Introduction to Data Center Networks

Mohit P. Tahiliani

Assistant Professor

Department of Computer Science and Engineering

National Institute of Technology Karnataka, Surathkal, India

tahiliani@nitk.edu.in

Overview

- Privately managed networks with full control on deployment
- Thousands of servers are co-located in a same geographical area
  - Data Centers have thousands of servers
  - One data center can be as big as three football grounds
  - Heat dissipation is a problem! Underwater data centers?
    - Let's go to Antarctica!
- Server to server communication is more frequent in Data Centers
- Switches are more common than routers
  - Buffers are shallow
- Round Trip Times are extremely small (in the order of microseconds)
- Different types of traffic: mice, cat and elephant

Different types of traffic and their requirements

DATA CENTER TRAFFIC: APPLICATIONS AND PERFORMANCE REQUIREMENTS

Traffic Type	Examples	Requirements
Mice traffic (< 100KB)	Google Search, Facebook	Short response times
Cat traffic (100KB-5MB)	Picasa, YouTube, Facebook photos	Low latency
Elephant traffic (> 5MB)	Software updates, Video On-demand	High throughput

- Mice traffic: small flows with bursty nature
- Cat traffic: medium sized flows with occasionally bursty nature
- Elephant traffic: long lasting flows

Challenges for TCP in Data Center Networks

- Fundamental changes in TCP
  - RTO cannot be in the order of milliseconds because the overall RTT in data centers is in hundreds of microseconds.
- Major challenges for TCP in Data Center Networks
  - TCP Incast
  - TCP Outcast
  - Queue buildup
  - Buffer pressure
  - Pseudo congestion effect

TCP Incast

- Partition/Aggregate application structure is commonly used in data center networks
  - Main use cases: search traffic, advertisement mining, etc

