today's RESEARCH is tomorrow's REALITY

Microsoft® Research TechFest 2010 the & in R&D

Cloud Faster: milliseconds matter

Albert Greenberg, Cheng Huang, Dave Maltz, Jitu Padhye, Parveen Patel, Murari Sridharan

Data Center Transport

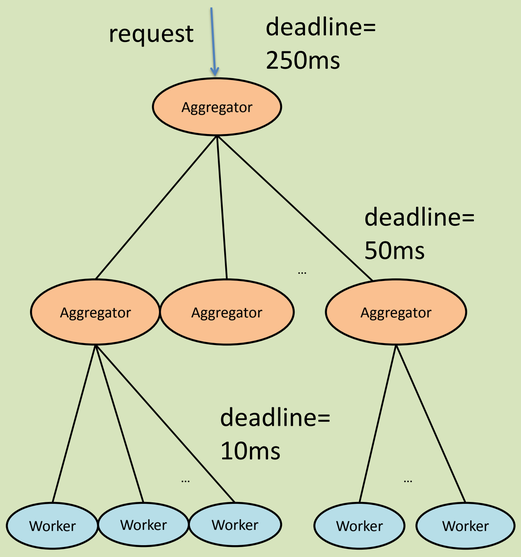
Goals

1. High Throughput
- ➢Continuous data update

2. High Burst Tolerance
- ➢The Partition/Aggregate pattern is common

3. Low Latency (milliseconds matter)
- ➢Soft real time app's run close to SLA's
- ➢Indeed, reduce network latency ➔ more time for the algorithms, and for better results

- Partition/Aggregate Pattern



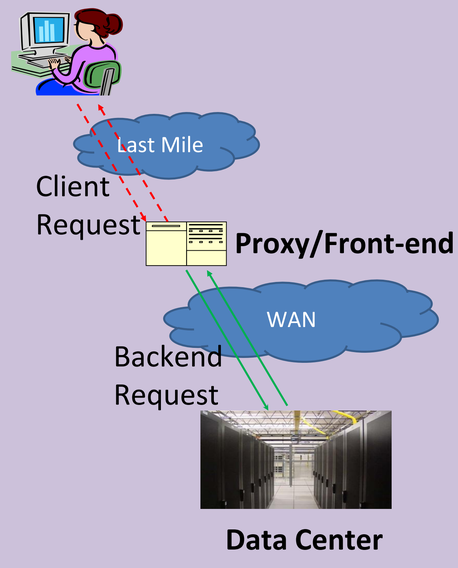
Wide Area Transport

Goals

1. Minimize connection setup time
- ➢Loss of initial packets leads to very long delays

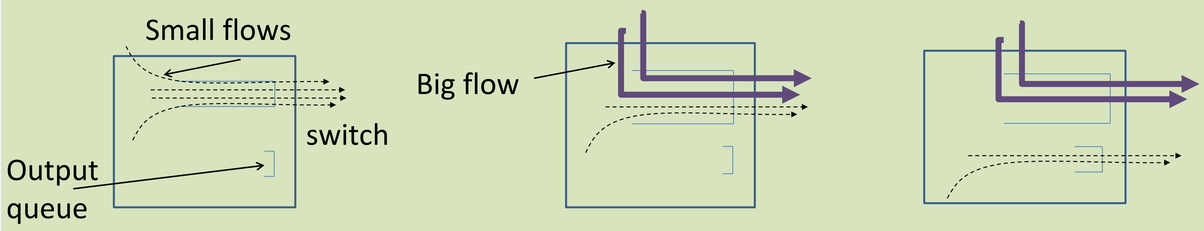
2. Minimize transfer delays
- ➢Short transactions take too long to ramp up
- ➢E.g., a search query result is only 17KB yet takes 4 RTTs

3. Faster loss recovery for clients
- ➢Clients experience high losses at the last mile
- ➢Recovery takes too long to complete



What Causes Problems?

Flow interaction in shared memory switches cause packet loss and delay



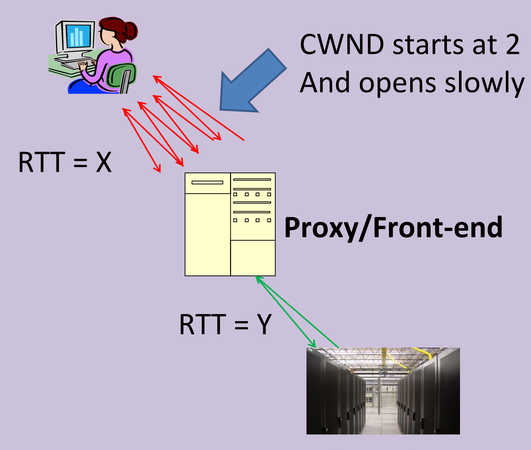
a) Incast
- ➢ Lots of small flows cause packet loss
- ➢ Aggregate in Partition/Aggregate

b) Queue Buildup
- ➢ Big flows build long queues
- ➢ Mix of low latency and big flows

c) Memory Pressure
- ➢ Big flows take away all the buffer in the switch

What Causes Problems?

