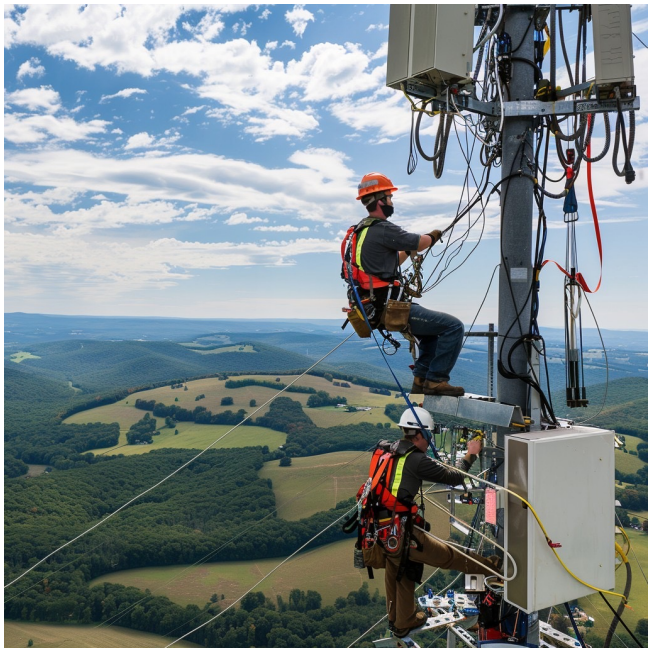
UiO : Faculty of Mathematics and Natural Sciences
University of Oslo

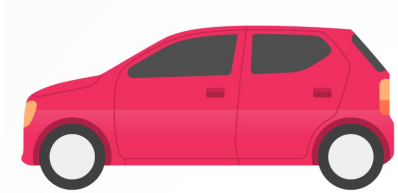
Reduce Unnecessary Hardware "Speed Upgrades" ("7G Should Be Software")

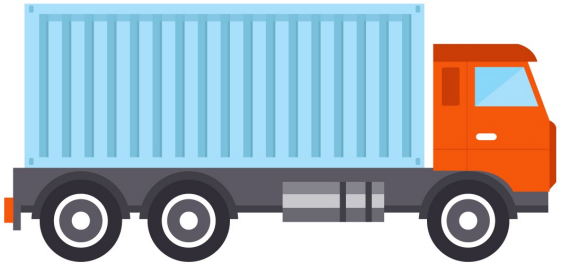
Michael Welzl

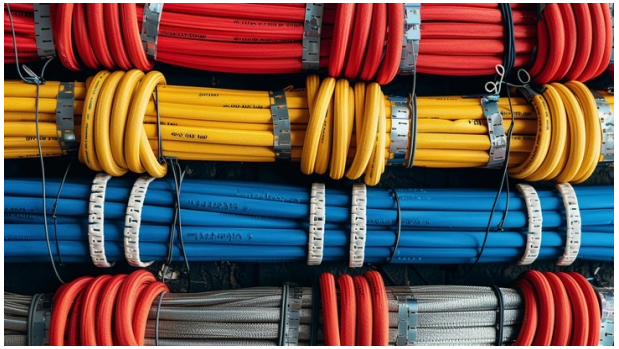
E-impact side meeting @ IETF-120
Vancouver, CA
24.7.2024