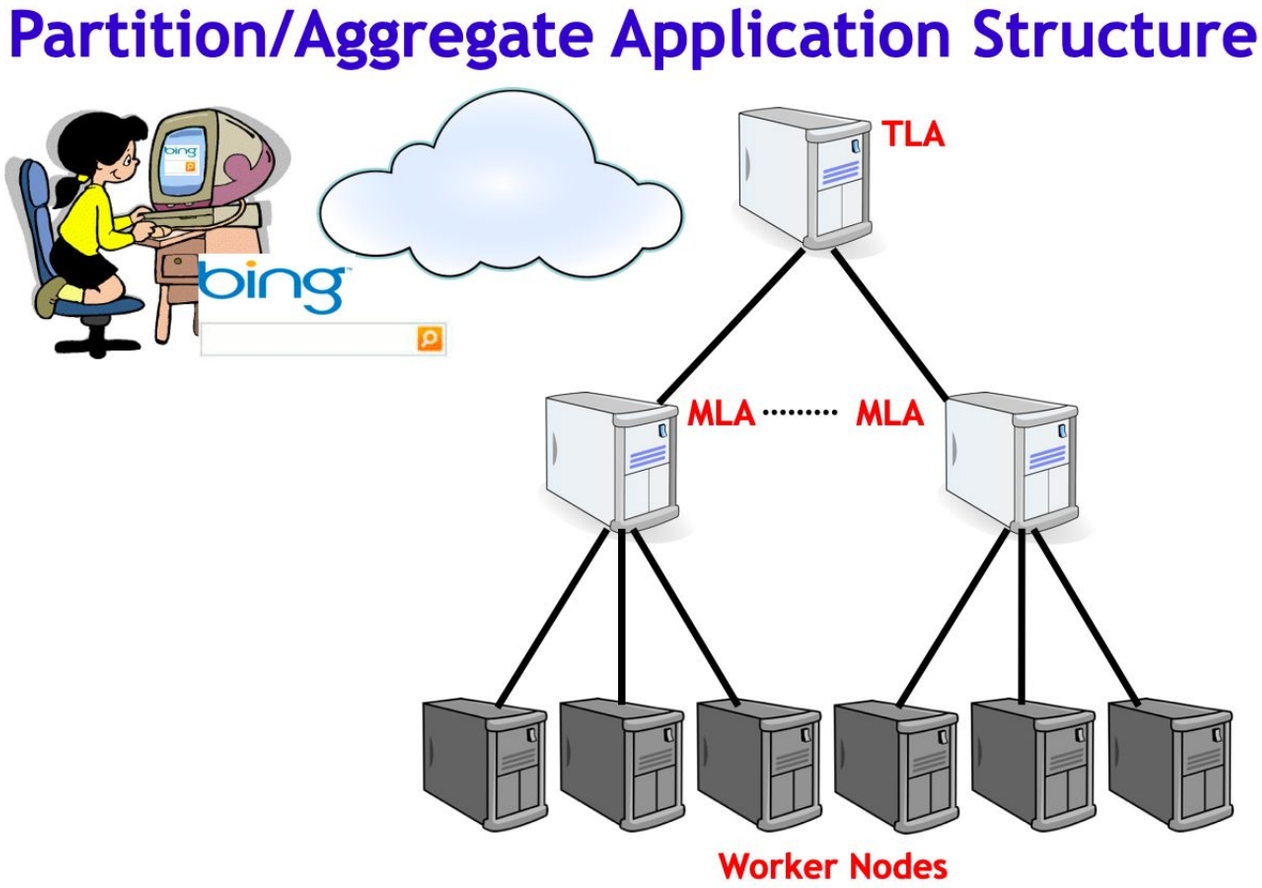

Image Credits: Original presentation slides from DCTCP authors

TCP Incast (contd ...)

