Evaluation of TCP Congestion Control for Public High-Performance Wide-Area Networks

Fatih Berkay Sarpkaya[1], Andrea Francini[2], Bilgehan Erman[2], Shivendra Panwar[1]

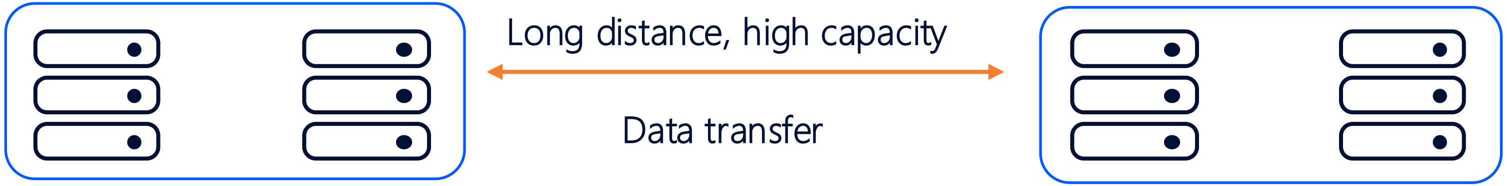
[1]NYU Tandon School of Engineering, Brooklyn, NY, USA & [2]Nokia Bell Labs, Murray Hill, NJ, USA

IEEE International Conference on High-Performance Switching and Routing (HPSR) 2026

NYU TANDON SCHOOL OF ENGINEERING

NOKIA BELL LABS

Our interest: Deterministic transfer of large data files



- Compute resources are tightly scheduled and expensive → Accurate prediction of ***Flow completion time (FCT)*** is essential

- The maximum FCT value is the defining metric for the file transfer contribution to the overall length of the HPC reservation.

Use case: Multiple sites

- Long-Distance inter-site (d > 1000 km)
  - *HPC for scientific research*
  - *Distributed storage*
  - *Multimedia content production*
  - *Data backup and disaster recovery*

High-Performance WAN (HP-WAN)

Our use-case: “Public” High-Performance Wide Area Networks

- HP-WANs such as ESnet, GEANT, and JANET now support broader communities
- This creates a new model: the data-transfer owner uses the network but does not control the full end-to-end path

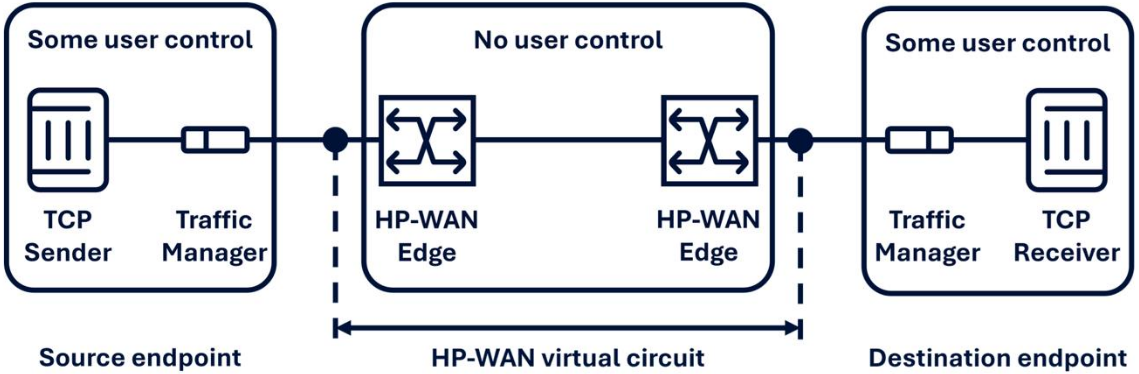


Figure: Logical representation of the massive-transfer data path.

Achieving predictable FCTs is a significant challenge

Accurate estimation of the FCT requires:

- Known and stable path capacity
    - A fixed-bandwidth virtual circuit (VC) provides traffic isolation through network traffic management mechanisms.
- Reliable end-to-end transport (i.e TCP), and a congestion control (CC) algorithm with stable throughput
    - Alternatives to TCP exist but not better suited to the public HP-WAN use case that we are considering (i.e RDMA over Converged Ethernet (RoCEv2)).

