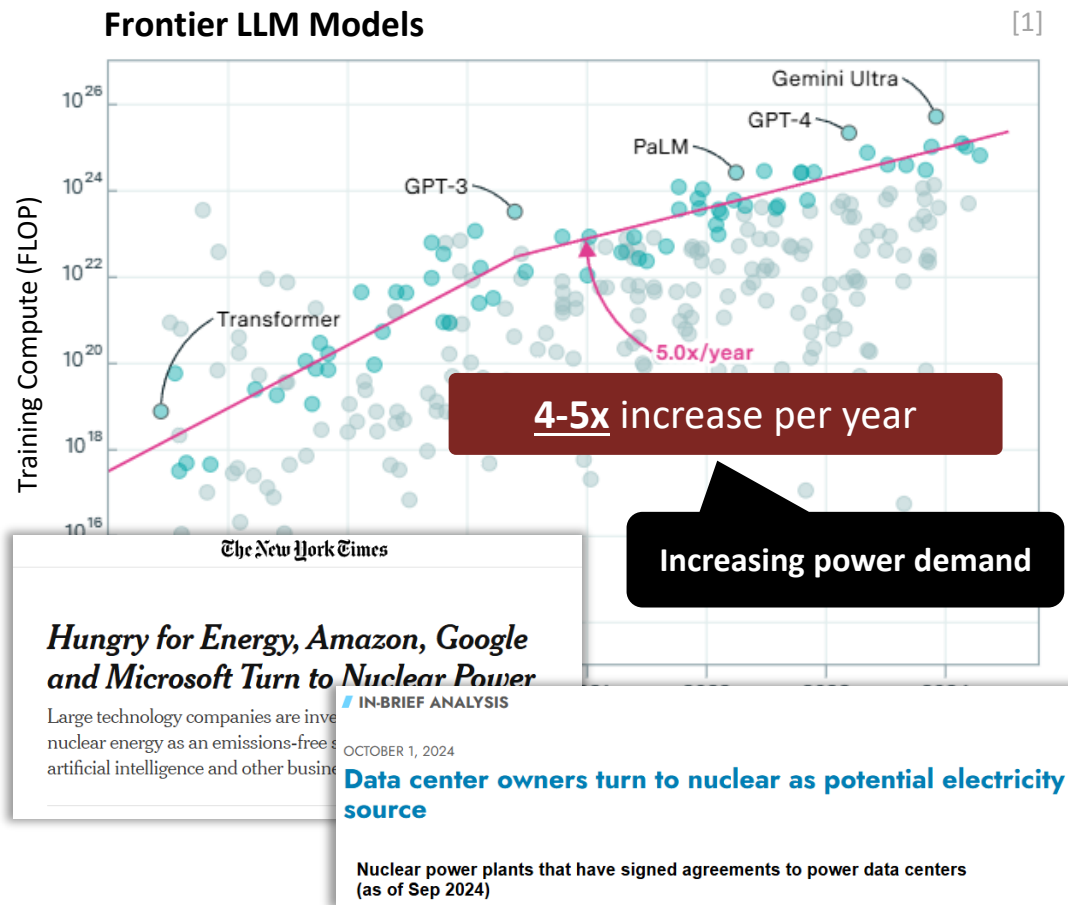


Uno: A One-Stop Solution for Inter- and Intra-Data Center Congestion Control and Reliable Connectivity

Tommaso Bonato* ^{1,3}, Sepehr Abdous* ^{2,3}, Abdul Kabbani ³, Ahmad Ghalayini ³, Nadeen Gebara ³, Terry Lam ³, Anup Agarwal ^{3,5}, Tiancheng Chen ¹, Zhuolong Yu ³, Konstantin Taranov ³, Mahmoud Elhaddad ³, Daniele De Sensi ⁴, Soudeh Ghorbani ², Torsten Hoefler ¹



