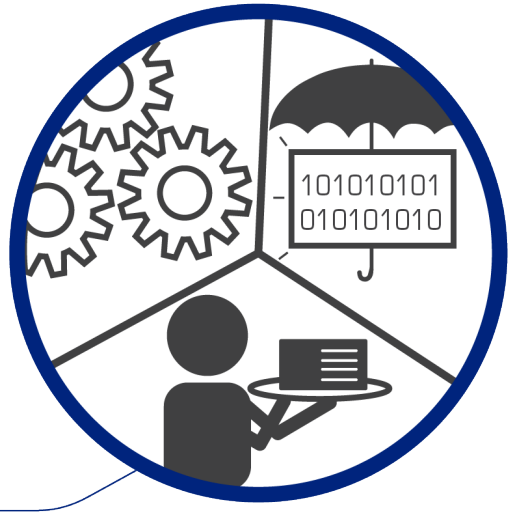
Buffer Sizing Revisited

SWITCH

Simon Leinen
simon.leinen@switch.ch

SwiNOG #35, Bern, 08 May 2019

Our core beliefs



Together for **greater capability**, **convenience and security** in the digital world.

How big should they be?

- $C \times RTT$ (bottleneck capacity × (“average”) round-trip time)
  - *High Performance TCP in ANSNET,* C. Villamizar et al. 1994
- $\frac{C \times RTT}{\sqrt{N}}$ (divided by number of flows)
  - *Sizing Router Buffers,* G. Appenzeller et al. 2004
- O(log W) (on the order of the log of TCP window size…
  few dozen packets (if you’re willing to run links below 100% utilization and TCP is “not overly bursty”)
  - *Routers with Very Small Buffers,* Enachescu et al. 2006
  - *Experimental Study of Router Buffer Sizing,* Beheshti et al. 2008
    for some caveats in real networks

Confused yet?

How big are they in practice?

- With commodity switching silicon, either quite small (9–64 MB) or quite large (2–4 GB)
  - Depending on whether internal or external memory is used
  - For external, you need lots of bandwidth, so several memory channels/chips, which increases the minimum size
- With large buffers, you ~halve chipset capacity for I/O

Why should I care?

- If your buffers are too big
  - Wasted CAPEX/bandwidth
  - Potential for negative performance effects from “bufferbloat”
- If buffers are too small
  - Packet drops are a fact of life and TCP will just slow down, but:
    - Might hit “innocent” flows
    - Might hit flows at a bad time
    - Can mess with your tail latencies
- What do you optimize for?
  - Cost (util.), page load times (avg./median/99%ile/99.99%ile/…?), “ping”…

What's in them?

- Really hard to tell!
- Buffer management is in the "fast path"

  ⇒hard to (micro)measure what really happens there
- Chipsets have some instrumentation (high-water marks, histograms)
  - In practice, hard/impossible for operators to access
  - OEM/OS vendors can help (please don't limit access to extra-cost analytics)
  - Possible IETF work to model buffers/queues for monitoring?
- Other ways to find out: INT/IOAM, passive/active measurement

What's New?

- New attempts at limiting queuing, e.g. L4S/DualQ
- More pacing at the high-speed edge? (BBR etc.)

Where can I learn more?

- Stanford buffer sizing page
- Ivan Pepelnjak's buffer sizing page
- Soothing animations for thinking about buffer sizes
- Next RIPE meeting, MAT session (TY Huang, Netflix)
- At an upcoming workshop (late 2019/early 2020?)
  - Pre-workshop happened in early March, slides hopefully forthcoming—lots of interesting measurements from various large operators (edge, content/cloud, ISP)

SWITCH

Working for a better digital world