Prior Work does not cover this specific scenario

- Prior work:
    - HP-WAN infrastructures
    - TCP/RDMA in private HP-WANs
    - TCP fairness and coexistence
- This work:
    - Public HP-WANs
    - Bandwidth-isolated VCs
    - Limited end-to-end path control
    - Focus on FCT efficiency and predictability
    - Compare TCP CC + ingress data-path configurations

FABRIC Experiment Design (1/2)

- L2PTP virtual circuits with guaranteed bandwidth between SmartNICs
- Ingress policers enforce each VC's reserved rate
- L2PTP traffic mapped to FABRIC's high-priority EF queue.
- 100 Gb/s inter-site links → 80 Gb/s reservable capacity → 4 × 20 Gb/s VCs
- RTTs across VC pairs: 7–16 ms

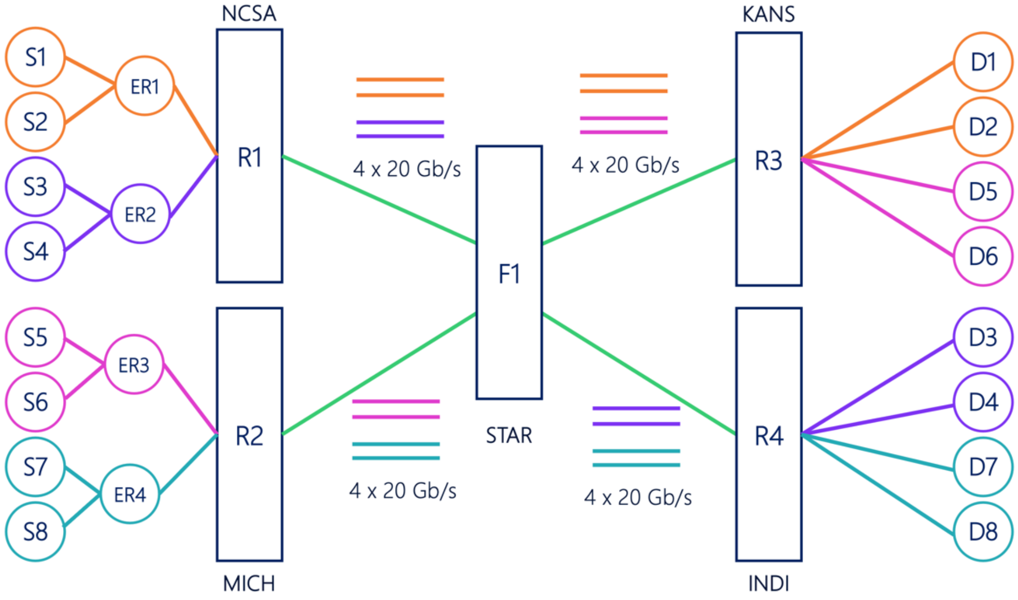


Figure: Experiment topology

FABRIC Experiment Design (2/2)

- Source-side edge-router VMs for optional per-VC shaping (DPDK)
- Bandwidth isolation only between VC endpoints
- Sender-to-router and receiver-side access segments not prioritized