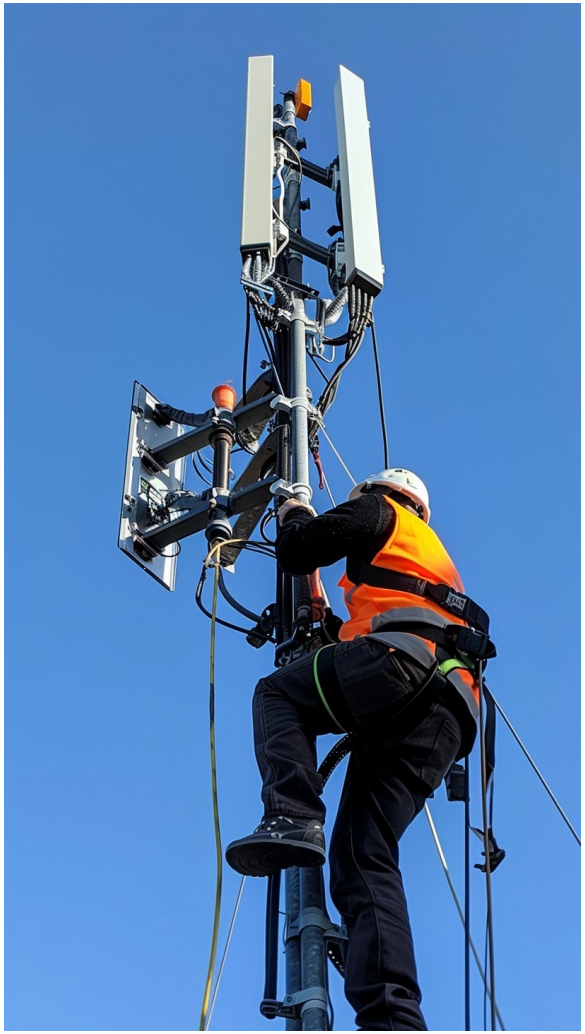
The need for speed...











Photos by Stockcake: https://stockcake.com/
Cliparts designed by Titusurya / Freepik

These “speed upgrades” can be meaningless.

Networks are underloaded !

Two major (categories of) reasons

1. Overprovisioning
   - "Killed" QoS in the 1990s: simpler and cheaper
   - Also helps resilience - probably again: the cheapest way to do it
   - But: is overprovisioning sustainable?
   - **Should** it be cheaper than QoS and Traffic Engineering?
   - Could there be ways to make traffic manipulation cheaper? (reduce manpower cost, reduce risk, ..? tax overprovisioning?)

2. Congestion control, application demands
   - E.g., 5G on your cell phone: do you feel the difference?
   - Next: some reasons for this

Transfers end in slow start

MEASUREMENTS—15-MINUTE LONG TRACES FROM EACH MONTH FROM JANUARY 2019 TO NOVEMBER 2021 AT THE TRANSIT LINK OF THE JAPANESE WIDE BACKBONE NETWORK.

Total number of flows	2555005
Flows with re-transmissions (client to server)	317754 (12.44%)
Flows with re-transmissions (server to client)	52738 (2.06%)
Total flows with re-transmissions	370492 (14.50%)

Michael Welzl, Peyman Teymoori, Safiqul Islam, David Hutchison, Stein Gjessing: "Future Internet Congestion Control: The Diminishing Feedback Problem", IEEE Communications Magazine, 2022.

