Networking with Deep Buffers

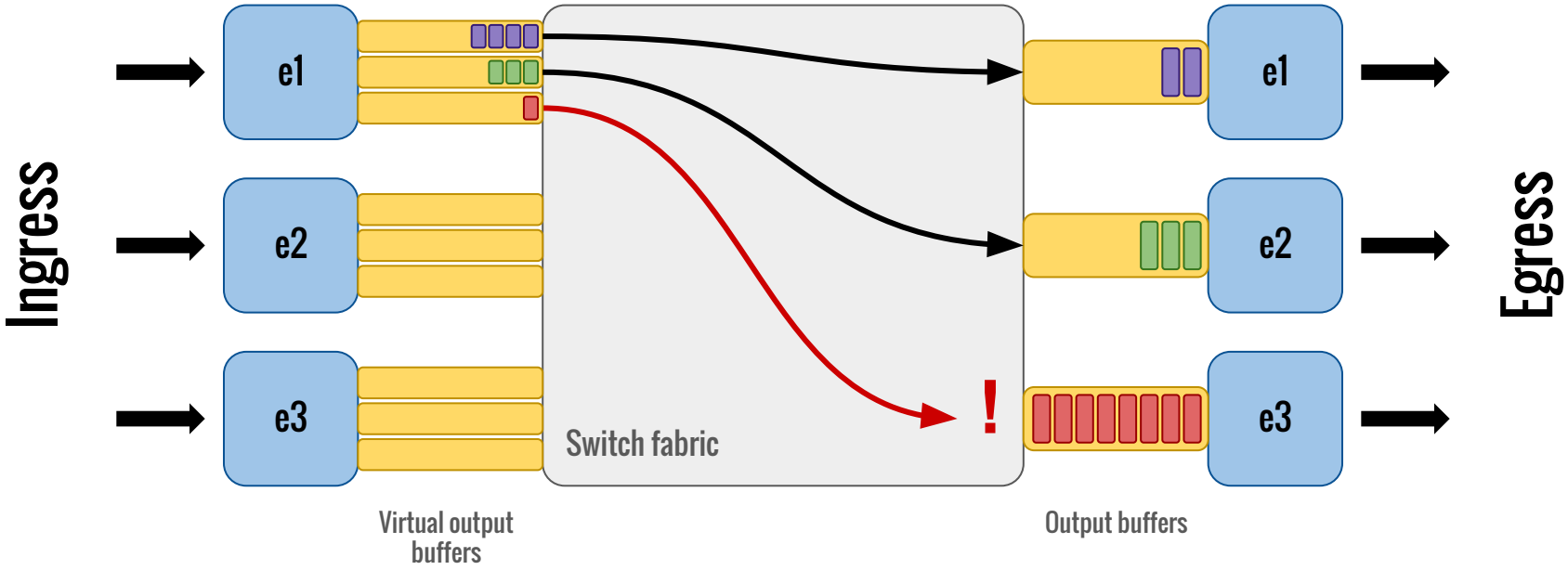
19/7/16

Aaron Levitan

Switching Asic Packet Buffer Size

- Broadcom (StrataXGS) Trident2
  - Internal 16 MB, shared between all ports
- Broadcom (StrataXGS) Tomahawk
  - Internal 16 MB, split into 4 MB per pipeline, shared within a pipeline
- Broadcom (Dune) Arad
  - External DDR3, up to 32 GB per asic, per port allocation
- Broadcom (Dune) Jericho
  - External GDDR5 or DDR4, up to 48 GB per asic, per port allocation
- Intel (Fulcrum) Alta
  - Internal 2-8MB, shared between all ports

Virtual Output Queues (VOQs)



