SWINBURNE UNIVERSITY OF TECHNOLOGY

Exploring High Speed TCP Incast Congestion Issues In Large Data Centres

Jonathan Kua

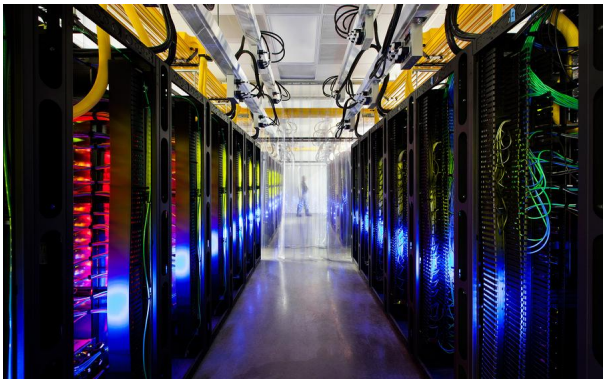
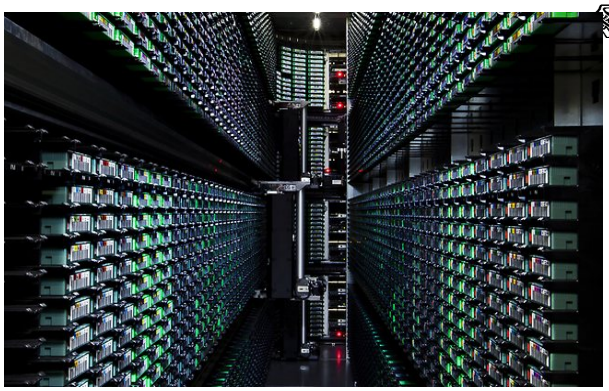
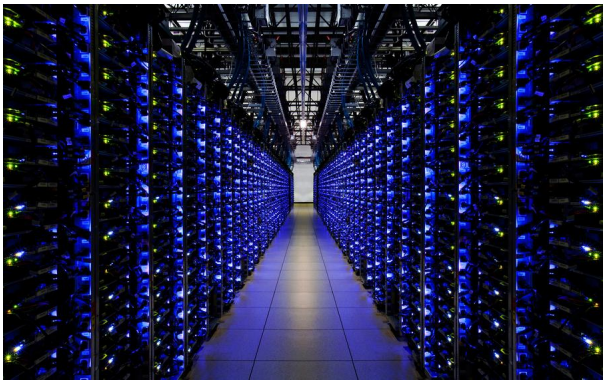
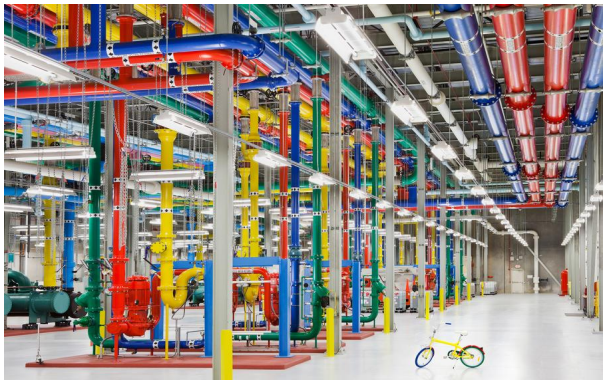
jkua@swin.edu.au

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

Where “The Internet” Lives

What Makes Google Google

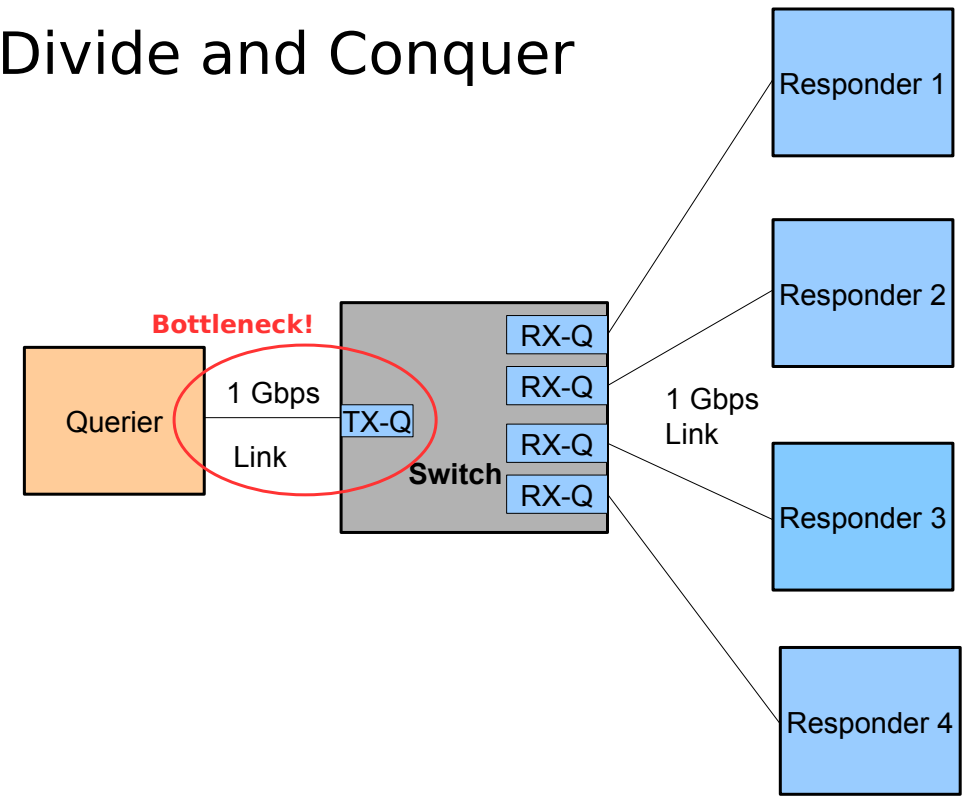
- 13 Data Centres, 900,000 servers
- In 2013, there are 2,161,530,000,000 Google search queries, 5,922,000,000 per day, indexing 20 billion web pages per day
- Free gmail storage to 425 million users
- Serve 6 billion YouTube videos per month
- How does Google handle these workloads?