Motivation



HOME > NEWS > IT HARDWARE & SEMICONDUCTORS

Training Google's Gemini Ultra data centers, and rising power demand

Tech Industry > Artificial Intelligence

What Google is doing to train Gemini Ultra

December 4, 2024

News By Anton Shilov published October 18, 2024

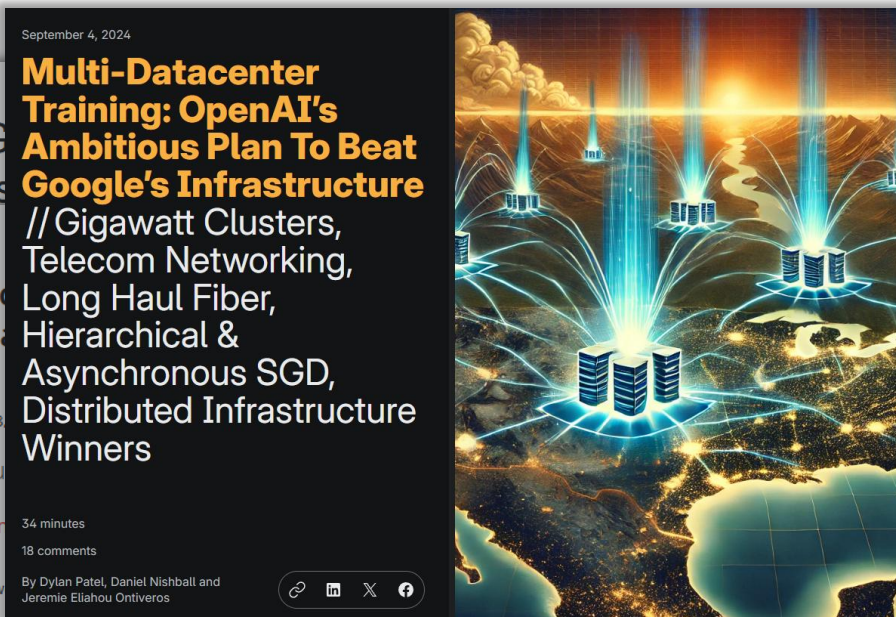
Microsoft Azure CTO: AI training is needed to approach power grid

While others seek a nuclear solution, Google is looking for a different approach.

34 minutes, 18 comments

By Dylan Patel, Daniel Nishball and Jeremie Eliahou Ontiveros

When you purchase through links on our site, we may earn a commission.



"We had to scale to more compute, and that compute was not available as part of one cluster. We had to go to multi-cluster training"

GPT4.5 Team
April 2025

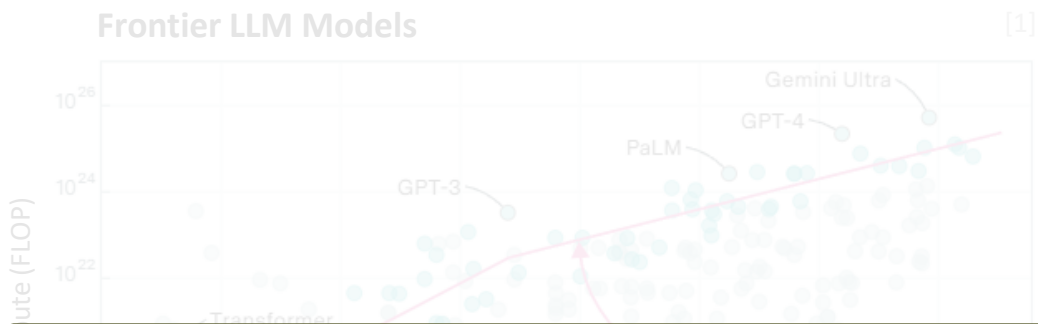
"We don't disclose exactly the details of how many locations but Gemini Ultra was trained across multiple sites, and multiple clusters within those sites."

Thomas Kurian, CEO Google Cloud
December 2023

Problem: local power limitation, increased complexity

Solution: training across multiple datacenters

Motivation



HOME > NEWS > IT HARDWARE & SEMICONDUCTORS

Training Google's Gemini data centers, and rising power demand

Tech Industry > Artificial Intelligence

What Google's Gemini data centers, and rising power demand

December 18, 2024

News By Anton Shilov published October 18, 2024

September 4, 2024

Multi-Datcenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure

// Gigawatt Clusters, Telecom Networking, Long Haul Fiber, Hierarchical & Asynchronous SGD, Distributed Infrastructure

Winners



Significant inter- and intra-DC traffic co-existence inside DCs
Efficient communication becomes challenging

The New York Times

Hungry for Energy, Amazon, Google and Microsoft Turn to Nuclear Power

IN-BRIEF ANALYSIS

OCTOBER 1, 2024

Data center owners turn to nuclear as potential electricity source

Increasing power demand

Nuclear power plants that have signed agreements to power data centers (as of Sep 2024)

"We had to scale to more compute, and that compute was not available as part of one cluster. We had to go to multi-cluster training"

GPT4.5 Team
April 2025

"We don't disclose exactly the details of how many locations but Gemini Ultra was trained across multiple sites, and multiple clusters within those sites."

