UiO : Faculty of Mathematics and Natural Sciences
University of Oslo

Department of Informatics
Networks and Distributed Systems (ND) group

Some Observations On Inter-DC Congestion Control

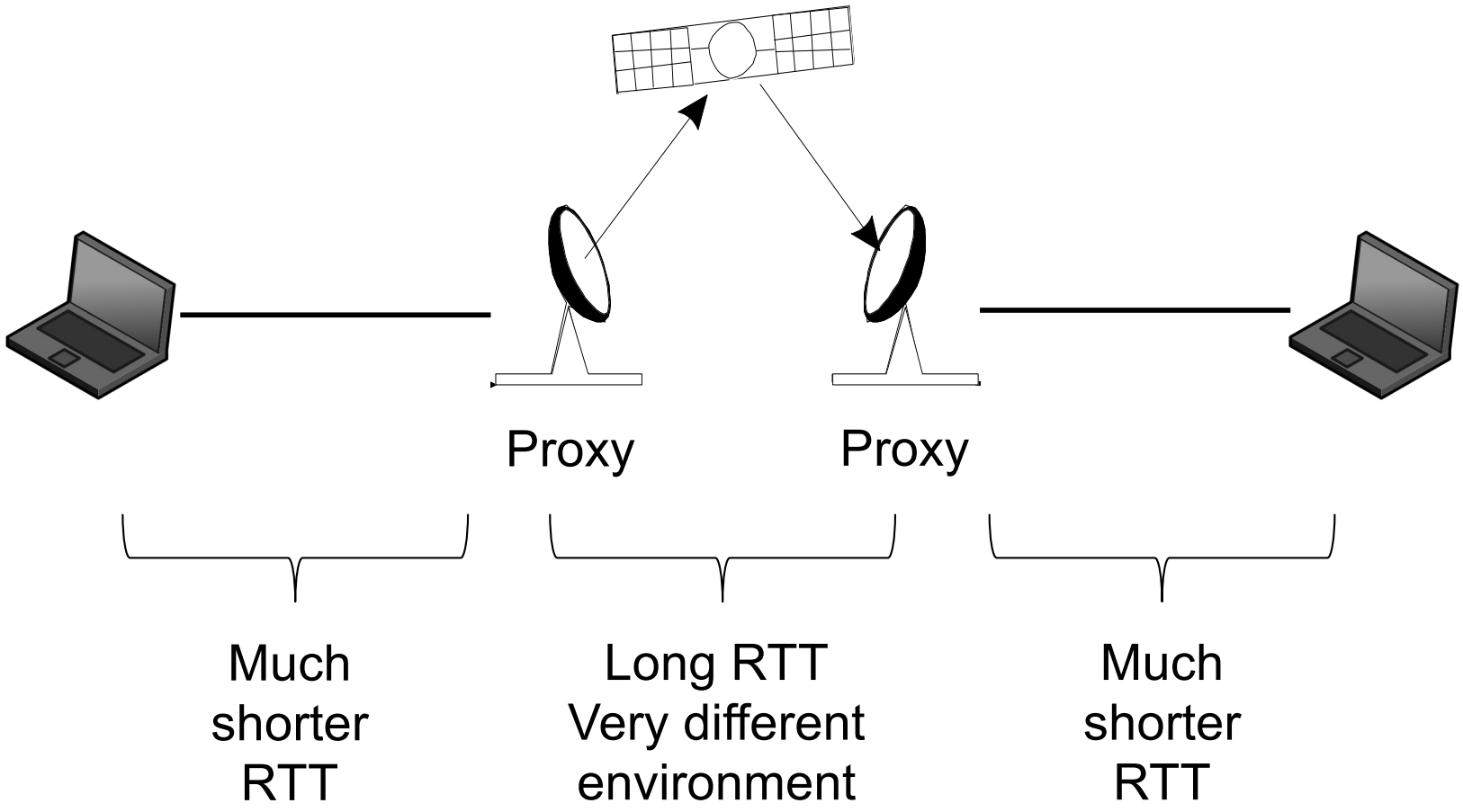
Michael Welzl

Inter-DC side meeting
IETF-120