Incast and Microburst

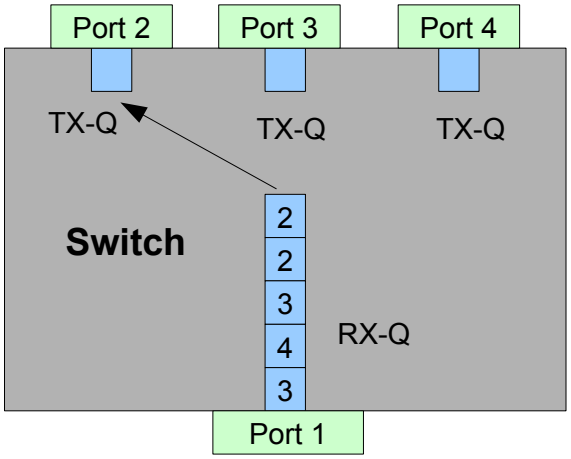
- "You can't pour two buckets of manure into one bucket"- Scott Fritchie's grandfather
- Network pathology that affects many-to-one communication pattern
  - High speed and low latency data centres
- Causes: Response synchronization and TCP timeouts
- Effect: Catastrophic throughput collapse
- Cluster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

Incast Topology

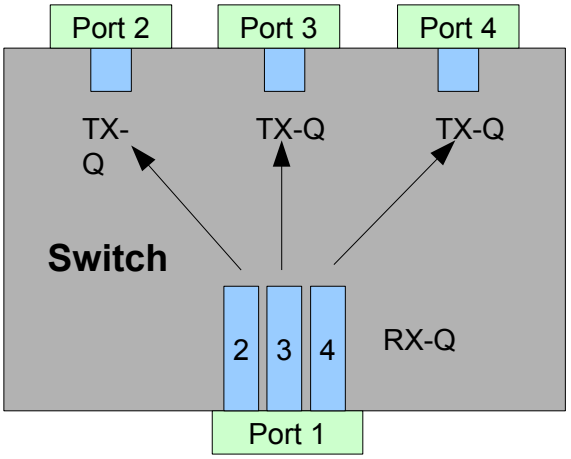
- Divide and Conquer



Virtual Output Queuing (VOQ)

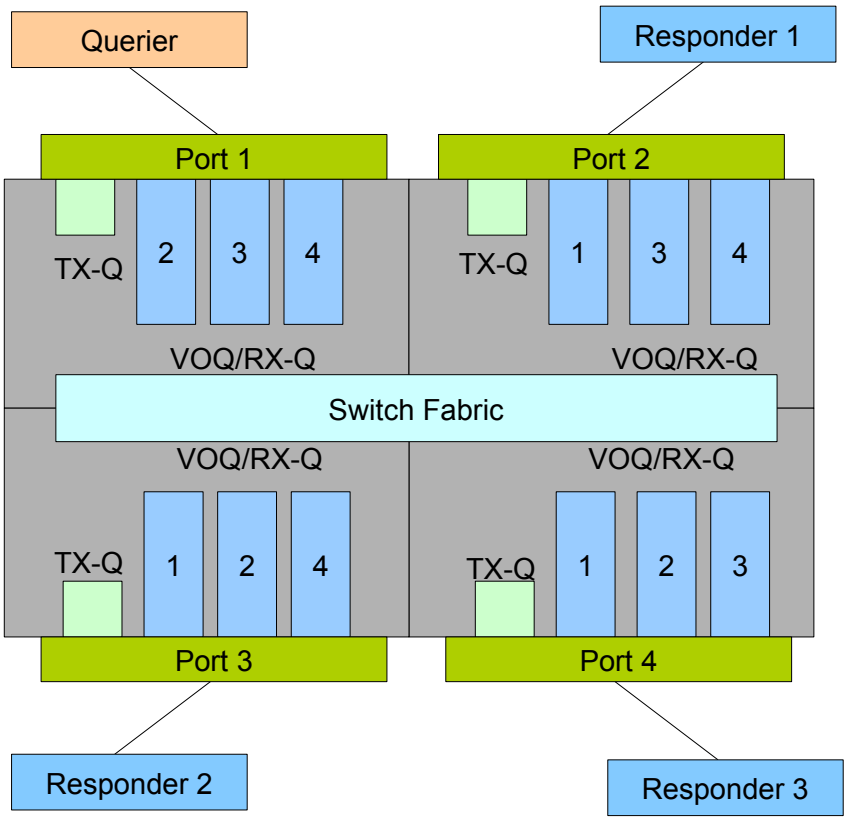


Traditional Input Queue Switch Model
Head of Line (HoL) Blocking



Switch Model with VOQ

Switch Model

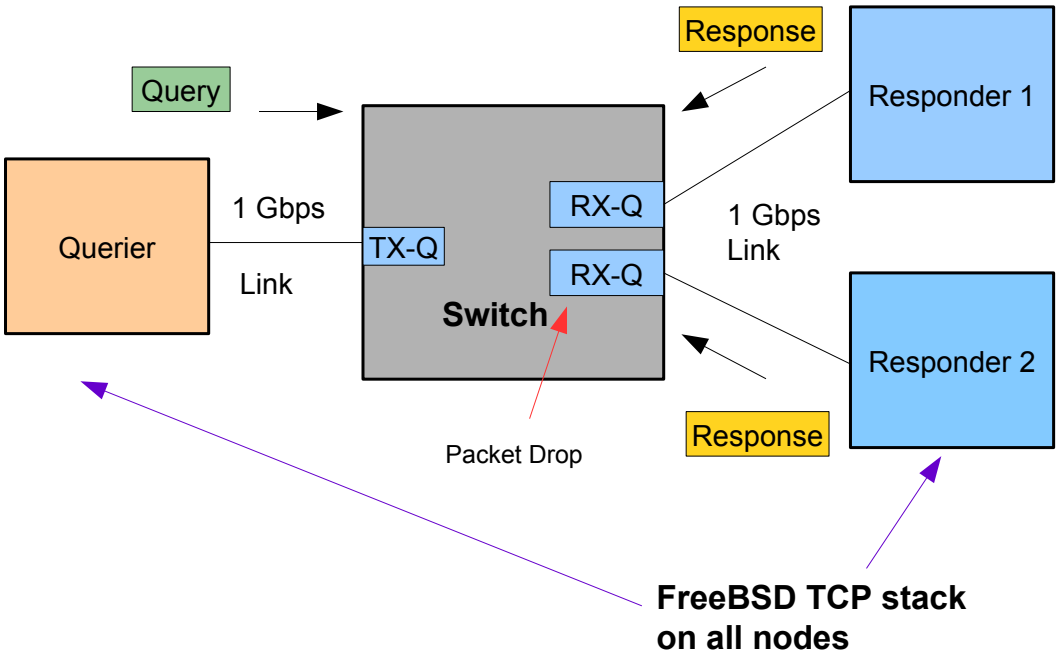


CAIA's Simulator (NS-3/NSC)

