Exploring High Speed TCP Incast Congestion Issues In Large Data Centres

Jonathan Kua
jkua@swin.edu.au

Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

Overview

- Data Centres and Incast/Microburst
- Switch Model with Virtual Output Queuing (VOQ)
- CAIA's NS-3/FreeBSD Simulator
- Results and Analysis
- M/D/1 Queuing Model
- Possible Solutions
- Further Work
What Makes Google

- 13 Data Centres, 900,000 servers
- In 2013, there are 2,161,530,000,000 Google search queries, 5,922,000,000 per day, indexing 20 billion web pages per day
- Free gmail storage to 425 million users
- Serve 6 billion YouTube videos per month
- How does Google handle these workloads?
Incast and Microburst

- “You can't pour two buckets of manure into one bucket” - Scott Fritchie's grandfather
- Network pathology that affects many-to-one communication pattern
  - High speed and low latency data centres
- Causes: Response synchronization and TCP timeouts
- Effect: Catastrophic throughput collapse
- Cluster storage, websearch, MapReduce

Incast Topology

- Divide and Conquer
Virtual Output Queuing (VOQ)

Traditional Input Queue Switch Model
Head of Line (HoL) Blocking

Switch Model with VOQ

Switch Model
CAIA's Simulator (NS-3/NSC)

- CAIA's NS-3/FreeBSD Network Stack in Virtual Machine (Version 0.1)
- Released on Sep 26, 2013
- Network Simulation Cradle (NSC)
  - Uses FreeBSD-9 Stable TCP stack
  - TCP New Reno SACK
- Long live TCP connection
  - Set initial cwnd to infinite
- Backpressure and backplane latency

Experimental Work: Testbed

- Experiment Parameters

![Diagram showing network components and connections with FreeBSD TCP stack on all nodes]
Results and Analysis

- Vary Number of Responders
- Vary Response Size to RX-Q Ratio
- Measurements:
  - Query Response Completion Time
  - Goodput
  - Peak Queue Occupancy and Packet Losses
  - RX-Q Drop and Queue Occupancy Correlation

Completion Time

Query and Response Completion Time vs Number of Responders
(Response Size = 1RX-Q, RX-Q = 256Cells x 384 Bytes)

Packet Loss starts to occur at 15 responders
Completion Time Distribution

Query Response Completion Time vs Number of Responders Boxplot
(Response Size = 1RX–Q, RX–Q = 256 Cells x 384 Bytes)

Goodput

Goodput vs Number of Responders
(Response Size = 1RX–Q, RX–Q = 256 Cells x 384 Bytes)
Peak Queue and Packet Loss

Peak Queue Occupancy vs Number of Responders
(Response Size = 1RX-Q, RX-Q = 256 Cells x 384 Bytes)

Total Packet Loss vs Number of Responders
(Response Size = 1RX-Q, RX-Q = 256 Cells x 384 Bytes)

Packet Loss starts to occur

A Deeper Look

- 15 responders

RX–Q Queue Occupancy Correlation Graph Between Individual Responders

Receive Queue Correlation Graph Between Individual Responders

Only 2 responders drop packets
A Deeper Look

- 50 responders

Receive Queue Correlation Graph Between Individual Responders

- 57 responders

Receive Queue Correlation Graph Between Individual Responders

Not all responders drop packets
A Deeper Look

- 60 responders

![RX-Q Queue Occupancy Correlation Graph Between Individual Responders](image1)

Packet Drop are slightly uncorrelated in time

![Receive Queue Correlation Graph Between Individual Responders](image2)

Completion Time

Query and Response Completion Time vs Response Size

(20 Responders, RX-Q = 48 Cells x 384 Bytes)

![Query and Response Completion Time vs Response Size Graph](image3)

Multiple retransmissions

Jump in completion time at 0.95 RX-Q
Goodput

Query and Response Goodput vs Response Size
(20 Responders RX-Q = 48 Cells x 384 Bytes)

Throughput Collapse at 0.95 RX-Q

Peak Queue and Packet Loss

Peak Queue Occupancy vs Response Size
(20 Responders RX-Q = 48 Cells x 384 Bytes)

All RX-Qs are filled

Total Packet Loss vs Response Sizes
(20 Responders RX-Q = 48 Cells x 384 Bytes)

Packet Loss starts to occur at 0.95 RX-Q
A Deeper Look

- Response Size = 0.95 RX-Q

Receive Queue Correlation Graph Between Individual Responders
(20 Responders, Response Size = 0.95 RX-Q)

Packet Drop are slightly uncorrelated in time

Putting It All Together

Peak Queue Occupancy vs Response Size for Different Number of Responders
(RX-Q = 48 Cells x 384 Bytes)
Putting It All Together

Total Packet Loss vs Response Sizes for Different Number of Responders (RX-Q = 48 Cells x 384 Bytes)

Query Response Completion Time vs Response Size for Different Number of Responders (RX-Q = 48 Cells x 384 Bytes)
M/D/1 Queuing Model

- Theoretical analysis of system under load
  - Continuous queries (Poisson distribution) from the querier
    \[ T = \frac{1}{\mu} + \frac{\rho}{(2\rho(1-\rho))}; \rho = \frac{\lambda}{\mu} \]
    - \( \mu \) is the completion rate
    - \( \lambda \) is the arrival rate
  - Arrival rate < Service rate

Assume arrival rate is 5 queries/sec

MD1 Model Time vs Response Size
(Response Size = 1RXQ, RX=Q = 48 Cells x 384 Bytes)
Possible Solutions

- Congestion control and fine grain OS timers
  - Reducing the minimum value of RTO, facilitate submilisecond RTO values, randomizing RTO timers, disabling TCP delayed ACKs.
- Increase size of switch buffers
- Application level scheduling of requests
- DCTCP – Server adjust cwnd
- ICTCP – Receiver adjust receive window

Further Work

- Fix software bugs in the simulator
- Propose solutions for TCP Incast
  - Reactive: Faster control loops
  - Proactive: De-correlation and dithering
- Understand DCTCP and ICTCP implementations more thoroughly
- More accurate goodput measurements
- M/D/1 model verification
- Experiment reconciliation
Conclusion

- Incast is a real phenomenon in large data centres (not solved yet)
  - Synchronization and TCP retransmission timeouts
  - Response size, switch buffer size, number of responders
- Switch model and CAIA's simulator
- Results and analysis
  - High correlation and synchronization
- Possible solutions

Thoughts on Internship

- Great exposure to TCP Incast research
  - Steep learning curve
- Developed technical skills and soft skills
  - PCBSD, NS-3/NSC Simulator, R, Lyx
  - Analytical thinking and problem solving skills – Handling large amounts of data
  - Articulating and formulating ideas
  - Patience and discipline
- Absolutely invaluable and rewarding learning experience
A Big Thank You

I would like to express my heartfelt appreciation for all CAIA staff who provide support throughout my internship.

Thank you Prof. Grenville Armitage for your mentorship and giving me this internship opportunity.

Thank you Dr Philip Branch and Lawrence Stewart for your supervision, incredible patience and ongoing support.

References

- CAIA Incast Congestion Control Project  
  http://caia.swin.edu.au/urp/incast/

- CAIA's NS-3/FreeBSD Network Stack in Virtual Machine  
  (Version 0.1) http://caia.swin.edu.au/urp/incast/tools.html


References


Thank You
Questions & Answers

Questions?