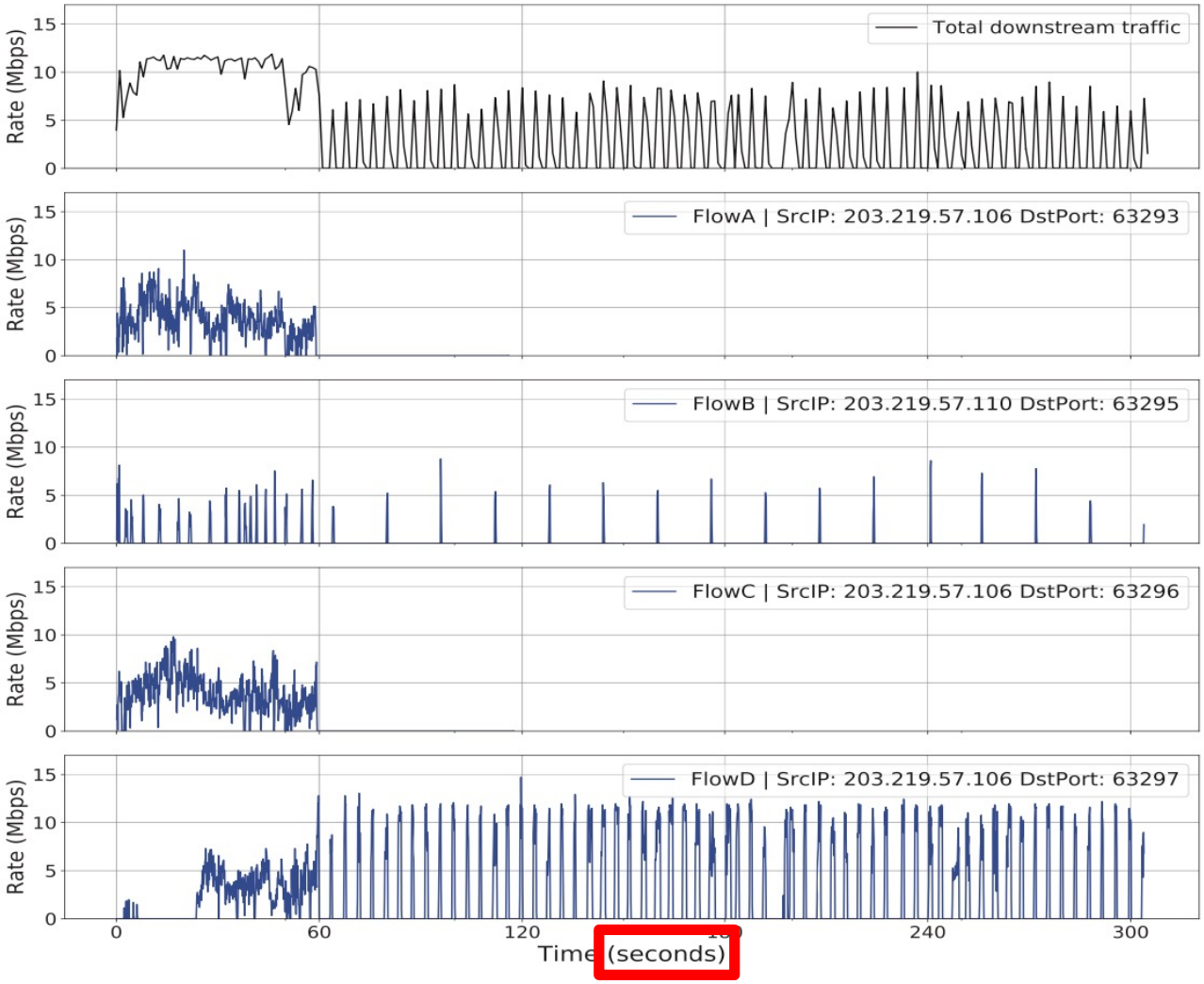
- In a large dataset consisting of traffic traces that were captured on a Tier-1 ISP backbone in Chicago between June 2008 and March 2016, 85% of all TCP connections carry between 100 B and 10 kB of data.

  S. Bauer, B. Jaeger, F. Helfert, P. Barias, and G. Carle, "On the evolution of internet flow characteristics," in Proc. ACM ANRW, 2021.

- In a measurement of one week in 2017 in the mobile search service of Baidu, Inc., 80.27% of all TCP connections terminate in slow start.

  X. Nie, Y. Zhao, Z. Li, G. Chen, K. Sui, J. Zhang, Z. Ye, and D. Pei, "Dynamic TCP initial windows and congestion control schemes through reinforcement learning," IEEE JSAC, vol. 37, no. 6, 2019.

Streaming video



Network profile of flows in a typical Netflix video stream

Taken from:

S. C. Madanapalli, H. Habibi Gharakhieli and V. Sivaraman, "Inferring Netflix User Experience from Broadband Network Measurement," 2019 Network Traffic Measurement and Analysis Conference (TMA), 2019, pp. 41-48.