- 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

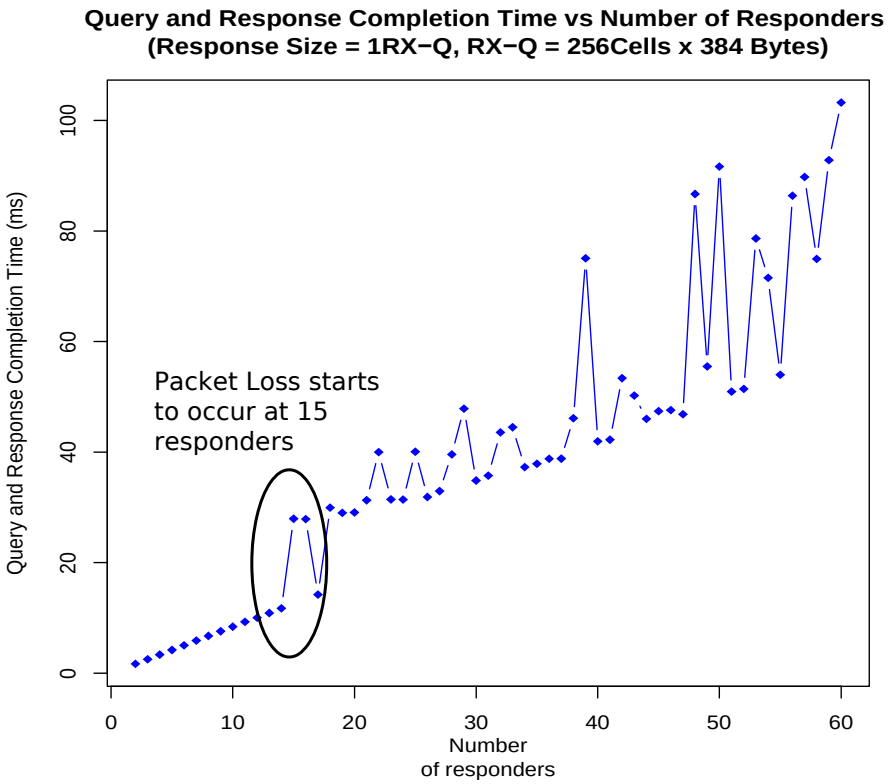
- Experiment Parameters



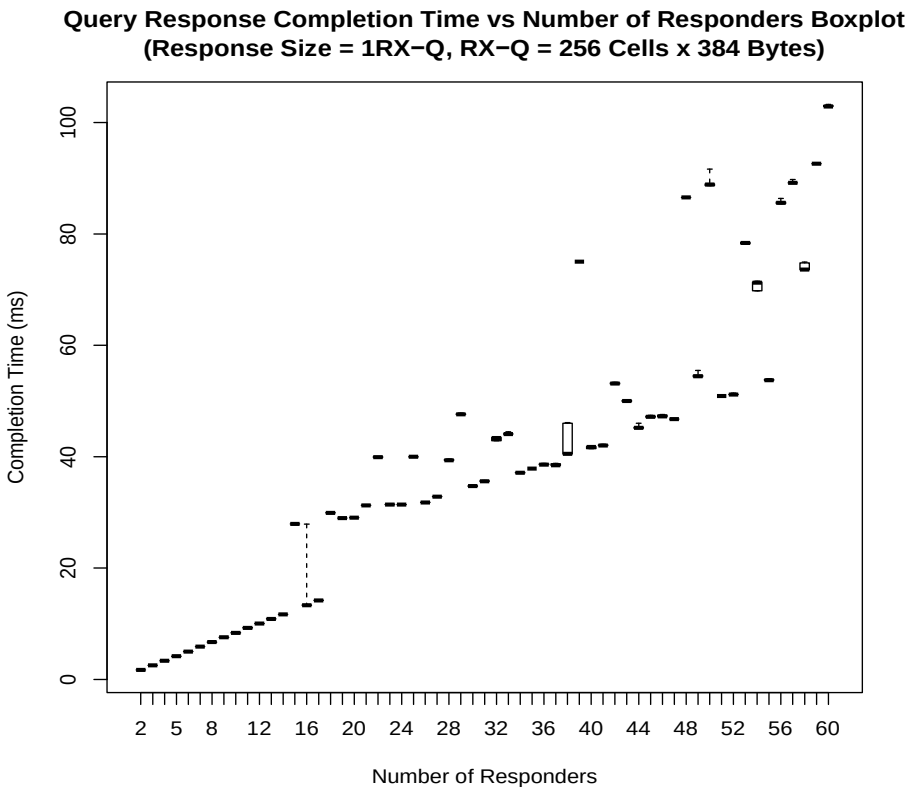
Results and Analysis

- 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

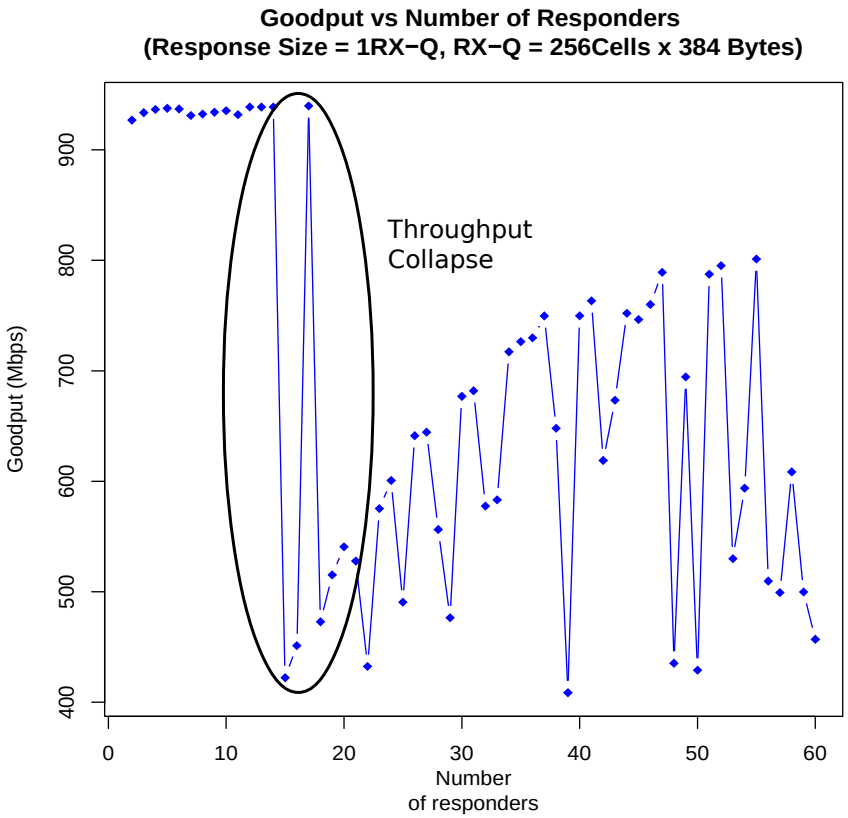
Completion Time



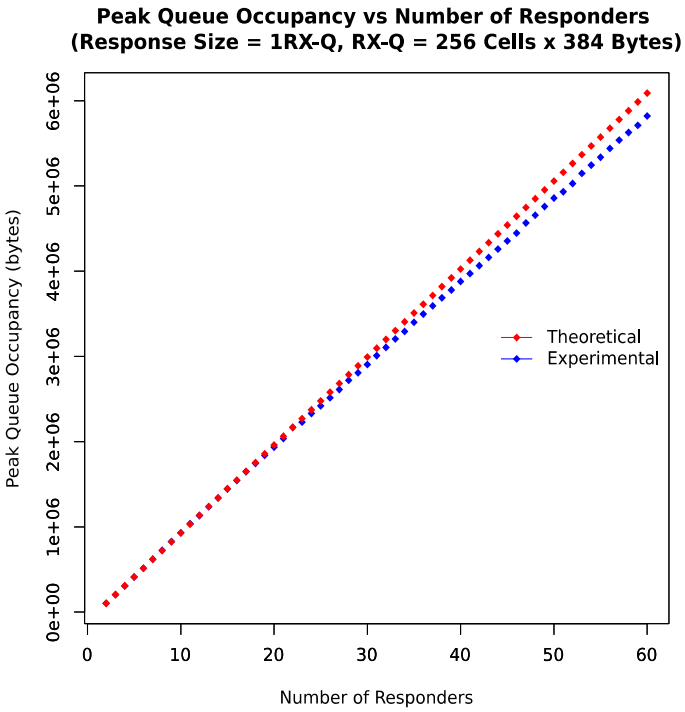
Completion Time Distribution

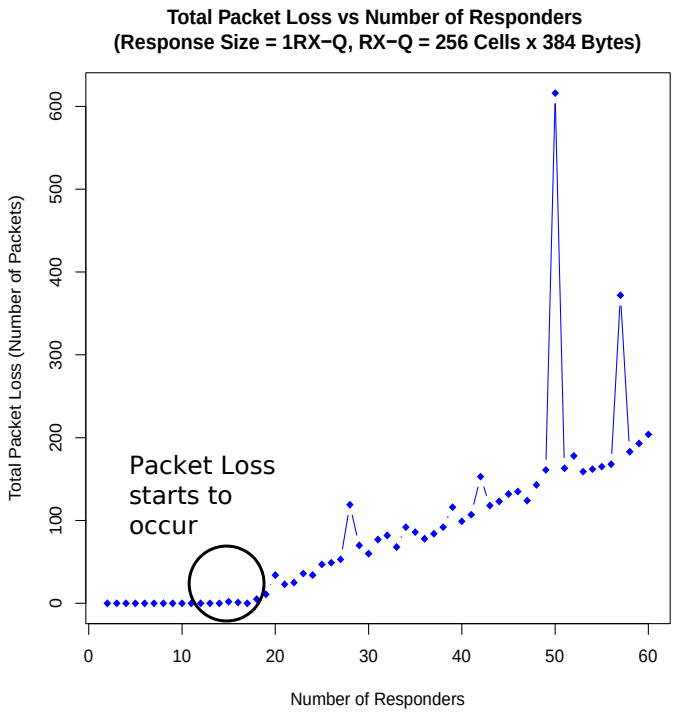


Goodput



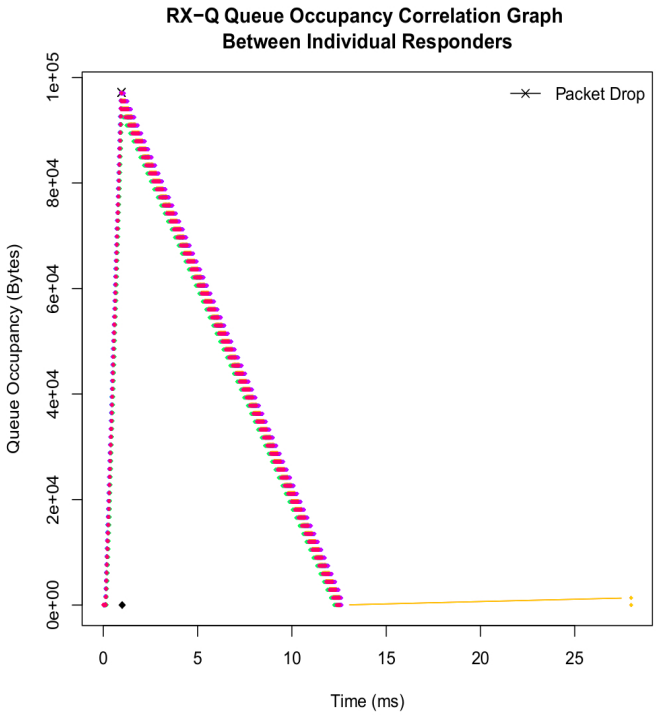
Peak Queue and Packet Loss