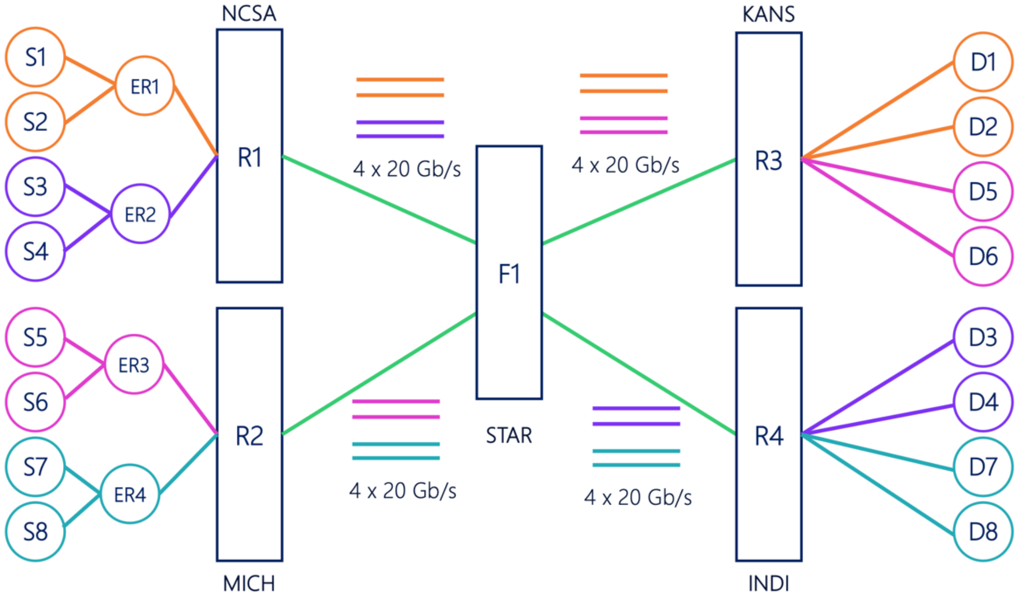


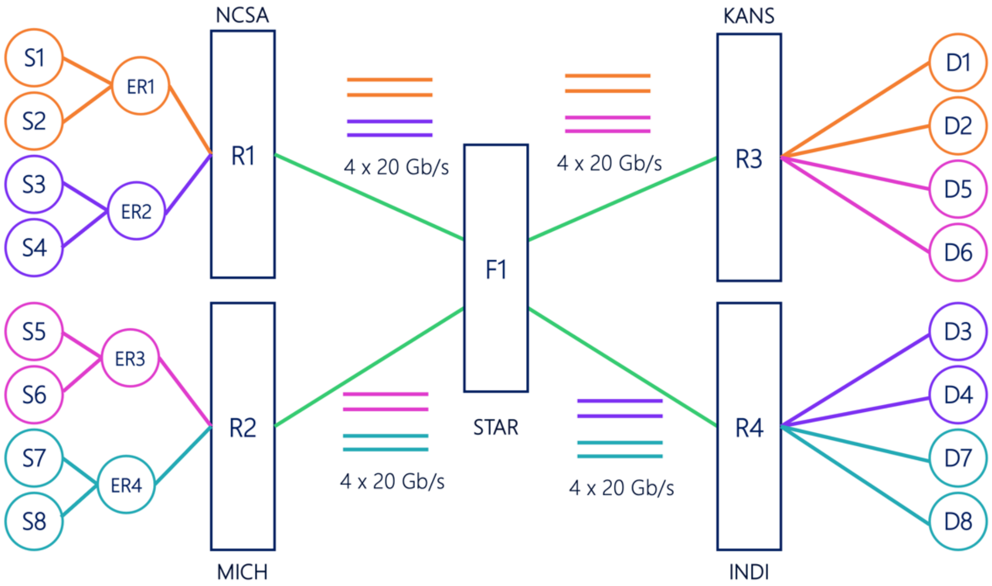
Figure: Experiment topology

TCP Congestion Control Algorithms

- We evaluate only congestion control (CC) algorithms designed for high-throughput bulk transfers.
- Specifically, we consider:
  - TCP CUBIC → the dominant CC algorithm in today's Internet
  - TCP BBRv1 → model-based, widely deployed in practice
  - TCP BBRv3 → currently used for Google's internal WAN traffic as well as google.com and YouTube services.