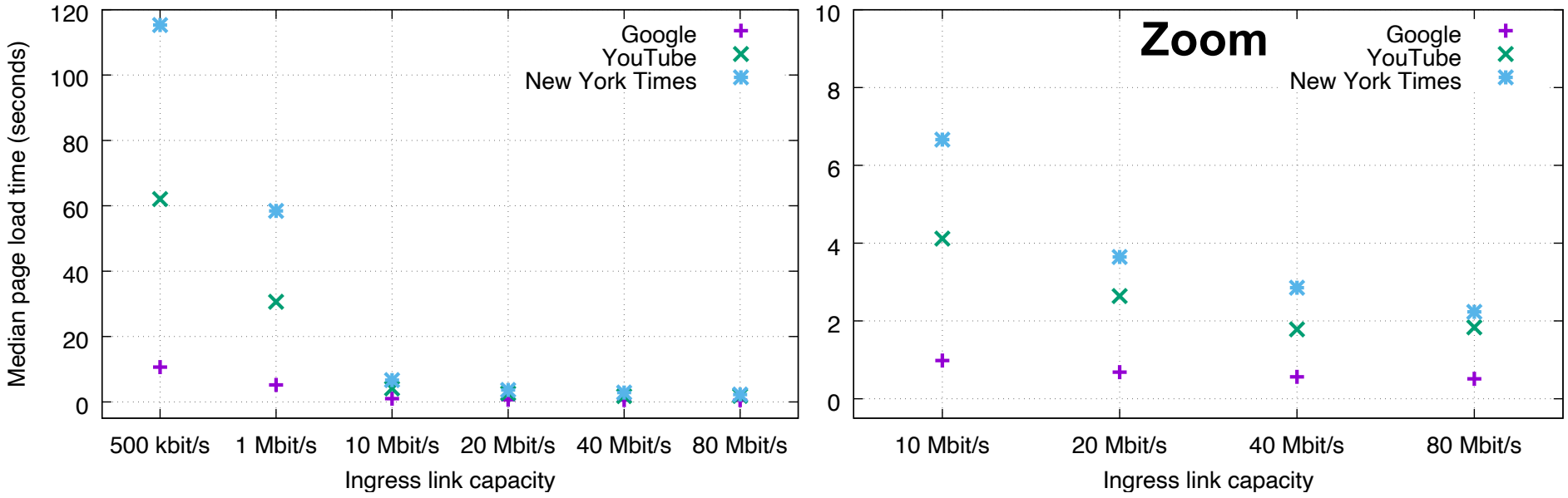
doi: 10.23919/TMA.2019.8784609.

What are the bandwidth limits of video?

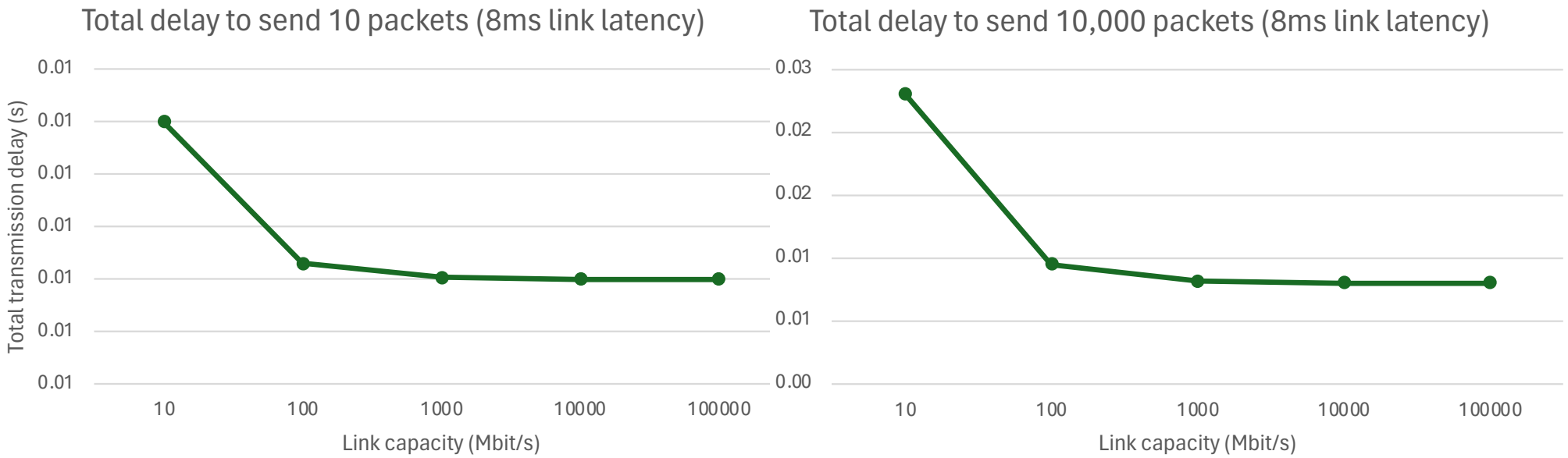
- Netflix: "The fixed ladder has predetermined 4K bitrates — 8, 10, 12 and 16 Mbps — it deterministically maxes out at 16 Mbps." https://netflixtechblog.com/all-of-netflixs-hdr-video-streaming-is-now-dynamically-optimized-e9e0cb15f2ba
  - 10 users watching 4K video (big screens!) in parallel: 160 Mbps
  - Today, this is not much! **5G**: designed for peak data rates up to **20 Gbps**
  - **WiFi 7**: theoretical max **40 Gbps.** **WiFi 6**: "only" **9.6 Gbps**
- Interactive: typically delay-based congestion control, avoiding to exceed the capacity limit
  - E.g., IETF “RMCAT” (Real-Time Media Congestion Avoidance Techniques) controls, designed for WebRTC (GCC, NADA, SCReAM, ..)
- (When) will future applications really exceed today's limits?