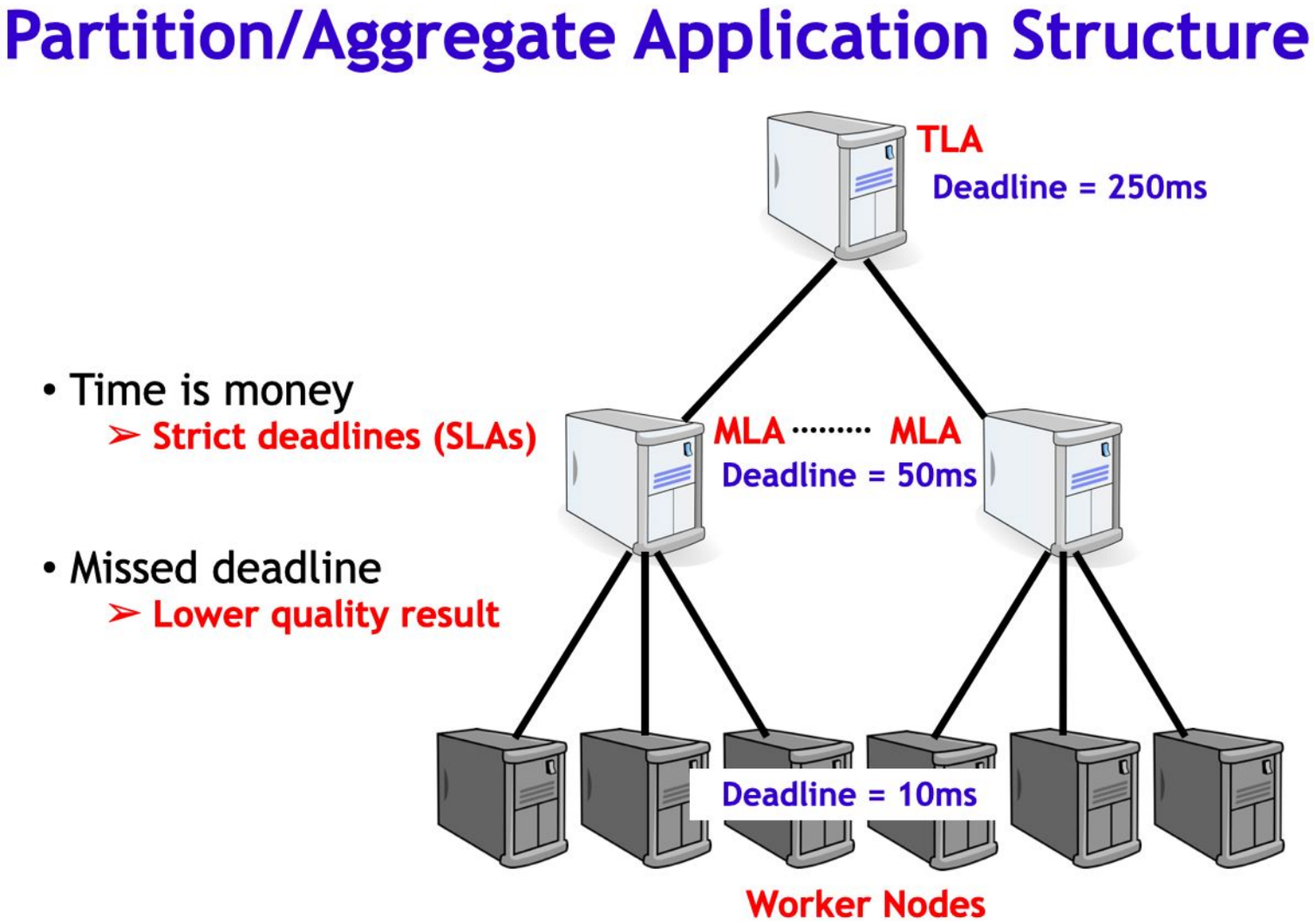

Image Credits: Original presentation slides from DCTCP authors

TCP Outcast

- Two sets of flows: mice and elephant competing for the same output port
- Elephant flows lead to port blackout for the mice flows
- It occurs in data centers that employ switches with droptail queue discipline