Experiment Details / Metric

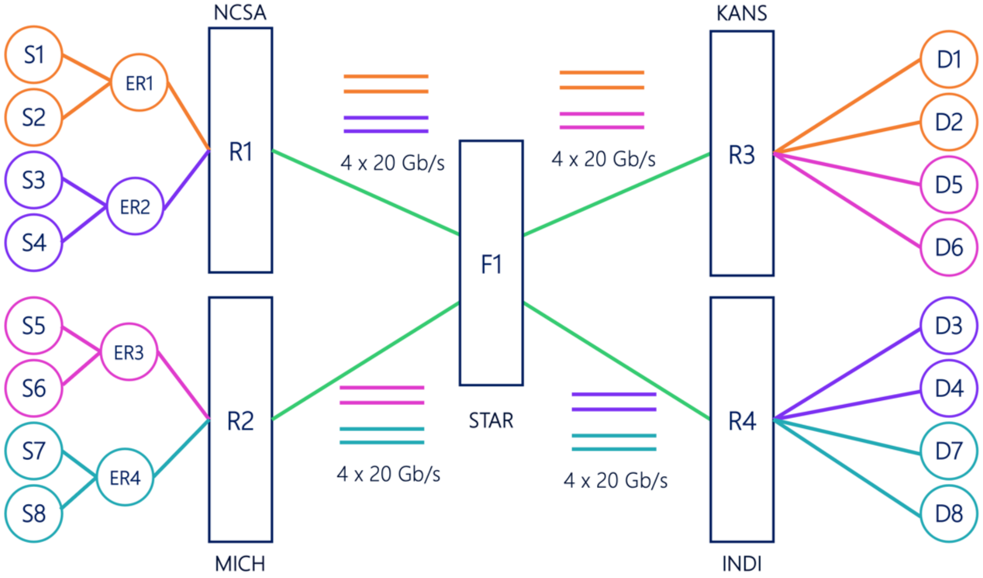
- 8 concurrent VCs, each at 20 Gb/s
- 1 TB transfer per VC
  - Ideal FCT ≈ 8 minutes
- 1 vs. 10 parallel TCP flows
  - Single-flow vs. striped transfer
- Metric: FCT efficiency → $\frac{ideal_{fct}}{measured_{fct}}$
- 20 independent runs per scenario at different times



Data-Path Configurations

	ER1–4 (sender-side edge routers)	R1–4 (routers at VC endpoints)
Linux forwarding	Linux kernel forwarding, no shaping	Linux kernel forwarding, no shaping
DPDK forwarding	DPDK fast forwarding, no shaping	DPDK fast forwarding, no shaping
DPDK shaping	DPDK fast forwarding, per-VC shaping	DPDK fast forwarding, no shaping

Table: Summary of the three data-path configurations used in the experiments



Experiment Results

- Which combination of data-path configuration and CC algorithm provides the most efficient and predictable massive data transfers over public HP-WAN VCs?

System Performance Limits

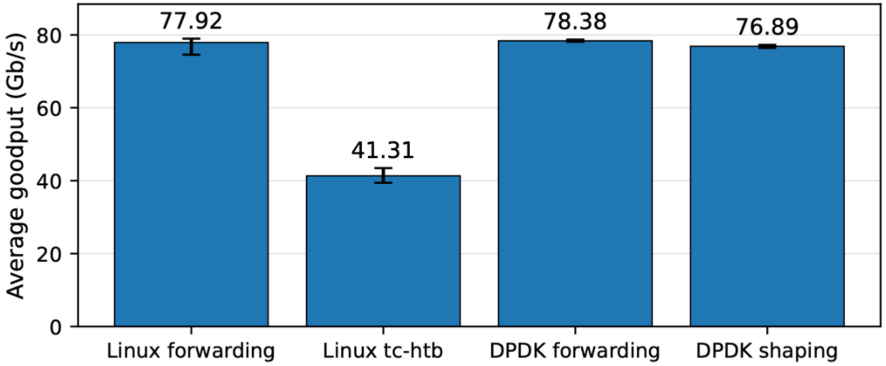


Figure: Goodput performance of router VM configuration options. (10 BBRv1 flows) (80 Gb/s tunnels)