25.7.2024

Disclaimer

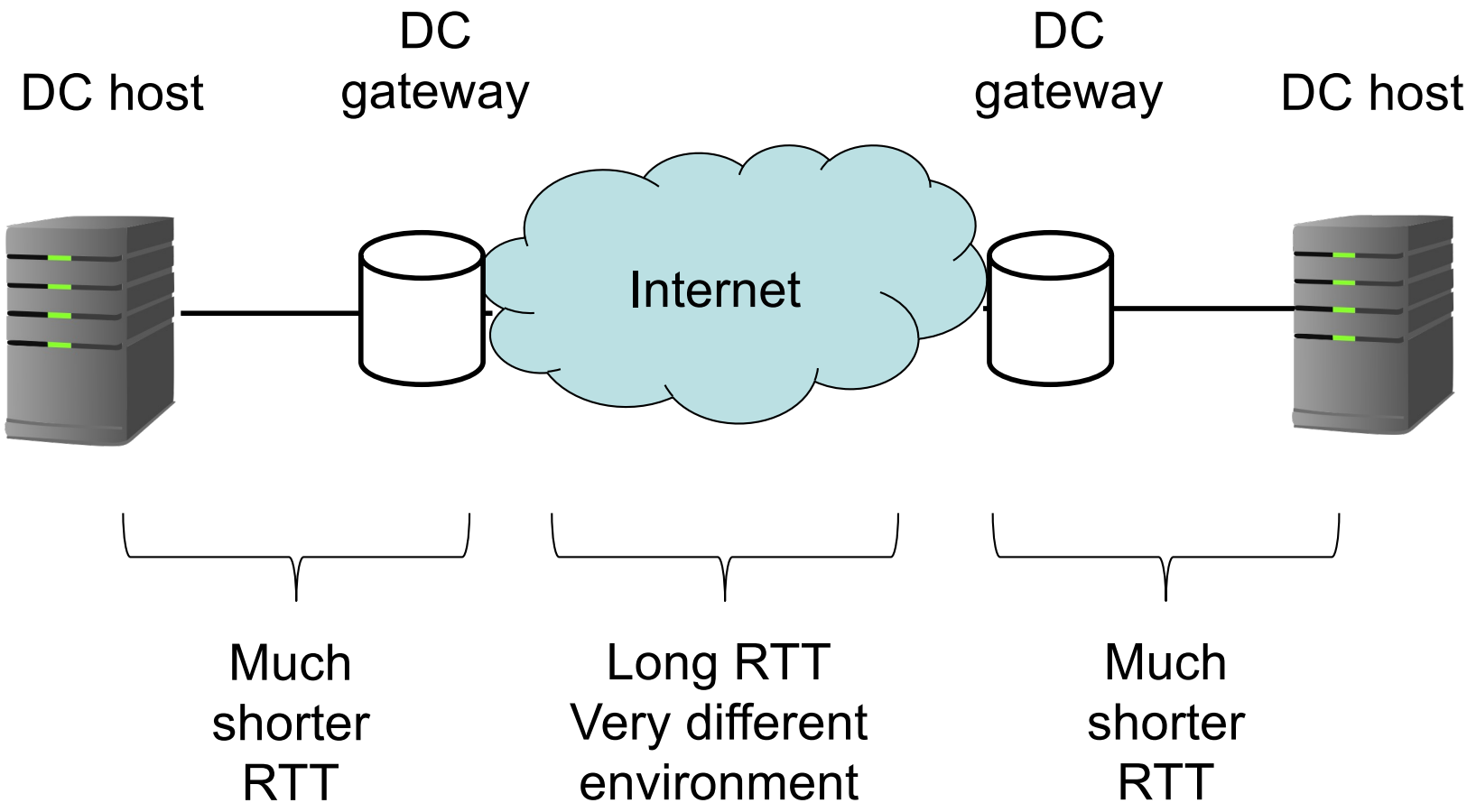
- I know nothing about the industry's practice
  - Maybe people do better things than I say here.
  - ...or maybe worse.
  - Or maybe they implement my suggestions?
- Would like to learn, any pointers are welcome!

