queue Game



- Students: Write S/L in buffer (if space) or in DROPPED bin (if no space)
- Amir: Move packet from buffer to sent box

Trial run: One sender, every 3-5 seconds

↳ Show what this looks like in queue

For real now...

- Two flows, both send every 1-2 seconds

- One long flow, sending as quickly as possible
  - ↳ Some packets drop but throughput is good
  - ↳ How does TCP handle this? Try again.

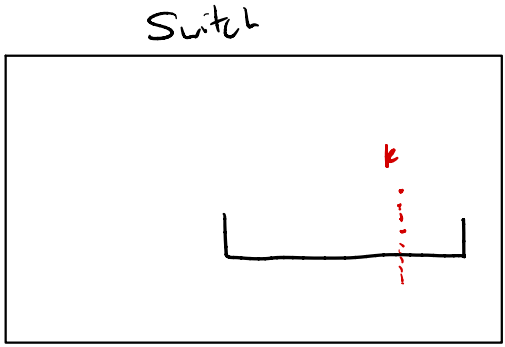
- One long flow, one short flow
  - ↳ Queue length grows ⇒ latency!
    "bufferbloat"

- Everyone sends at same time
  - ↳ Lots of packets drop ("incast")

→ What could we do to fix this problem?

→ What does DCTCP do to fix this problem?

DCTCP



Switch

As soon as queue grows beyond threshold, (small!) K switch sets flag (↳ ECN) on each packet

Normal TCP: Wait until queue is full

Sender

Uses ECN flags to estimate whether queue size is > K

↳ If so, gradually back off (adjust window size)

↳ If not, continue as normal

Normal TCP: Cut window in half

Why would this not work on Internet?

- Practical: Need to modify both ends of the connection.
- Convergence time depends on RTT
  - ↳ Small in DC (0.1 ms)
  - ↳ Big in Internet (50 ms)

Time for new flow to get its fair share of bandwidth.

3-4x slower than TCP

- Feedback is too slow... by the time sender gets CN, queue may be empty.
  - ↳ Instantaneous queue length not a good signal (e.g. traffic on road 8 hrs away)

- Less clarity about types of flows (?)