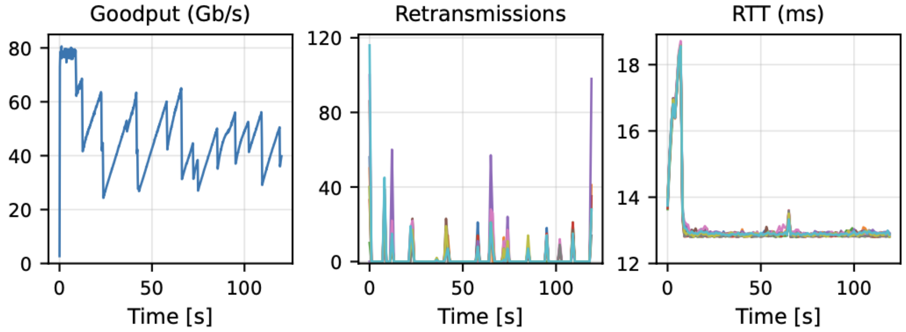


Figure: 10 TCP CUBIC flows with DPDK shaping at 80 Gb/s.

DPDK shaping is the best data-path configuration – FCT Efficiency

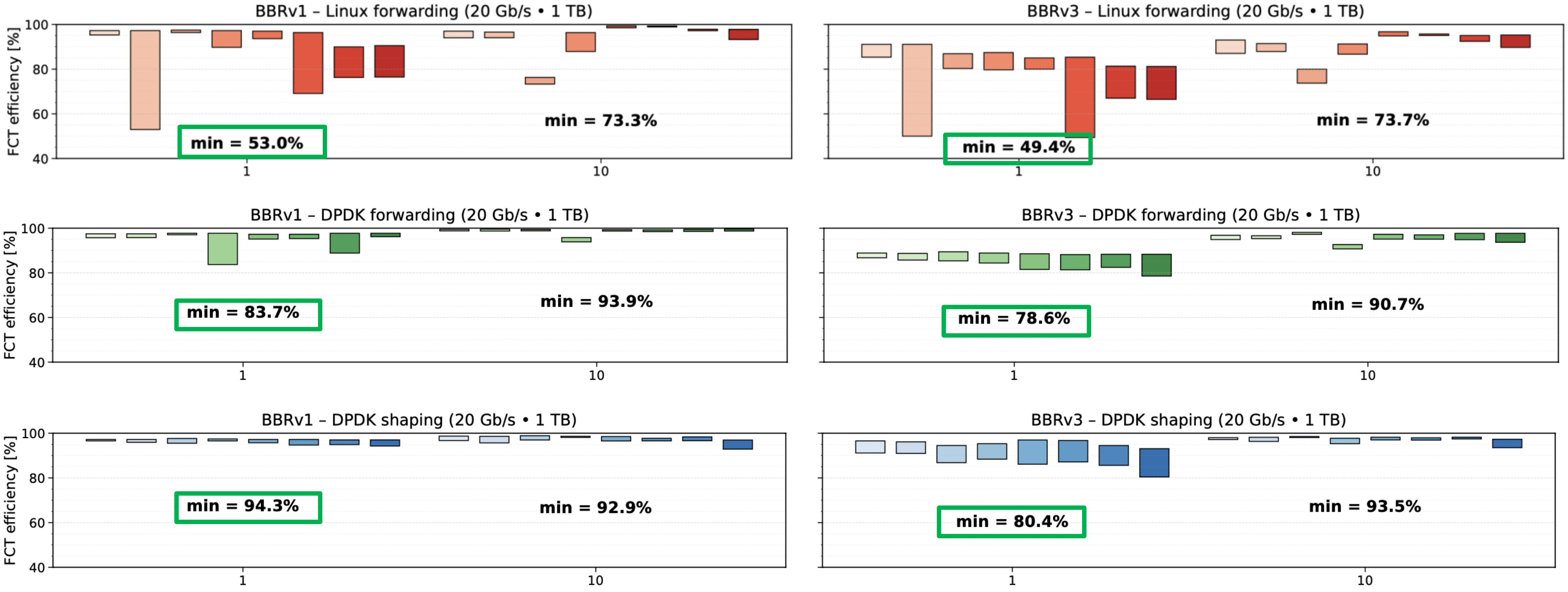


Figure: Bar edges show the min–max values across 20 trials for each sender. Annotations report the minimum FCT efficiency, aggregated over the full set of VCs in each scenario.