VOQs

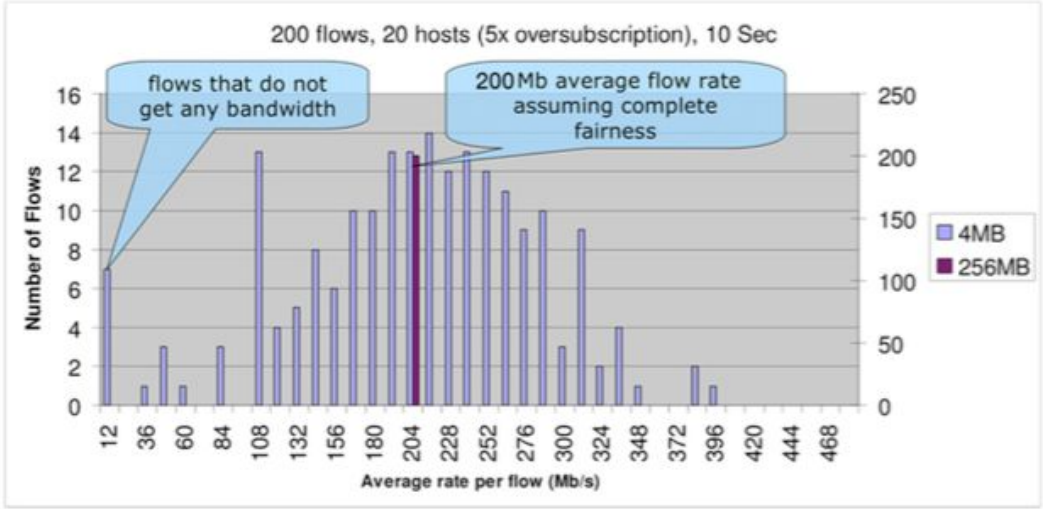
- Per traffic class, per port
- Example tail drop settings
  - 50MB for 10G
  - 200MB for 40G
  - 500MB for 100G

Simulation whitepaper and case study

- https://www.arista.com/assets/data/pdf/Whitepapers/BigDataBigBuffers-WP.pdf
- http://architects.dzone.com/articles/platform-performance-gains

TOR Simulation

- ns-2 network simulator with standard tcp/ip settings
- 20 servers with 10G connection to TOR
- 10 flows per server
- 40G uplink, 5:1 over subscription
- Compare small buffer vs big buffer switch



TCP/IP Bandwidth Capture Effect

- TCP drops packet at the switch if no buffers available
- Whether buffers available depends on when server flow arrives
- Large variation in bandwidth when using small shared buffers
- Fair allocation when using per port large buffers
- Overall uplink throughput unaffected
- In data center applications tail latency affects workload throughput

Results at DataSift

- Real-life example instead of simulation
- Replaced 1G small buffer TOR with 1G big buffer TOR
- 6x improvement

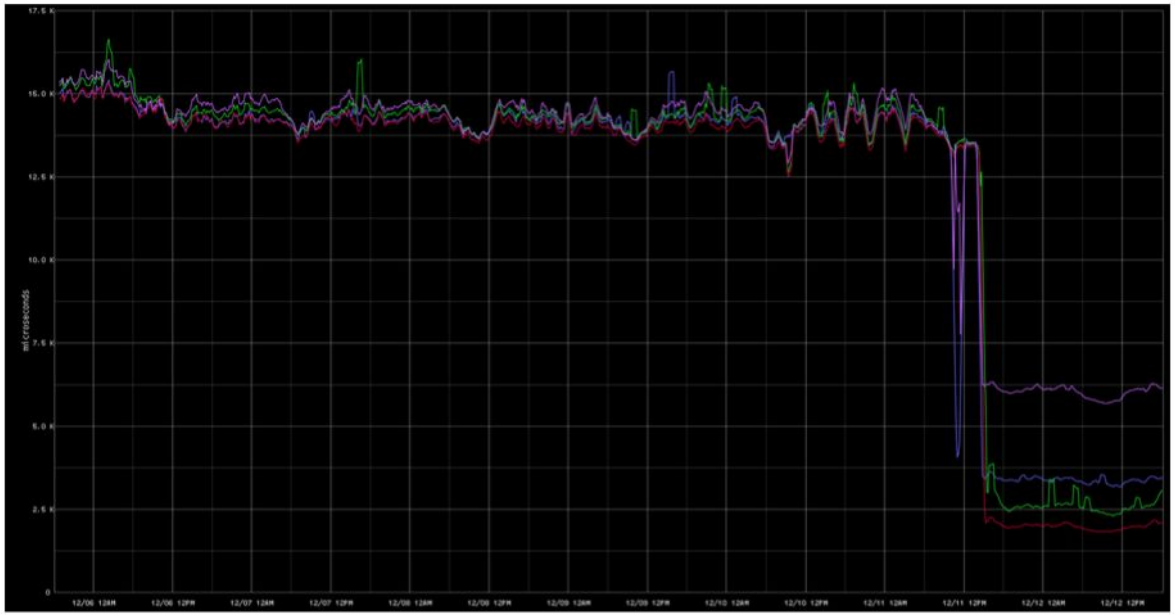


Figure 5: Average latency seen in a Big Data cluster before and after introducing big buffer switches [5].
http://architects.dzone.com/articles/platform-performance-gains