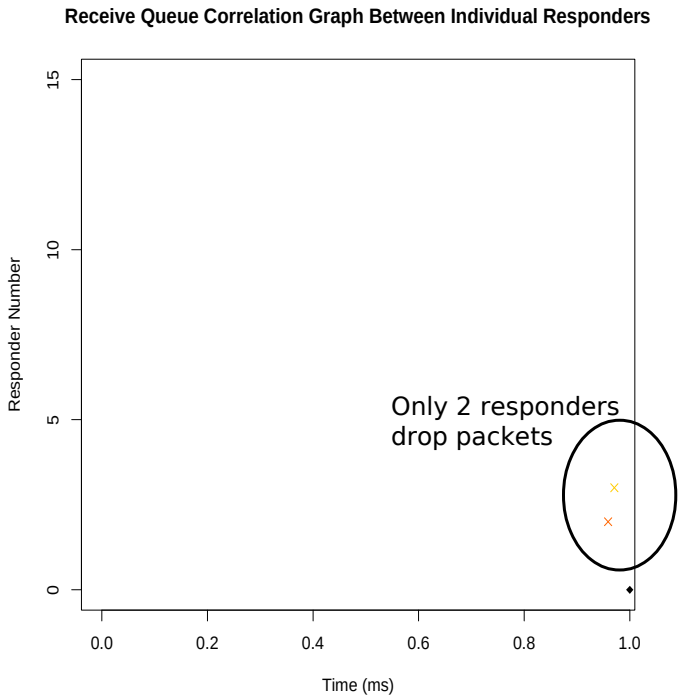




A Deeper Look

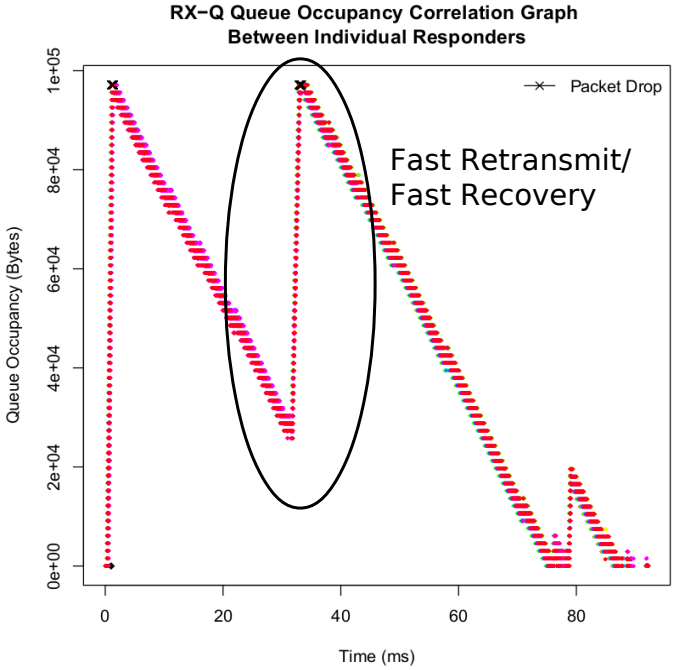
- 15 responders

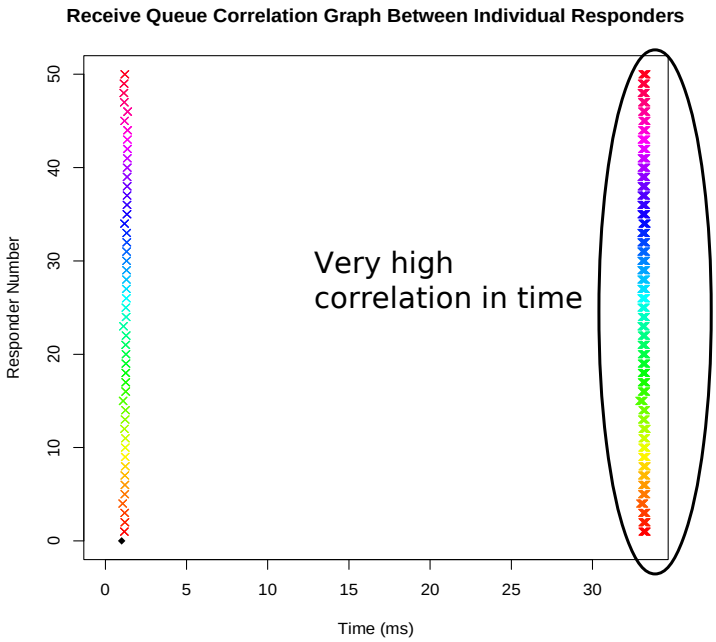




A Deeper Look

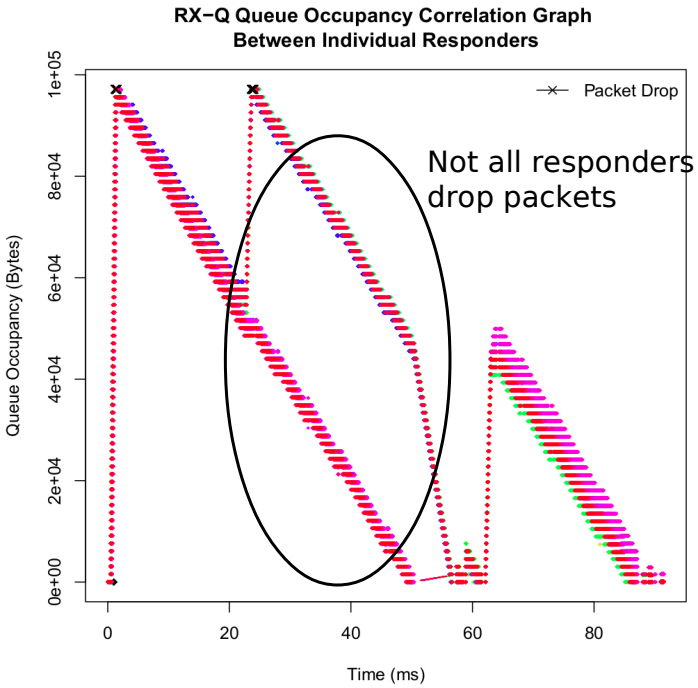
- 50 responders

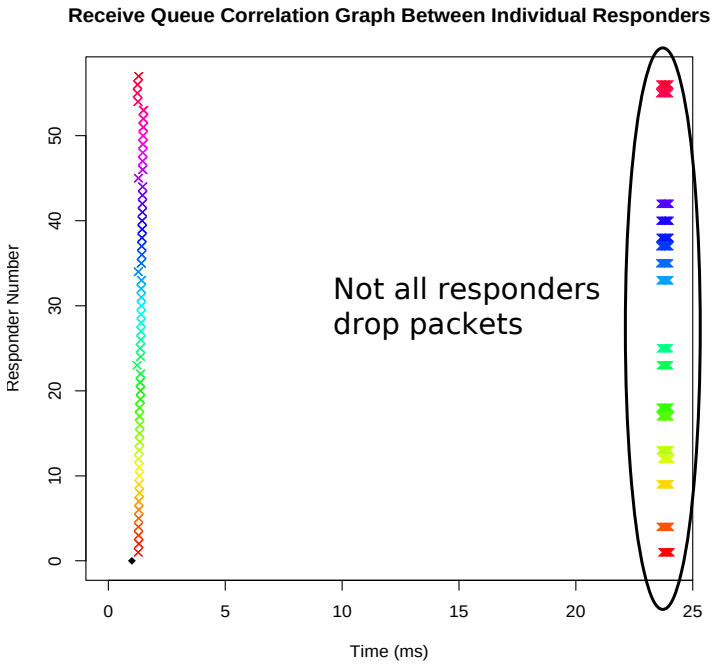




A Deeper Look

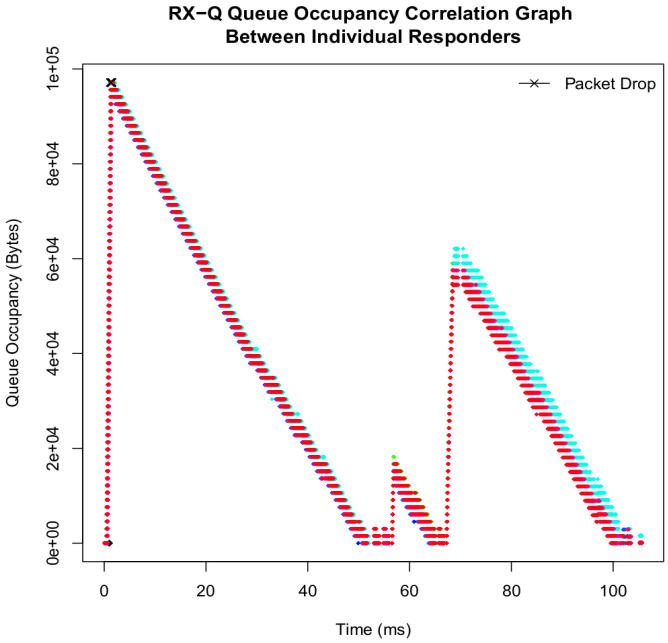
- 57 responders

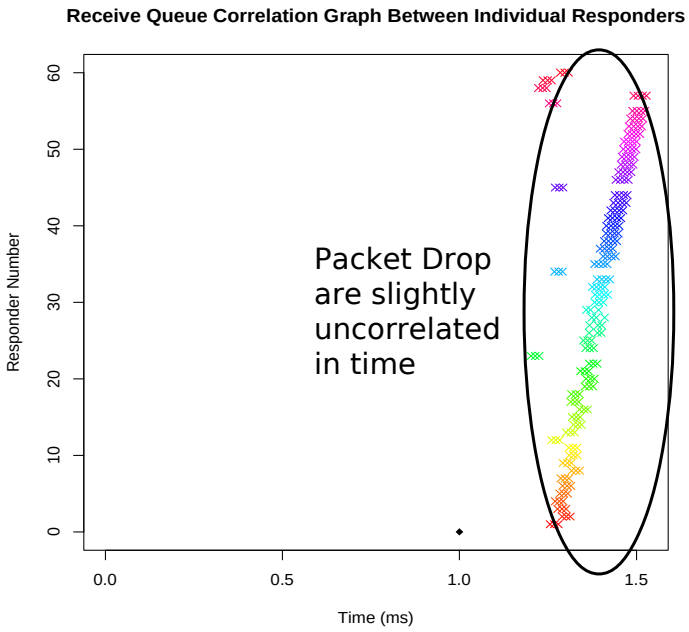




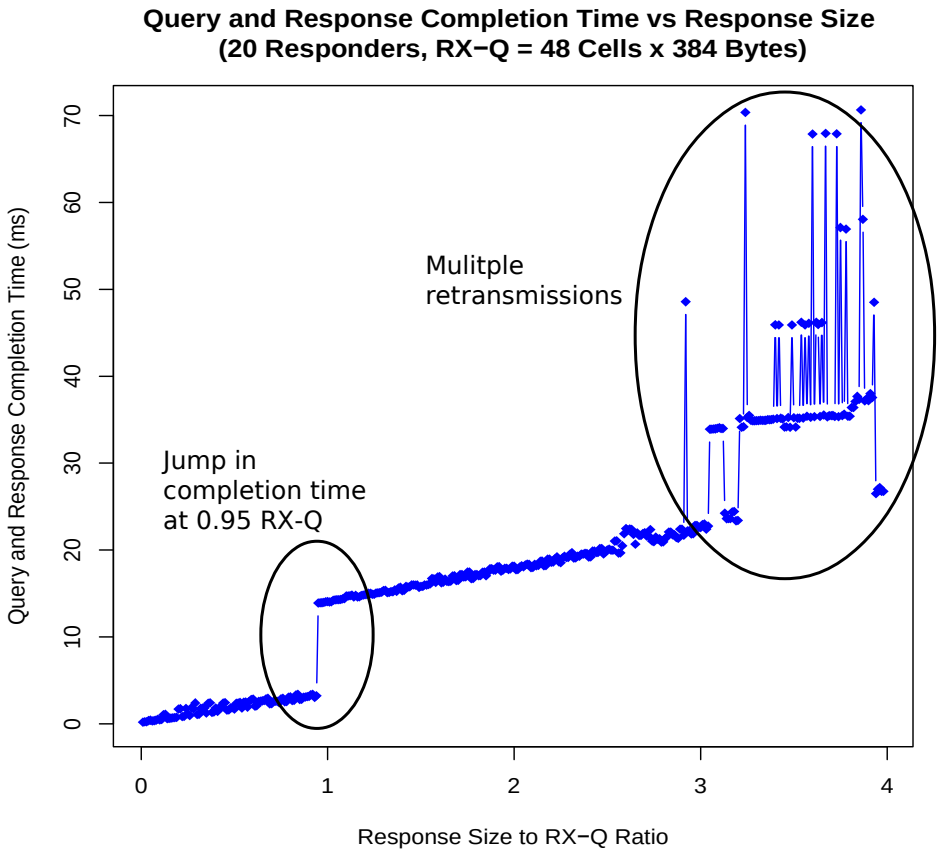
A Deeper Look

- 60 responders

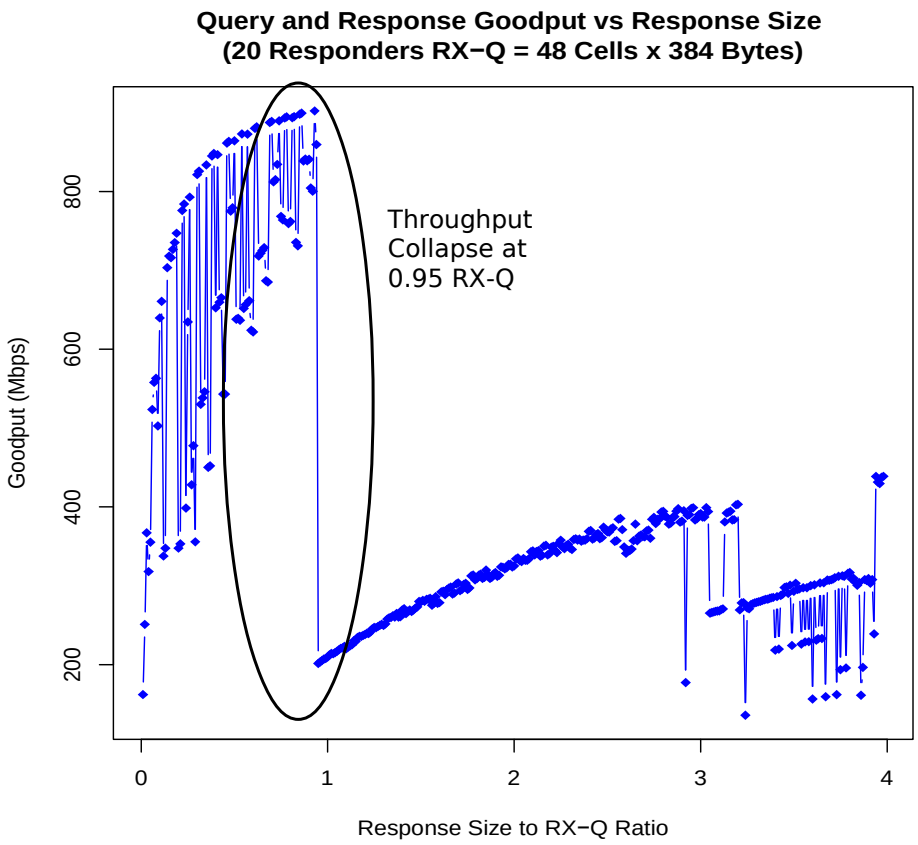