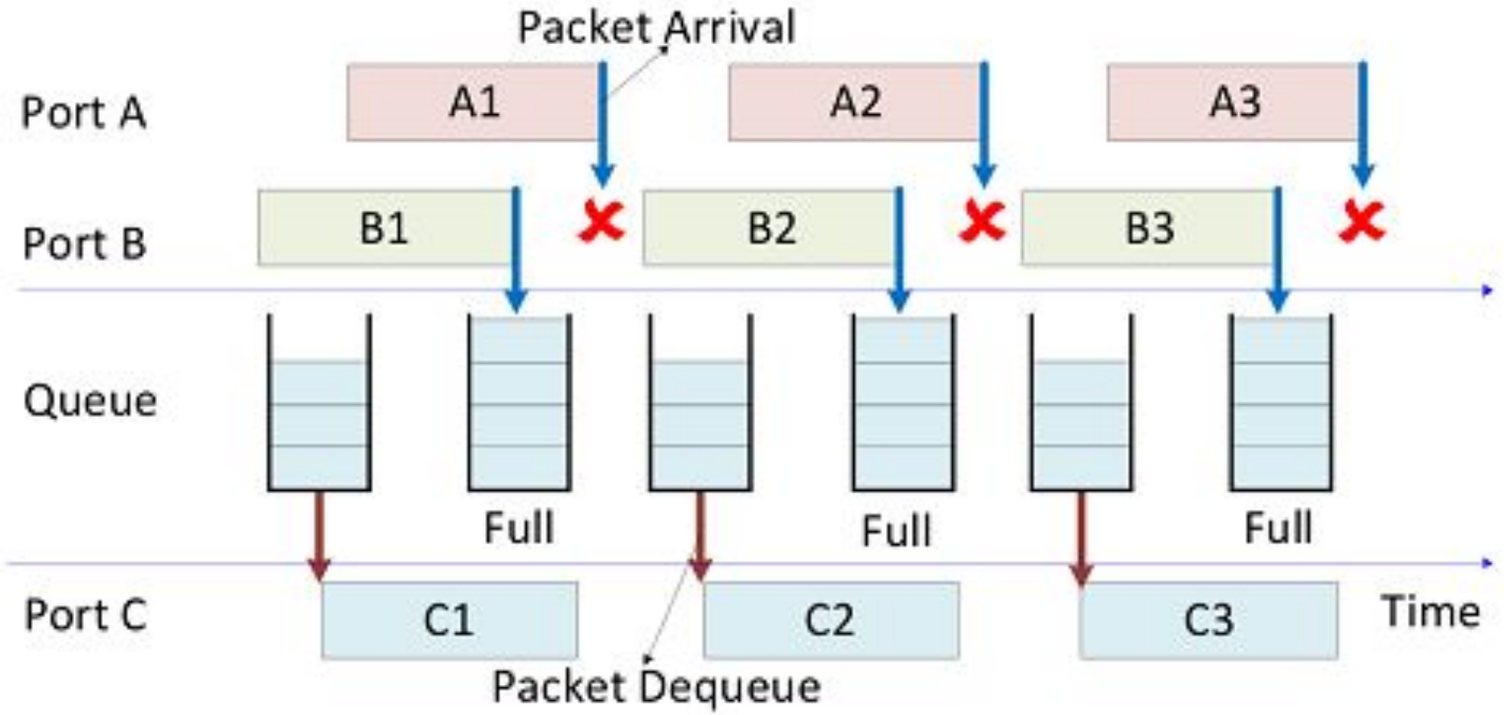

Image Credits: Prakash, P., Dixit, A., Hu, Y.C. and Kompella, R., 2012, April. The TCP Outcast Problem: Exposing Unfairness in Data Center Networks. In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (pp. 30-30).

Queue buildup and Buffer pressure

- Elephant flows keep the buffers full; hence there is no room to accommodate bursts
- Mice flows are either dropped or encounter high queuing delays