TCP over Satellite



Common to implement satellite-specific transport / CC between proxies.

Inter-DC communication



Having proxies at DC gateways, and doing something different between them, probably makes sense here, too!

Without proxies

- Intra-DC congestion control (CC) has been heavily optimized and is not compatible with Internet CC
  - An all-encompassing e2e compromise-loop is likely not very efficient
- E2E connections compete with each other, fairness is the result of CC competition
  - Not only limited control of fairness, but also competition side-effects: queuing delay, packet loss

With proxies

- QUIC (and before it, SCTP) has taught us about the benefits of multi-streaming

- So, it could be best to map all inter-DC traffic on one connection
  - Allows to share congestion control state
  - Allows to precisely prioritize flows
  - Can be a massive benefit for short flows (get a share of an already large cwnd): *stay tuned for two more slides*
  - But there are challenges too...

Challenge #1: QUIC

- PEP-style QUIC proxying is doable, but requires changes to QUIC
  - Or using MASQUE

- Change QUIC to delegate security context to proxy

M. Kosek, B. Spies and J. Ott, "Secure Middlebox-Assisted QUIC," 2023 IFIP Networking Conference (IFIP Networking), Barcelona, Spain, 2023, pp. 1-9.

- Minimally change QUIC to use a service offered by a protocol-independent daemon or library that ACKs packets without decrypting them

Gina Yuan, Matthew Sotoudeh, David K. Zhang, Michael Welzl, David Mazières, Keith Winstein: "Sidekick: In-Network Assistance for Secure End-to-End Transport Protocols", Usenix NSDI '24, Santa Clara, CA, USA, 16-18 April 2024. ***Outstanding paper award & Community award***