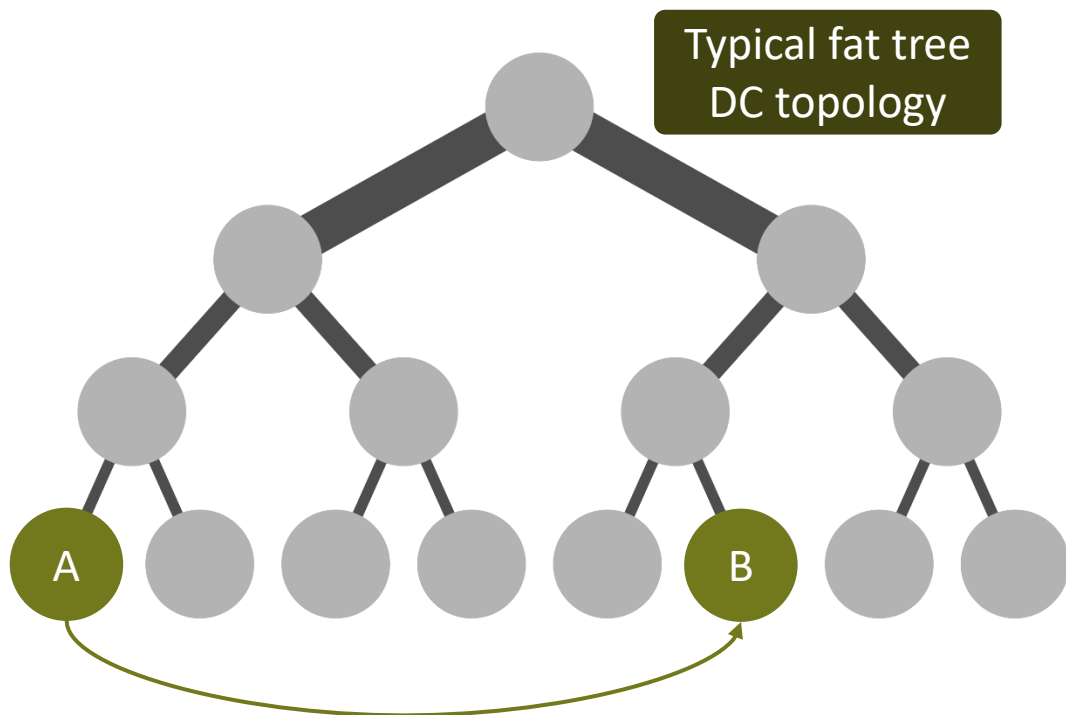
Thomas Kurian, CEO Google Cloud
December 2023

