It’s the latency, stupid!

Toke Høiland-Jørgensen
Principal Kernel Engineer, Red Hat

Driving IT
November 4th, 2022

It's the Latency, Stupid

Stuart Cheshire, May 1996.

(Revised periodically)

Copyright © Stuart Cheshire 1996-2001

Years ago David Cheriton at Stanford taught me something that seemed very obvious at the time -- that if you have a network link with low bandwidth then it's an easy matter of putting several in parallel to make a combined link with higher bandwidth, but if you have a network link with bad latency then no amount of money can turn any number of them into a link with good latency.

It's now many years later, and this obvious fact seems lost on the most companies making networking hardware and software for the home. I think it's time it was explained again in writing.

Fact One: Making more bandwidth is easy.

Imagine you live in a world where the only network connection you can get to your house is a 33kbit/sec modem running over a telephone line. Imagine that this is not enough for your needs. You have a problem.

The solution is easy. You can get two telephone lines, and use them together in parallel, giving you a total of 66kbit/sec. If you need even more you can get ten telephone lines, giving you 330kbit/sec. Sure, it's expensive, and having ten modems in a pile is inconvenient, and you may have to write your own networking software to share the data evenly between the ten lines, but if it was important enough to you, you could get it done.

It may not be cheap, but at least it's possible.

People with ISDN lines can already do this. It's called "bonding" and it uses two 56 (or 64) kbit/sec ISDN channels in parallel to give you a combined throughput of 112 (or 128) kbit/sec.

http://www.stuartcheshire.org/rants/Latency.html

https://www.youtube.com/watch?v=_fNp37zFn9Q

Which of these is faster?

Which of these is faster?