- **Slow ramp up even in best case**
  - ➢Total delay: n* X + Y
  - ➢High overhead for short transactions
- **Very long latencies if packets lost**
  - ➢If SYN or SYN-ACK is lost
    - ➢3 second timeout
  - ➢If packet is lost, timeout is likely
    - ➢Since window is small
    - ➢Default minimum timeout is 200ms
    - ➢Even if RTT to proxy is just 10ms!



Solution: Data Center TCP

- Switch side
  - ➢Mark packets when queue occupancy exceeds a small threshold K.

Example responses

ECN Marks	DCTCP	TCP
1 0 1 1 1 1 0 1 1 1	Cut window by 40%	Cut window by 50%
0 0 0 0 0 0 0 0 0 1	Cut window by 5%	Cut window by 50%

- Source side
  - ➢ Estimate the *fraction* of packets marked (α) as a measure of congestion

Let M be the fraction of ACKs with ECN bit = 1 in each RTT, then

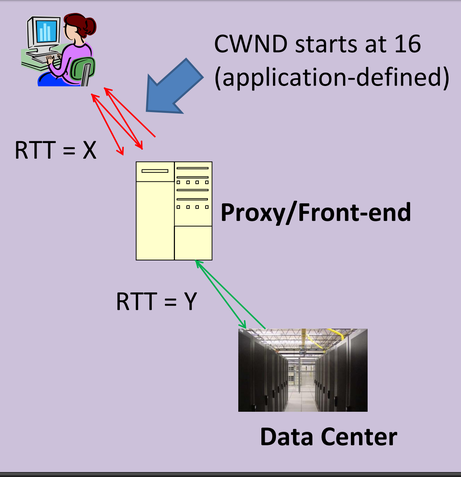
$$\alpha \leftarrow (1-g)\times\alpha + g\times M$$

- ➢Window decreases are adaptive and proportional

$$Cwnd \leftarrow Cwnd \times (1-\alpha/2)$$

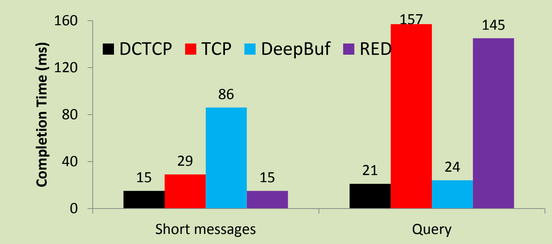
Solution: Wide Area TCP Optimizations

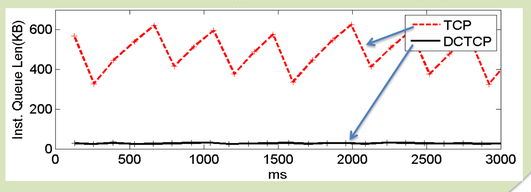
- **Quick ramp up**
  - ➢Increase ICW
  - ➢Delay drops to 2 * X + Y
- **Quick loss repair and FEC**
  - ➢Avoid loss penalties by duplicating small critical packets
    - ➢Proactively retransmit SYN-ACK three times
  - ➢Recover faster from losses
    - ➢Reduce MinRTO to 100ms
    - ➢Reduce Initial RTO to 500ms



DCTCP Achieves All Three Goals

1. **High throughput**
   - ➢Creating multi-bit feedback at TCP sources
2. **Low Latency (milliseconds matter)**
   - ➢Small buffer occupancies due to early and aggressive ECN marking
3. **Burst tolerance**
   - ➢Sources react before packets are dropped
   - ➢Large buffer headroom for bursts



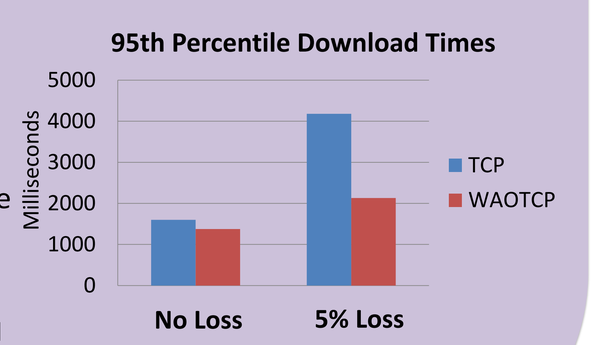


Wide Area TCP Achieves All Three Goals

1. **Faster connection setup**
2. **Lower transfer delays**
3. **Faster loss recovery**

Web apps are built around short messages
Reducing their latency improves user experience

- **Complementary Work**
  - ➢Google's SPDY protocol minimizes HTTP overhead
  - ➢Wide Area TCP minimizes network transfer time – benefits all applications



Microsoft Research

Microsoft