Designing accessible latency metrics

Toke Høiland-Jørgensen

25th September 2013

Quantifying latency

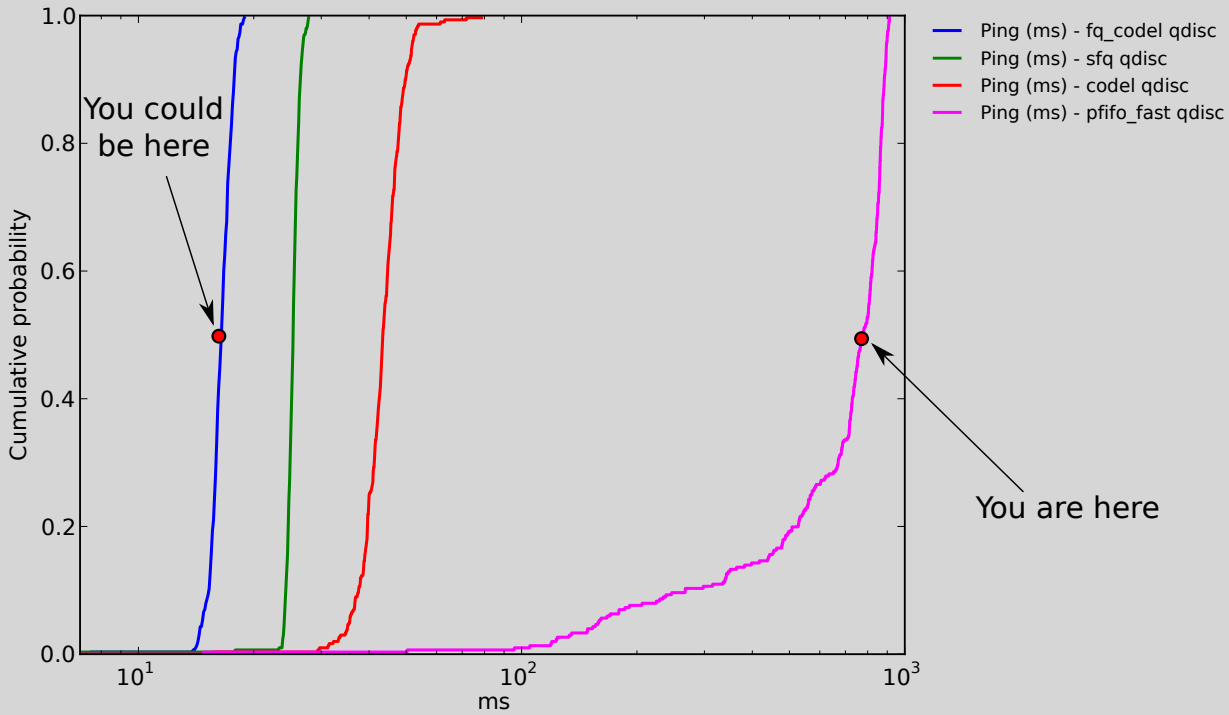
- Bufferbloat is (getting) accepted in technical circles
- But mostly unknown to end-users

- Can be explained with some care
- But how to quantify?

What do we have now?

- The RRUL test
  - Eight TCP streams to induce load
  - Measure UDP and ICMP ping times
  - Comparison using CDF plots

Example



What to measure?

- Minimum unloaded latency
  - To where? ISP, nearest exchange, major site(s), E2E?
  - From where? User device(s), CPE equipment?
  - Uni/bidirectional? ICMP, UDP, load a full website?

- Latency under (saturated) load
  - Needs reliable method to induce load
  - Really hitting ”worst case” probably hard
  - And what about outliers?
  - Sampling frequency

- The ratio between the two
  - A ”load degradation factor”?
  - Logarithmic, linear, normalised?

Communicating the metric

- What is important for the user to know?
  - An absolute measurement (ms) or a relative one (score)?
- Are bigger numbers better?
  - Roundtrips/second rather than milliseconds of latency?
- Should minimum latency and degradation be combined into one metric? How?

What can the metric be used for?

Goals (short term?)

- Expose bufferbloat in the network
- Enable the consumer to influence latency/bandwidth tradeoff decisions
- Create incentives for improvement

- User (self-)information (like speedtest.net)
- ISP comparison charts
- Regulation requirements (e.g. bounds)
- QoS definition in contracts etc.
- Benchmarking in systems engineering