High bandwidth

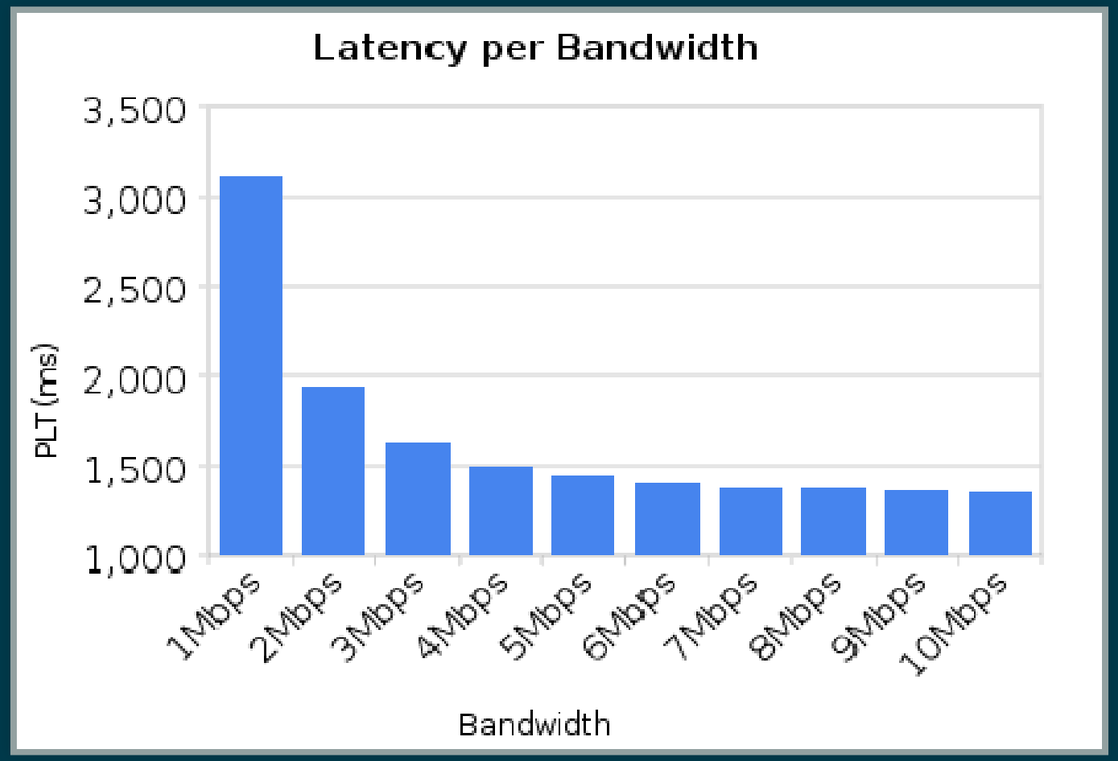
Low latency

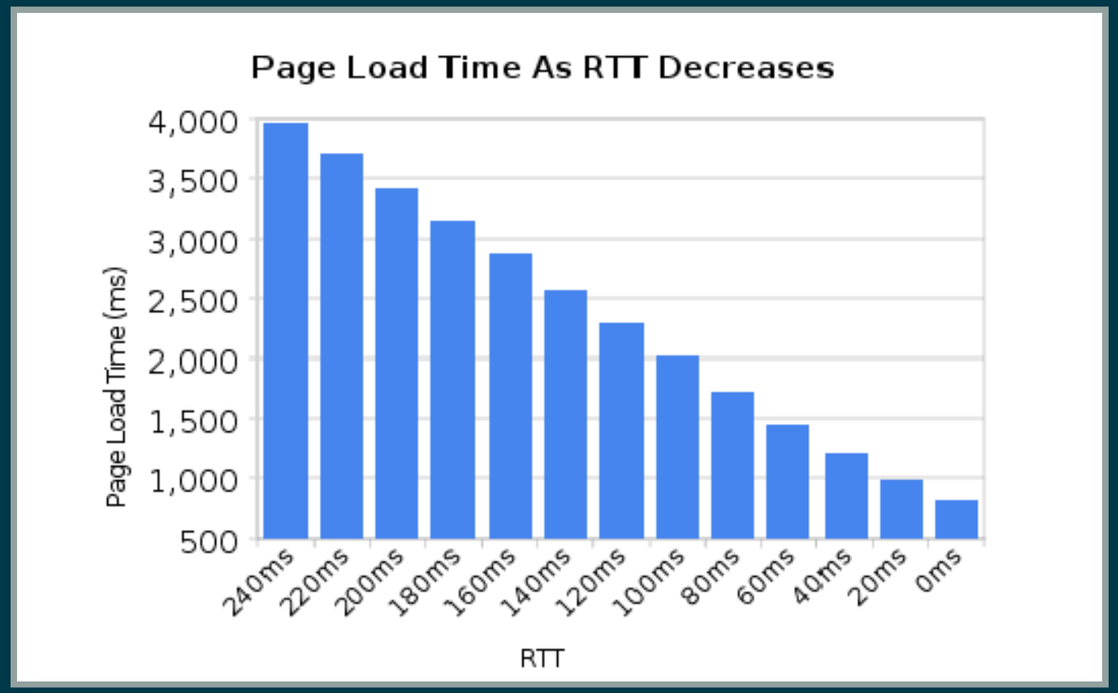
Why is this important?

“My theory here is when an interface is faster, you feel good, and ultimately what that comes down to is you feel in control. The [application] isn’t controlling me, I’m controlling it. Ultimately that feeling of control translates to happiness in everyone. In order to increase the happiness in the world, we all have to keep working on this.”