DPDK shaping is the best data-path configuration – Overhead

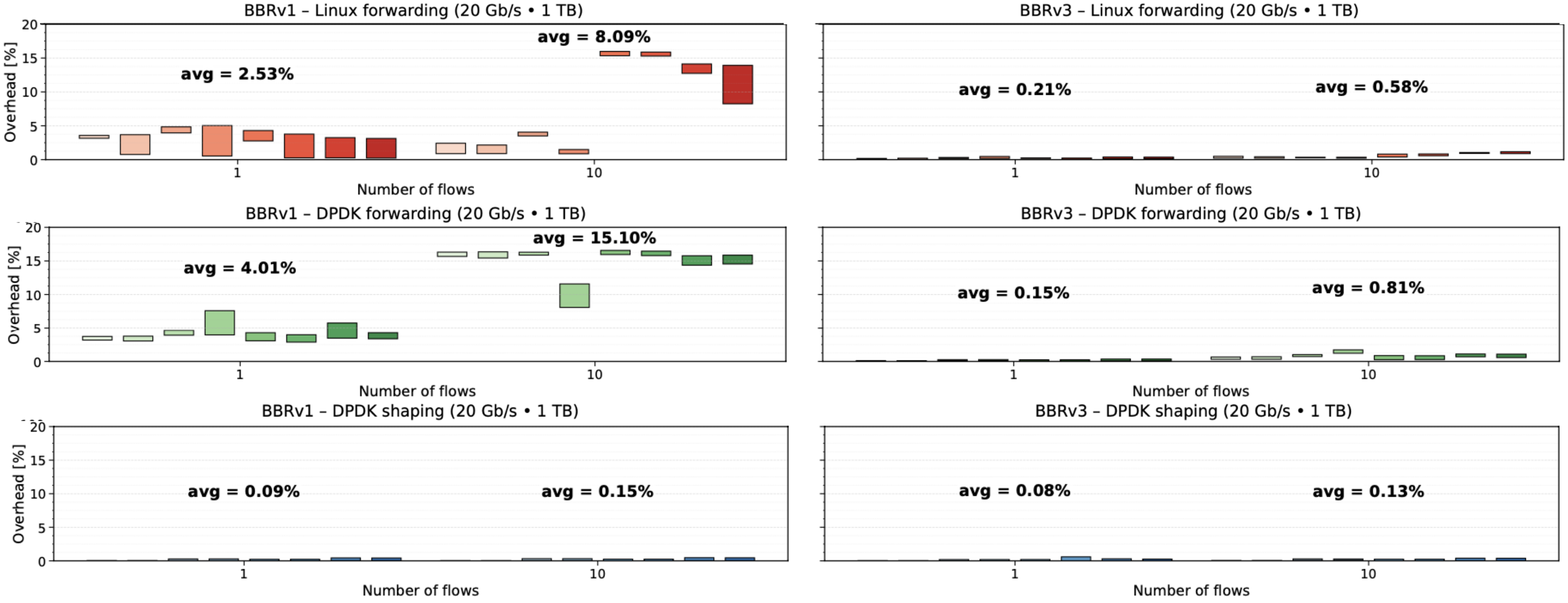
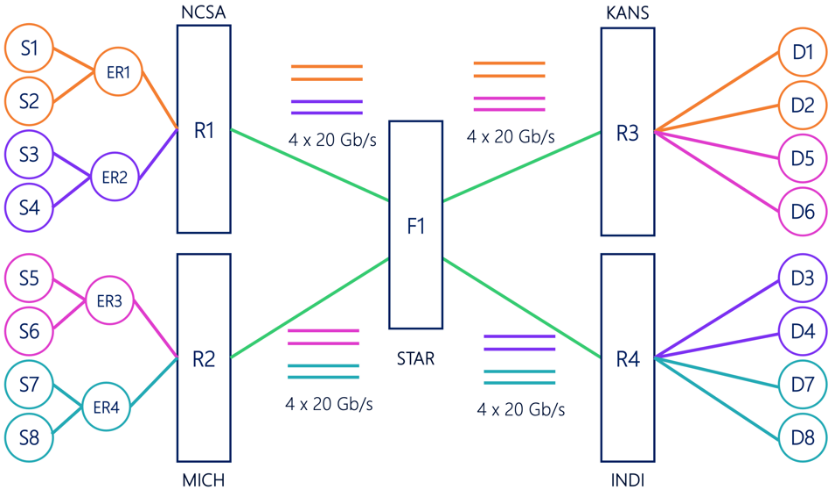