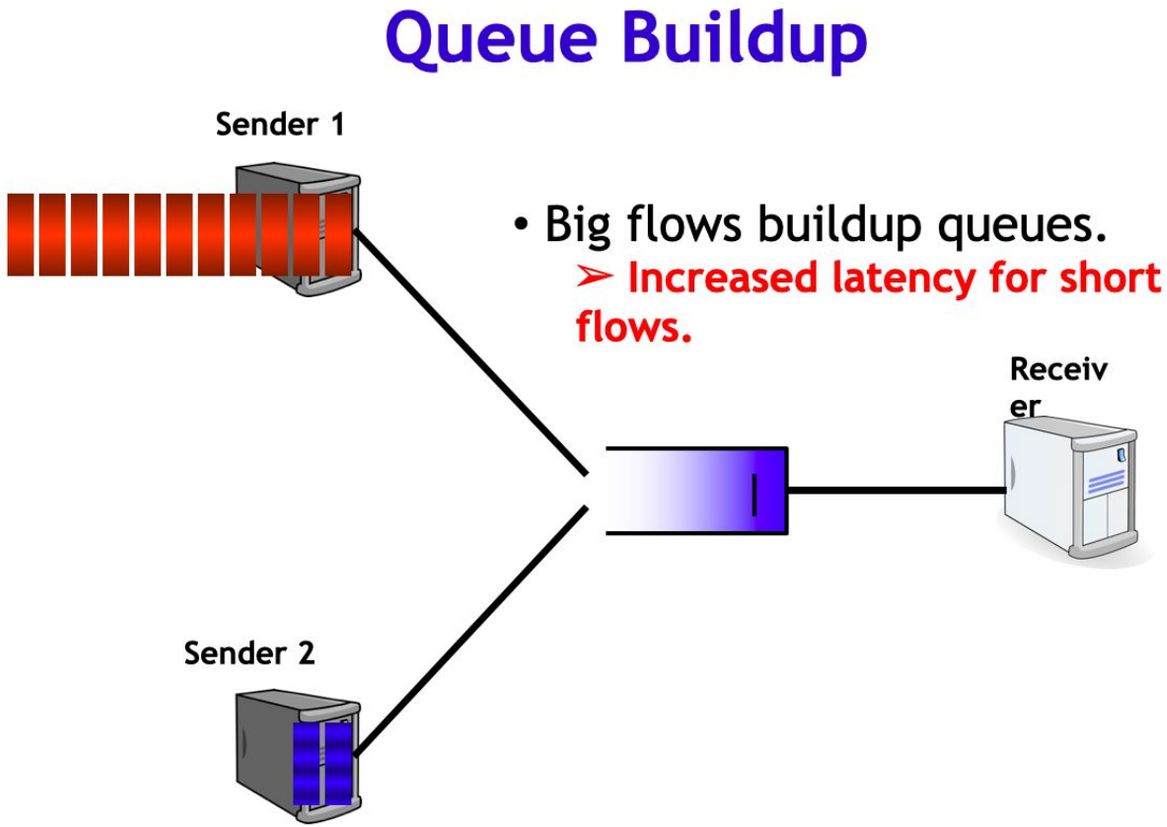

Image Credits: Original presentation slides from DCTCP authors

Pseudo congestion effect

- Virtualization is a key technology for cloud computing
- Hypervisors are crucial to schedule several VMs to gain access to physical resources
- Hypervisor scheduling latency can be very high sometimes!
  - Starvation of a VM
  - Long wait times for a VM
- Due to high scheduling latency, a packet may not get processed soon enough
  - RTO expires
  - TCP sender considers this as an indication of high congestion and resets the cwnd
- This is called pseudo congestion effect, because there is no ‘actual’ congestion

Summary

TCP Impairment	Causes
TCP Incast	Shallow buffers in switches and Bursts of mice traffic resulting from many-to-one communication pattern.
TCP Outcast	Usage of tail-drop mechanism in switches.
Queue Buildup	Persistently full queues in switches due to elephant traffic.
Buffer Pressure	Persistently full queues in switches due to elephant traffic and Bursty nature of mice traffic.
Pseudo-Congestion Effect	Hypervisor scheduling latency.

Proposed Solutions (updated till 2014-15)

TCP Variants proposed for Data Center Networks	Modifies Sender	Modifies Receiver	Modifies Switch	Solves TCP Incast	Solves TCP Outcast	Solves Queue build-up	Solves Buffer pressure	Is Deadline Aware	Detects pseudocongestion	Implementation
TCP with FG-RTO	√	x	x	√	x	x	x	x	x	Testbed and ns-2
TCP with FG-RTO + Delayed ACKs disabled	√	x	x	√	x	x	x	x	x	Testbed and ns-2
DCTCP	√	√	√	√	x	√	√	x	x	Testbed and ns-2
ICTCP	x	√	x	√	x	x	x	x	x	Testbed
IA-TCP	x	√	x	√	x	x	x	x	x	ns-2
D2TCP	√	√	√	√	x	√	√	√	x	Testbed and ns-3
TCP-FITDC	√	√	√	√	x	√	√	x	x	Modeling and ns-2
TDCTCP	√	√	√	√	x	√	√	x	x	OMNeT++
TCP with GIP	x	√	x	√	x	x	x	x	x	Testbed and ns-2
PVTCP	√	√	x	√	x	x	x	x	√	Testbed

Recommended Reading

Tahiliani, R. P., Tahiliani, M. P. and Sekaran, K. C., 2012, December. TCP Variants for Data Center Networks: A Comparative Study. In 2012 International Symposium on Cloud and Services Computing (pp. 57-62). IEEE.

Tahiliani, R. P., Tahiliani, M. P. and Sekaran, K. C., 2015. TCP Congestion Control in Data Center Networks. In Handbook on Data Centers (pp. 485-505). Springer, New York, NY.