Matt Mullenweg, WordPress

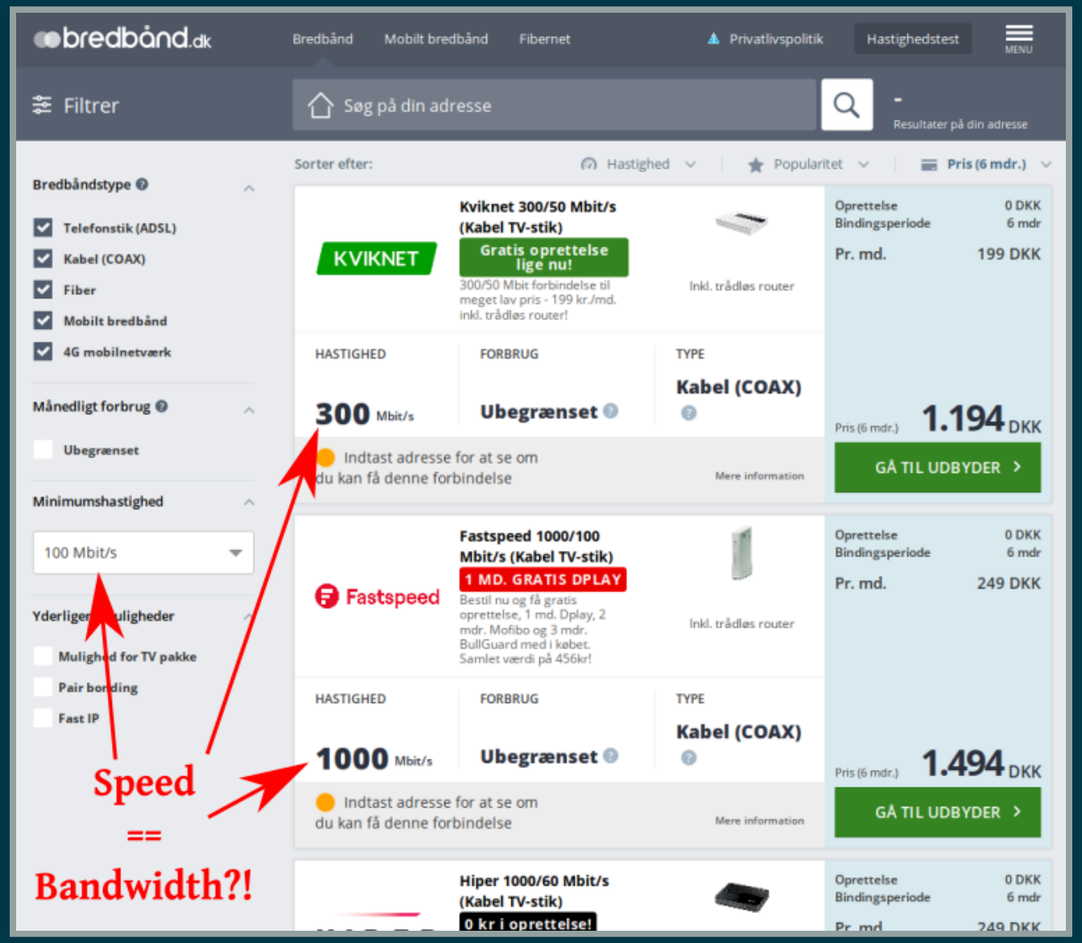
The web: Page load time





Source: https://youtu.be/TNBkxA313kk (from 2011)

So why are we still doing this?



How to add bandwidth to a low-capacity link

- Link bonding
- Compression
- Caching
- Scaling down content size
- Avoiding unneeded transfers

How to remove delay from a high-latency link

- Time travel

How to remove delay from a high-latency link

- Time travel

Latency cannot be removed it can only be prevented!

