Datacenter Network Congestion
Synopsis, Causes & (potential) Cures

Lawrence Stewart
lastewart@swin.edu.au
Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

Outline

1 Datacenter Networks
   • Why are they interesting?

2 Datacenter Congestion
   • What causes it?
   • What is a microburst?
   • What is incast?
   • What problems are associated with incast?
   • How do we address microbursts & incast?
Why are datacenter networks interesting?

- Scale
- Architecture (topology, N-tier)
- High bandwidth, low latency
- Hardware, software & protocol mix
- Traffic mix (background bulk, priority user)
- Business requirements (response latency)
- Single administrative domain

More on scale & architecture

http://www.juniper.net/techpubs/images/g041164.gif
More on hardware, software & protocol mix

- Typically commodity based
- Protocol stack: Ethernet + IP + TCP
- x86 servers connected at 1 or 10Gbps
- Standard operating systems
- Standard Ethernet switching and IP routing

What causes congestion in the datacenter?

- Interaction between clustered workloads, network protocol behaviour & hardware
The TCP feedback control loop

A (In)Cast of Thousands

- First articulated by clustered storage vendor Panasas in 2004
- Triggered by microbursts

http://www.pdl.cmu.edu/Incast/
Bad for business

- Impedance mismatch between TCP and network latency means RTOs are a disaster
- Efficiency of network and machine cycles drops
- Response time increases

Mitigating microbursts & incast

Changes to TCP:
- Fine grained timers
- Tweak congestion control
- Multipath

Changes elsewhere:
- Datacenter bridging (PFC)
More on congestion control tweaks

- **ECN refresher:**

![ECN diagram](http://www.potaroo.net/papers/ipj/2000-v3-n2-tcp-perf/figure11.gif)

- **Datacenter TCP (DCTCP):**
  - **Server-side**
  - Sender infers amount of congestion from rate of ECN marks
  - Adjusts congestion window proportionally to rate

- **Incast Congestion Control for TCP (ICTCP):**
  - **Receiver-side**
  - Uses flow fate sharing information to adjust receive window

Watch this space for CAIA’s contribution to the area...