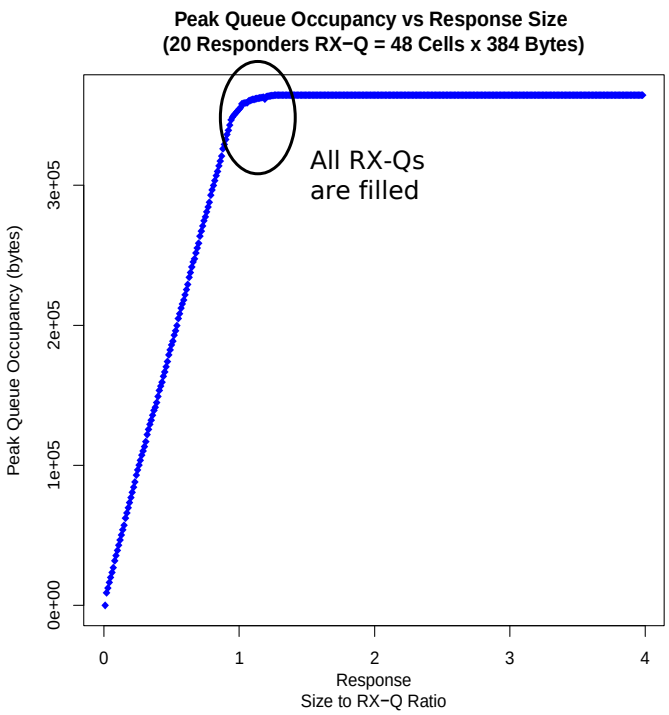
Completion Time

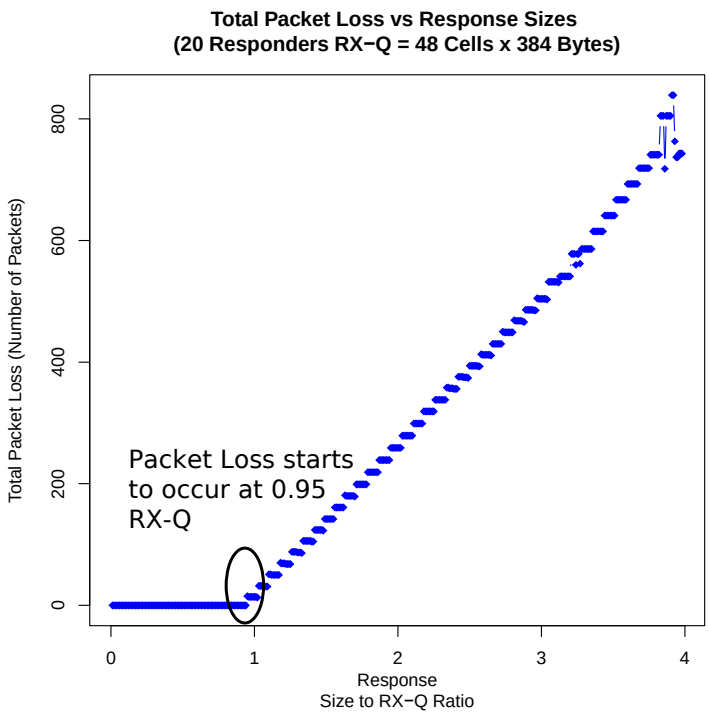


Goodput



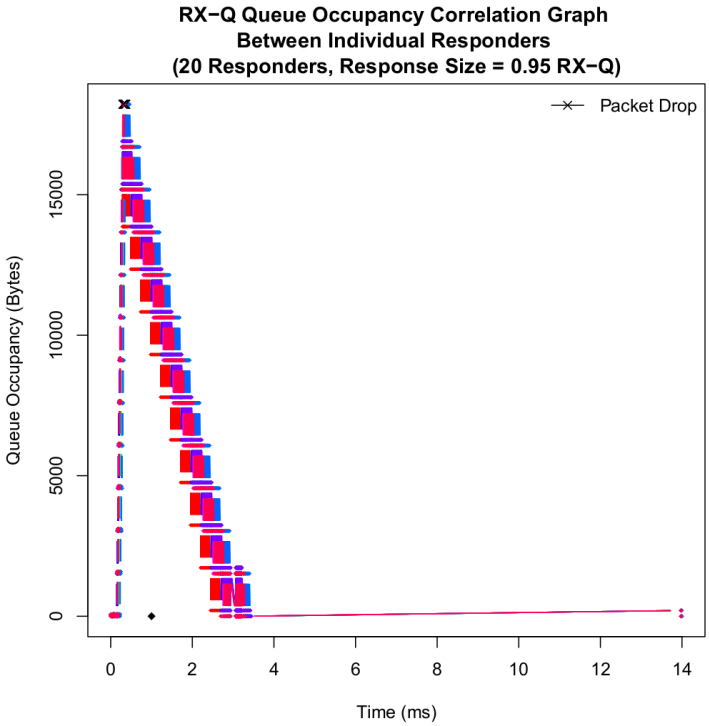
Peak Queue and Packet Loss

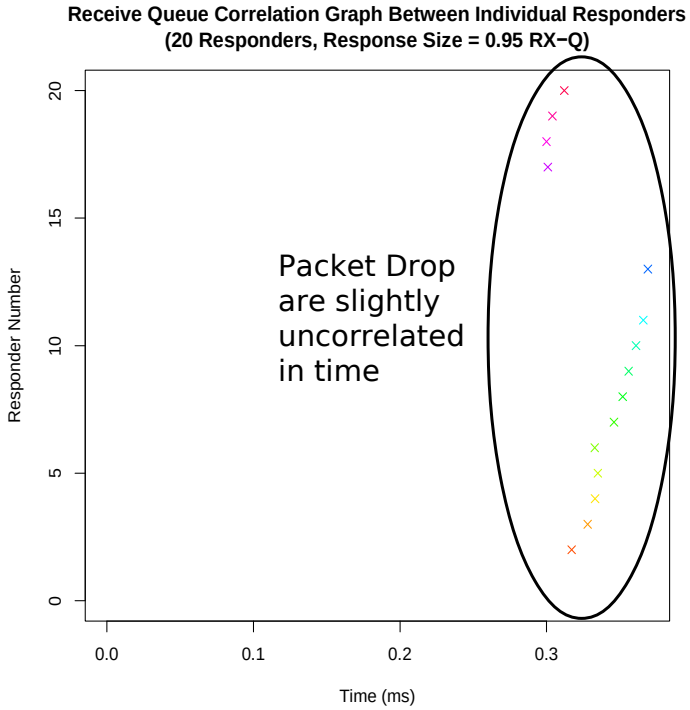




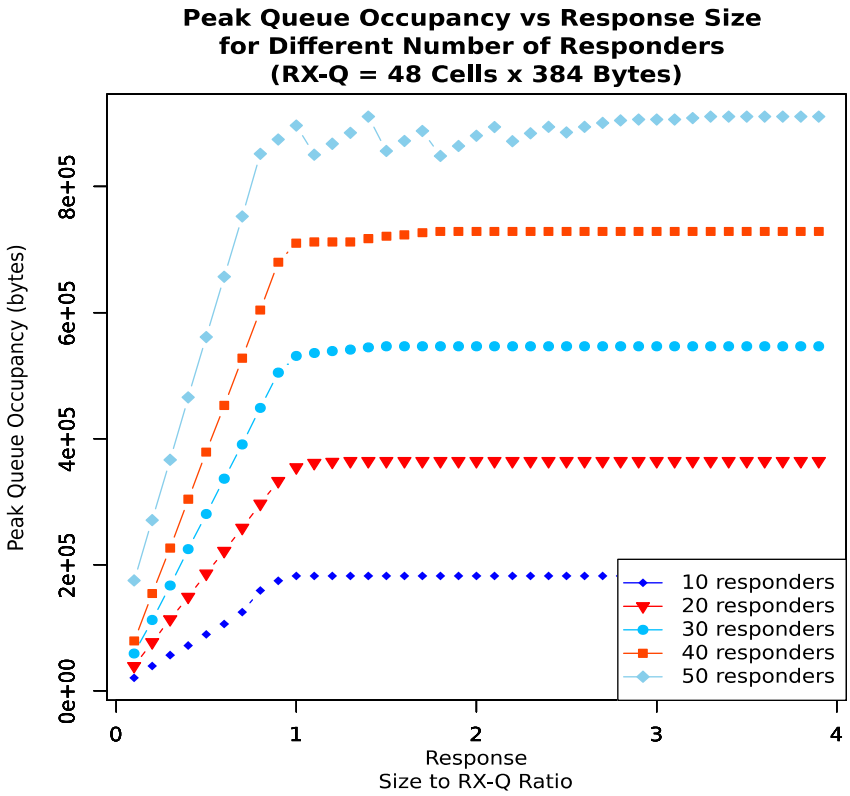
A Deeper Look

- Response Size = 0.95 RX-Q

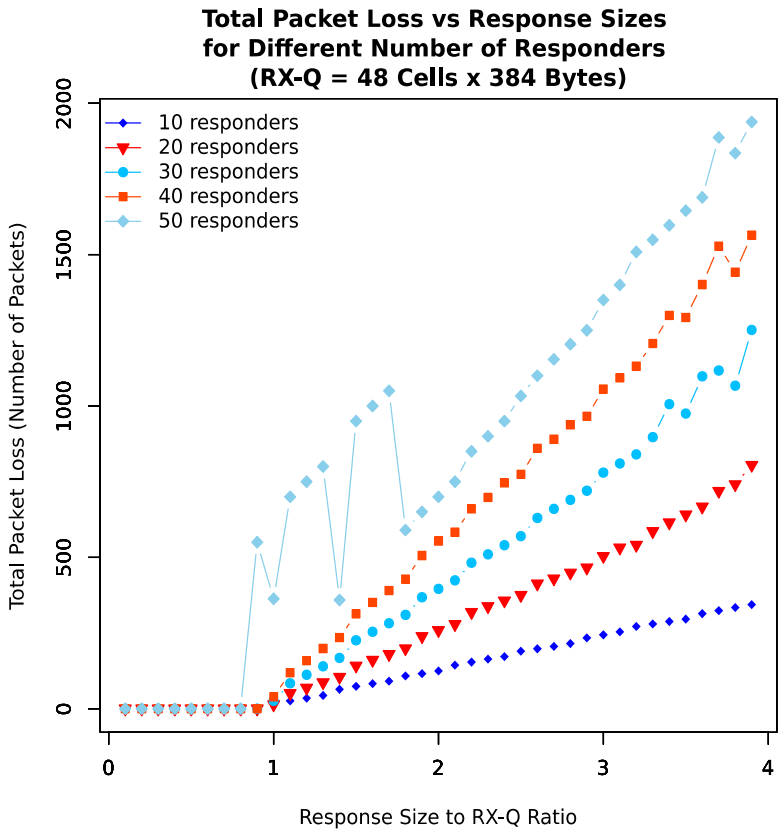




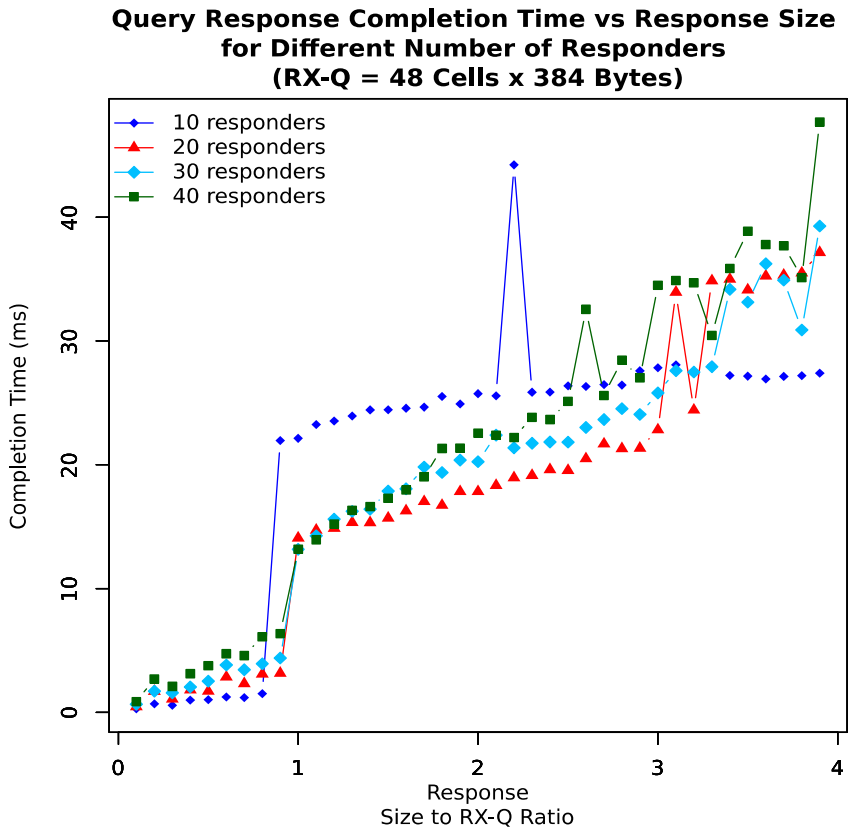
Putting It All Together



Putting It All Together



Putting It All Together

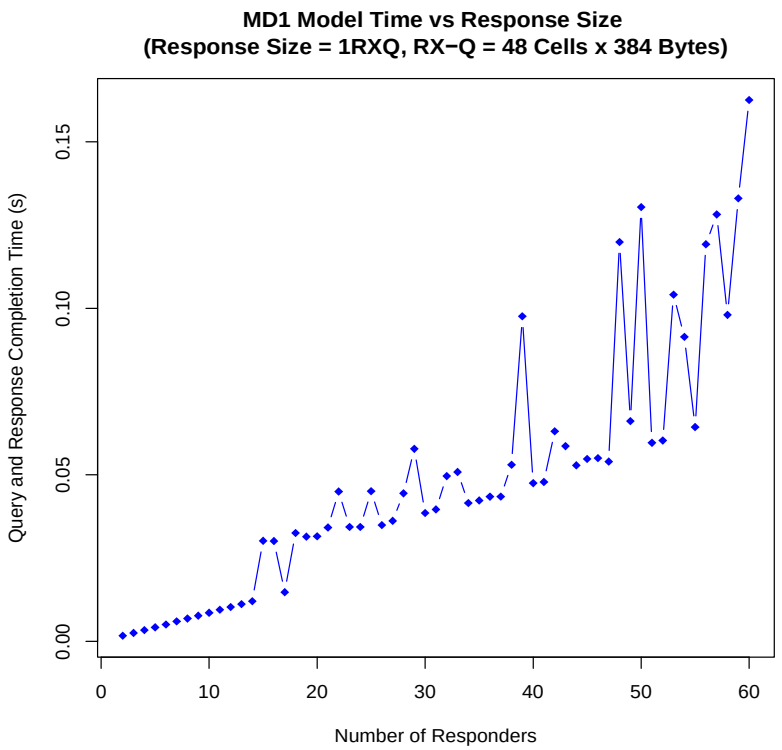