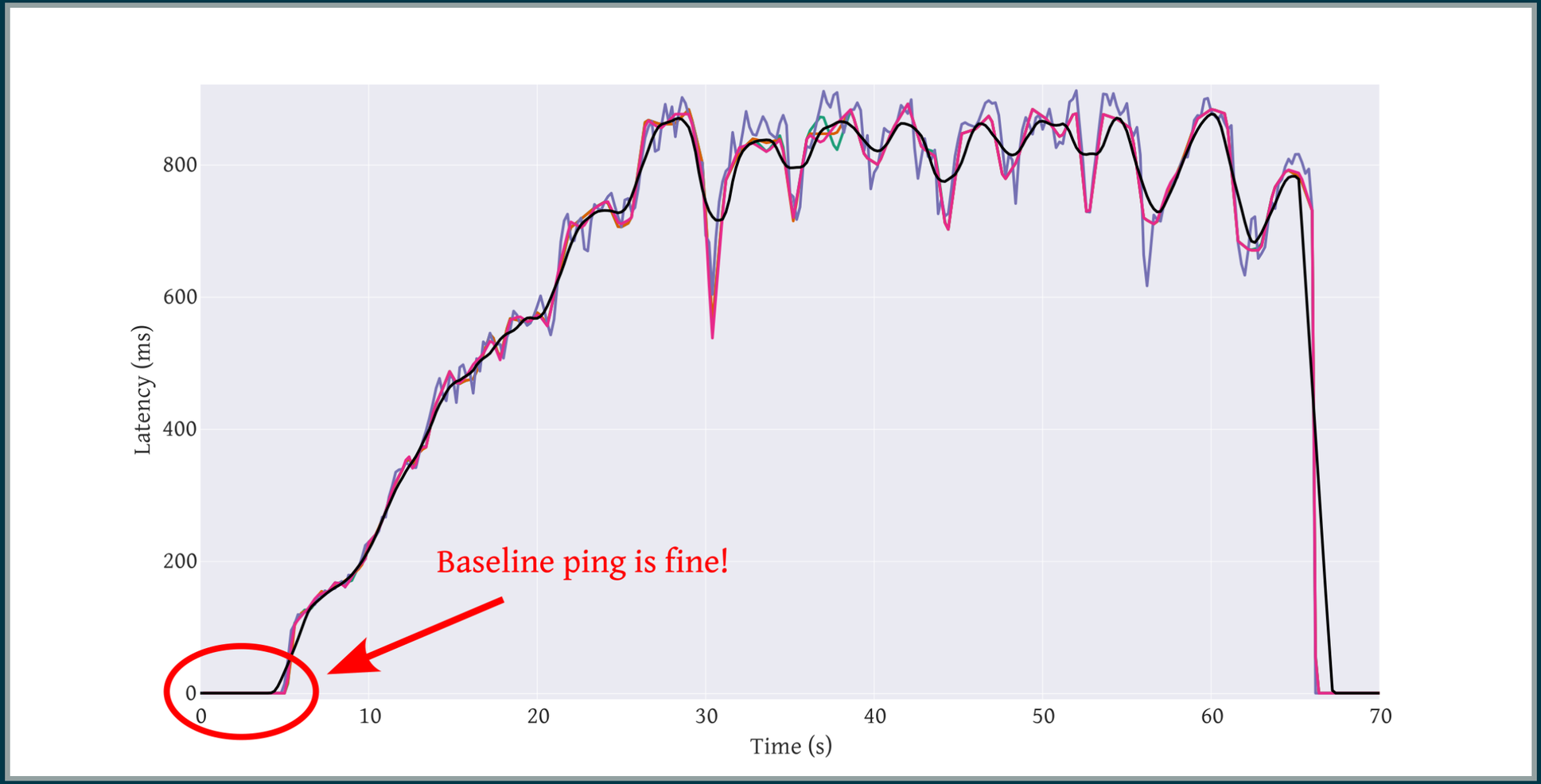
Sources of network latency

- Session setup delay
  - DNS lookups, TCP handshakes, TLS negotiation
- Serialisation delay
  - Time to transfer the data – only point affected by bandwidth
- Propagation delay
  - Media acquisition, physical distance, (bad) routing, queueing delay
- Processing delay
  - OS scheduling, head-of-line blocking, bulk processing, RPC latency

