# Data Center Networks

Arbob Ahmad

Computer Science Department
Carnegie Mellon University

15-744 Paper Presentation
October 22, 2010

# Questions

Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication

1. The authors suggest that, rather than developing a new datacenter-specific transport protocol, TCP should be modified because TCP is a mature and well-understood protocol. To what extent do you think this is justified?
2. In analyzing the impact of spurious timeouts, the authors measured bulk-data TCP transfers. How would this analysis differ if the authors had measured small tranfers (e.g., web page fetches)? Are the authors justified in focusing on bulk-data transfers?
3. Hardware interrupts for the hrtimers could cause overhead if timeouts are frequent. The authors leave this as future work, but argue that a small overhead may be acceptable if the alternative is an idle period for the server. Are there situations where the overhead may not be preferable?

# Question 1: TCP advantages

- TCP is a de facto standard
- TCP is well-understood so that the effects of small changes to TCP are more easily studied than an entirely new protocol
- TCP modifications are minor
- Good implementations of TCP already exist
- Hardware support for TCP datacenters
- TCP is adaptable. It may remain effective as datacenters change

# Question 1: New Protocol Advantages

- New protocol could be optimized for the specific features of a high throughput network
- Datacenter isolation makes it a good candidate for new protocols

# Question 2: Workload

The workload used in this paper differs from that in previous work by the same authors

- This workload reads a fixed size data block striped across N servers
- The previous workload had a fixed fragment size read from each server

Related work analyzes this choice of workload

- The workload is chosen because it is representative of communication patterns in popular distributed storage systems
- The earlier workload may be more representative of communication patterns in other applications involving bulk data transfers
- A study of datacenter traffic did not observe the incast collapse found in this paper even though the factors for this to occur existed in their experiments and these experiments are representative of a wide variety of datacenter loads

# Question 2: Workload Discussion

- Small transfers such as web page fetches were probably not studied because the load profile of a web server is significantly different than that of a file or database server
- Bulk transfers are studied because these are most likely to cause incast collapse
- Results depend on workload so testing with different workloads including the one from previous work would have been helpful in analyzing the method

# Question 3: Timer overhead

- There has been subsequent work on reducing the overhead of hrtimers such as "Analysis of High Resolution Timer Latency Using Kernel Analysis System in Embedded Systems" by Kwon, et al.
- The timer overhead seems justified by the empirical results in the paper, but the extensive evaluation left to future work would be informative

# Additional Questions

Can we use ECN to overcome these problems?

- Earlier work by the same authors found ECN and other existing TCP improvements did not substantially change the incast-induced throughput collapse.

The methods described in this paper seem to involve a lot of trial-and-error and hacking. Are there any cleaner solutions to this problem?