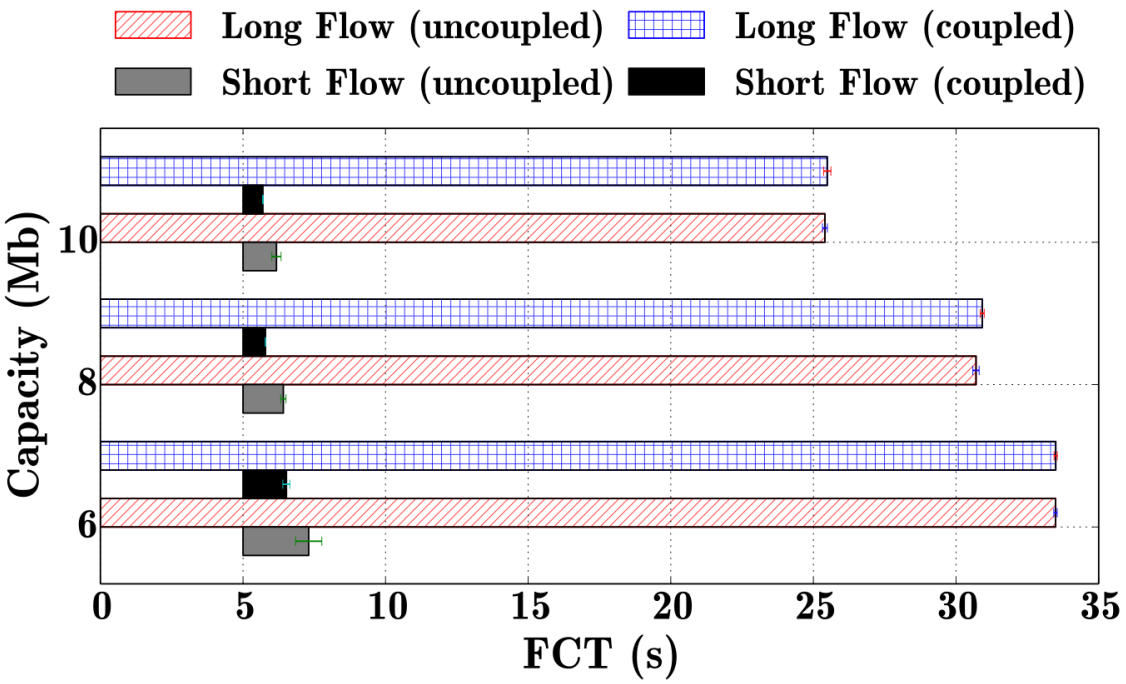
Challenge #2: many DC connections

- The single connection should be as aggressive as its muxed streams would be if they were e2e connections
  - In principle also true for QUIC... for sure needed with very many flows. Some cc. research needed (in the style of "MulTCP")

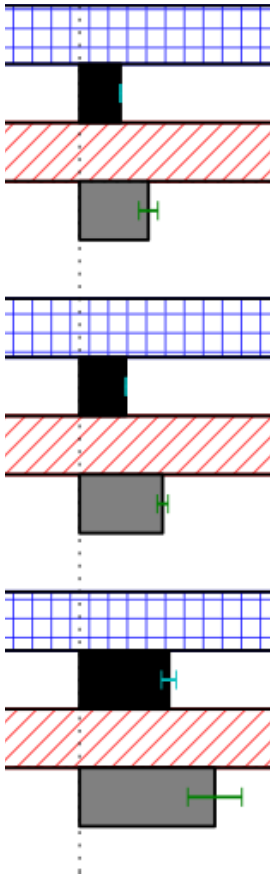
- Connection-splitting for many flows, mapping them onto streams, can be resource intensive
  - (Potentially large) per-flow buffers needed
  - To get the congestion control and prioritization benefits, more lightweight single-path congestion control is also possible
    - Similar to RFC 8699 (a lightweight emulation of the Congestion Manager (RFC 3124)), but this was for RMCAT flows
    - Works well for many controls, even a heterogeneous mix, and TCP too: Safiqul Islam's Ph.D. thesis

Benefits from sharing cc. information with other flows

Long TCP flow (25Mb) joined by a short flow (200 Kb) after 5 seconds



Zoom



Safiqul Islam, Michael Welzl, Kristian Hiorth, David Hayes, Grenville Armitage, Stein Gjessing: "ctrlTCP: Reducing Latency through Coupled, Heterogeneous Multi-Flow TCP Congestion Control", IEEE INFOCOM Global Internet Symposium (GI) workshop (GI 2018), Honolulu, HI, April 2018. DOI 10.1109/INFCOMW.2018.8406887

Best of workshop presentation award

Challenge #3: load balancing

- How should the transport header between the gateways look?
- Single 5-tuple: one path, possible to fully benefit from single-path CC coupling and prioritization
  - Cannot be applied across different 5-tuples
- Multiple 5-tuples: benefit from ECMP
  - Could actively control this: measure, find out how many 5–tuples to use; then, apply single-path CC. coupling per 5-tuple
  - Needs a strategy for flow mapping: e.g., good to use one 5-tuple when assigning priorities between a group of flows or wanting to avoid large RTT differences
  - Alternative: multipath CC., but then single-path CC. coupling can only work on subflows; find a compromise?

Summary: IETF perspective

- Proxying QUIC: needs changes to QUIC
- Some CC. work: as always, IETF not strictly needed for deployment, but there can still be interoperability problems, so the IETF generally wants standards
  - Single connection carrying many flows (being similarly aggressive) needs new CC. work, applicable to QUIC too
  - General & lightweight single-path congestion control coupling
- 5-tuple logic: unsure if IETF involvement is needed

Thank you!

Questions?