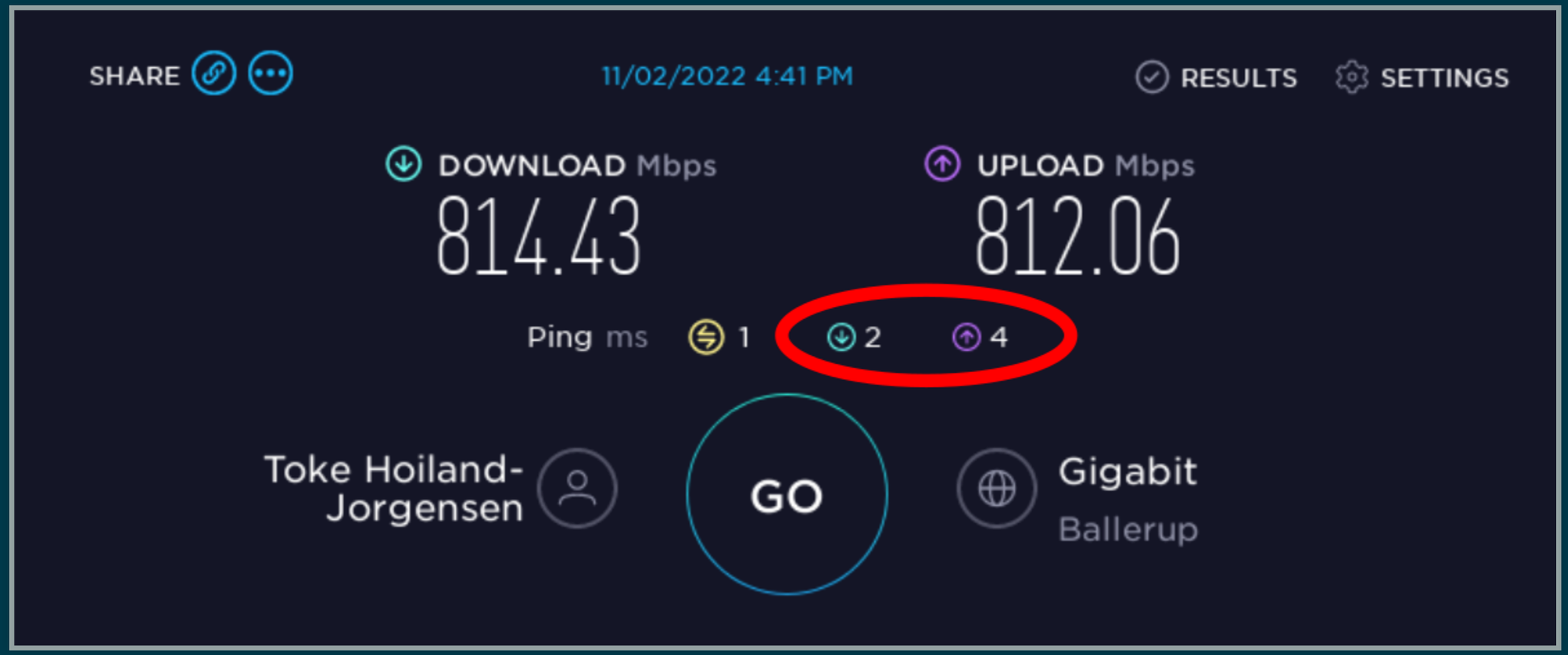
All of these need to be minimised!

Queueing delay

Bufferbloat: Latency under load



We build what we measure



It took a decade to get those two numbers added!

Summary

- Faster really means “low latency”, not “high throughput”
- It’s possible to add bandwidth, delay can only be prevented
- We build what we measure, so start measuring latency
  - Idle and under-load latency can be very different!

Faster systems make us feel in control.

Decrease latency - increase happiness!

End: Questions?

More reading:

- https://www.internetsociety.org/resources/doc/2013/reducing-internet-latency-2013-workshop-report/
- https://www.bufferbloat.net/projects/bloat/wiki/What_can_I_do_about_Bufferbloat/
- https://sci-hub.se/https://ieeexplore.ieee.org/document/6967689
- https://bufferbloat-and-beyond.net/
- https://www.waveform.com/tools/bufferbloat