M/D/1 Queuing Model

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

M/D/1 Queuing Model

- Assume arrival rate is 5 queries/sec



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
  - Proactve: 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
- A. Phanishayee, et al. (2008). 'Measurement and analysis of TCP hroughput collapse in cluster-based storage systems'. In FAST'08: Proceedings of the 6th USENIX Conference on File and Storage Technologies, pp. 1–14, Berkeley, CA, USA. USENIX Association.
- H. Tang, et al. (2004). 'The Panasas ActiveScale Storage Cluster - Delivering Scalable High Bandwidth Storage'. Sc 00.
- H. Wu, et al. (2013). 'ICTCP: incast congestion control for TCP in data-center networks'. IEEE/ACM Trans. Netw. 21(2):345–358.

References

- M. Alizadeh, et al. (2010). ‘Data center TCP (DCTCP)’. In Proceedings of the ACM SIGCOMM 2010 conference, SIGCOMM ’10, pp. 63–74, New York, NY, USA. ACM.
- V. Vasudevan, et al. (2009). ‘Safe and effective fine-grained TCP retransmissions for datacenter communication’. SIGCOMM Comput. Commun. Rev. 39(4):303–314.
- Y. Chen, et al. (2009). ‘Understanding TCP incast throughput collapse in datacenter networks’. In Proceedings of the 1st ACM workshop on Research on enterprise networking, WREN ’09, pp. 73–82, New York, NY, USA. ACM.
- Zhang, J.; Ren, F. & Lin, C. (2011), Modeling and understanding TCP incast in data center networks., in 'INFOCOM' , IEEE, , pp. 1377-1385 .

Thank You

Thank You!

Questions & Answers

Questions?