Figure: Bar edges show the min–max values across 20 trials for each sender. Annotations report the average overhead, aggregated over the full set of VCs in each scenario.

Non-congestive packet losses are persistent in our FABRIC setup

- Several alternative approaches to eliminate non-congestive losses:
    - per-VC shapers on R1–4
    - co-locating shapers with the senders
- Most losses likely occur downstream of the VC egress, close to the TCP receivers.
- Loss-robust CC improves predictability of massive transfers in public HP-WANs
- BBRv1 should be confined to isolated VCs due to poor coexistence with loss-based TCP



Within the tested bandwidth range, single-flow transfers are preferable

- BBRv1 + DPDK shaping: single-flow and 10-flow transfers perform similarly.
- Iperf3 emulates an ideal TCP striping application for packet distribution and reordering across parallel TCP flows.
- Practical striping applications may add coordination/reordering overhead that can degrade FCT, making single-flow transfers preferable for VC bandwidths up to at least 20 Gb/s.

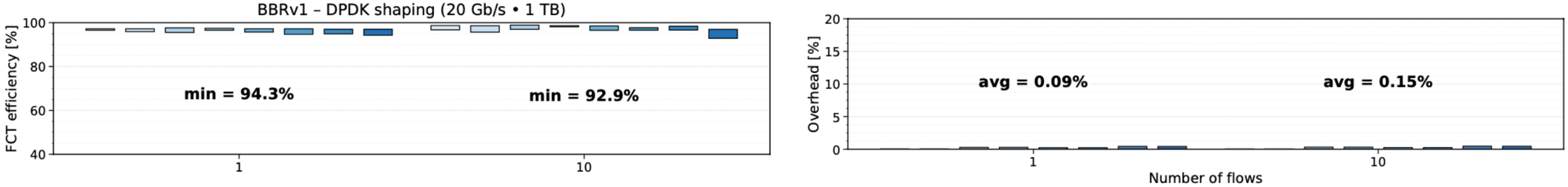


Figure: Bar edges show the min–max values across 20 trials for each sender. Annotations report the minimum FCT efficiency and the average overhead, aggregated over the full set of VCs in each scenario.

Conclusion

- We studied the completion-time efficiency and predictability of prominent TCP CC algorithms for massive data transfers over public HP-WANs.
- Using FABRIC L2PTP VCs, we compared Linux forwarding, DPDK forwarding, and DPDK shaping, and found that per-VC shaping improves FCT predictability with minimal retransmission overhead.
- Combined with per-VC shaping, BBRv1 provides more efficient and predictable file transfers than BBRv3 and CUBIC.
- Future work: isolate or prioritize VC traffic outside the HP-WAN tunnel to reduce cross-traffic interference near end-points.

Thank You for Listening !

Experiment Artifacts:

https://artifacts.fabric-testbed.net/artifacts/

Artifact Title: Evaluation of TCP Congestion Control for Public High-Performance Wide-Area Networks

Contact:

Fatih Berkay Sarpkaya: fbs6417@nyu.edu
Andrea Francini: andrea.francini@nokia-bell-labs.com