Capacity upgrade = higher "speed"? An experiment

- Median (20 tests): Selenium controlling Chrome, “complete page load”, no caching, from a client @ Univ. Oslo, connected with a 100 Mbps link
- ping to google.com and youtube.com: ~ 7.6 ms, nytimes.com ~ 0.5 ms
- Ingress capacity limited with token bucket: no buffer, just rate tolerance
  - Not very realistic, but overall tendency will be the same in a real network



Limited gains



- Capacity upgrades have a limited "speed" benefit
  - **Benefit is: same delay as before, but can send more data**

  => gain factor X requires data use growth factor X

- Total mobile traffic grows exponentially... but what about per-application traffic? Traffic / RTT is limited by TCP or QUIC.

1. **Physical capacity is not the bottleneck. (anymore)**

2. **Capacity upgrades hardly improve latency.**

3. **Applications do not demand more capacity.**
   - They don't even use anything close to what we have today.
   - They work with "best effort", so they live with what TCP (or QUIC) gives them.
   - Go on, dream of holograms... but applications do lag behind.
   - 7G, 8G, 9G, 10G will have higher and higher capacities, emptier and emptier...
   - How about taking a break with 7G, and letting apps catch up?

There are many arguments for why underload exists

Lloyd Brown, Yash Kothari, Akshay Narayan, Arvind Krishnamurthy, Aurojit Panda, Justine Sherry, and Scott Shenker. 2023. How I Learned to Stop Worrying About CCA Contention. In Proceedings of the 22nd ACM Workshop on Hot Topics in Networks (HotNets '23). https://doi.org/10.1145/3626111.3628204

- ... and for why congestion control will become increasingly worse at saturating the capacity !
  - No, we can't just "scale" it forever (e.g., scale IW): it's a deeper architectural problem

Michael Welzl, Peyman Teymoori, Safiqul Islam, David Hutchison, Stein Gjessing: "Future Internet Congestion Control: The Diminishing Feedback Problem", IEEE Communications Magazine, 2022.
https://doi.org/10.1109/MCOM.006.2200008

... and many questions

- Newer equipment is more energy efficient
- Life-Cycle Assessment (LCA) should dictate upgrading
  - **But: are networks upgraded following LCA?**
  - **How to find out?**

- If I'm right, and we should do fewer hardware updates:
  - **How to convince the industry?**
  - Surely "stop upgrading capacity" will get a lot of love from lower-layer folks (massive MIMO etc)
  - **How to convince funders to put money into this?**
    (e.g., re-architecting congestion control for higher speed)
  - **How to measure the sustainability gain?**