Problem: local power limitation, increased complexity

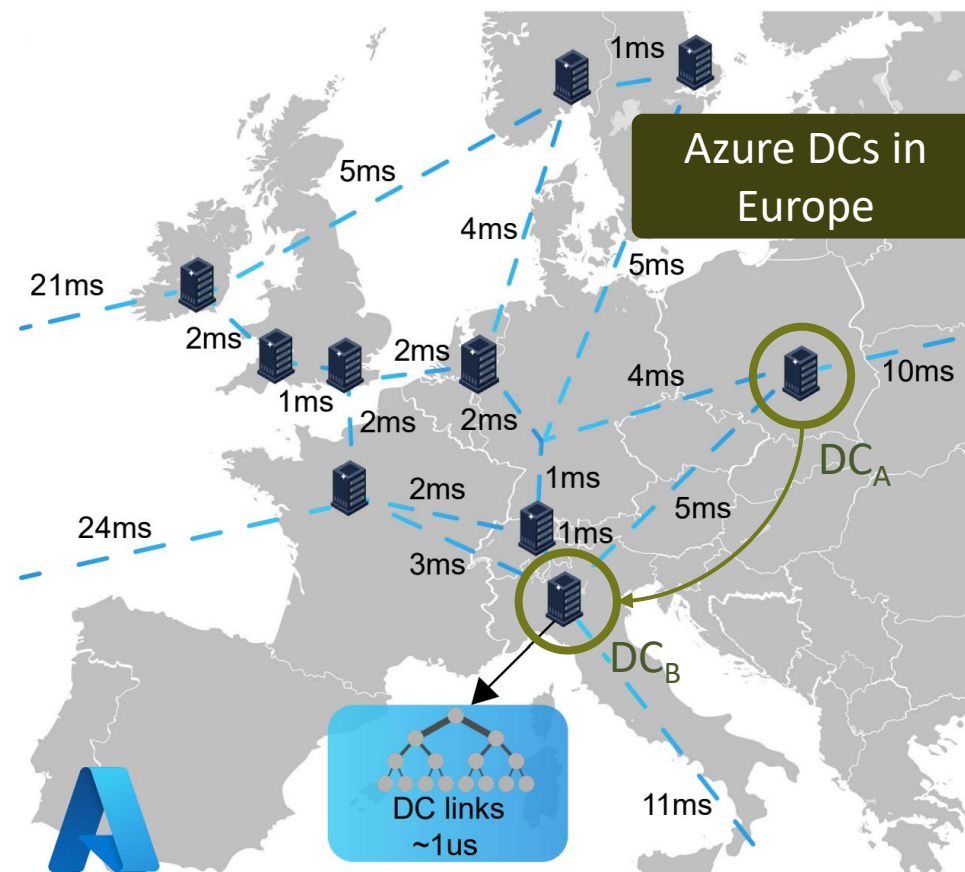
Solution: training across multiple datacenters

[1] Training Compute of Frontier AI Models Grows by 4-5x per Year, Report, [Epoch AI, 2024](#)

Motivation



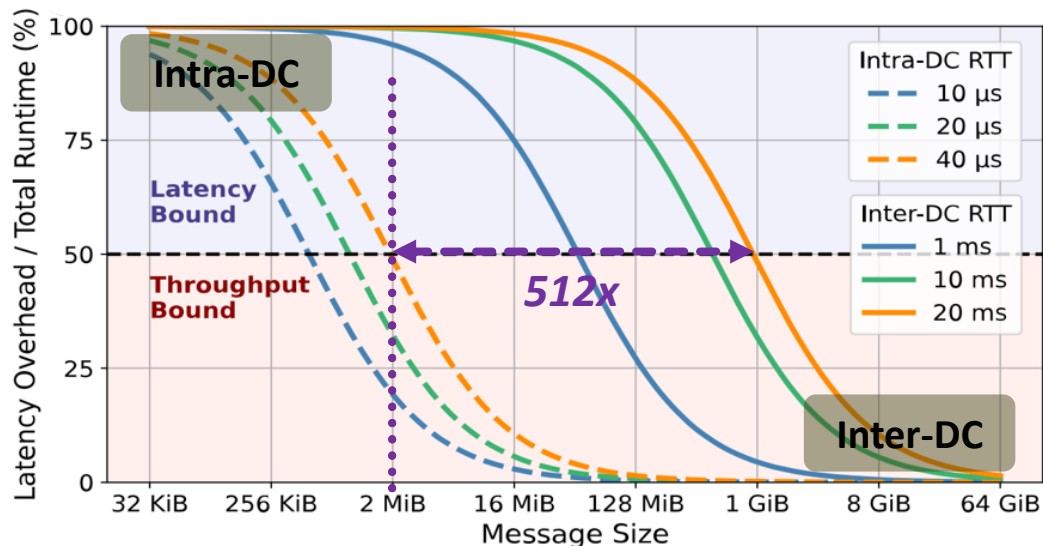
Short Cables (<100m) → ~1 μ s latency
RTT from A to B → ~12 μ s



Long cables (>100km) → >1ms latency
RTT from DC_A to DC_B → 10ms

1000x latency difference

Inter- and intra-DC RTTs have huge gaps



Slow loss detection and recovery

>1 ms RTT across datacenters

BDP <-> Buffer sizes mismatch

~ 100 MB buffering **per port** with 10 ms inter-DC RTT

Unfairness / slow convergence to fairness

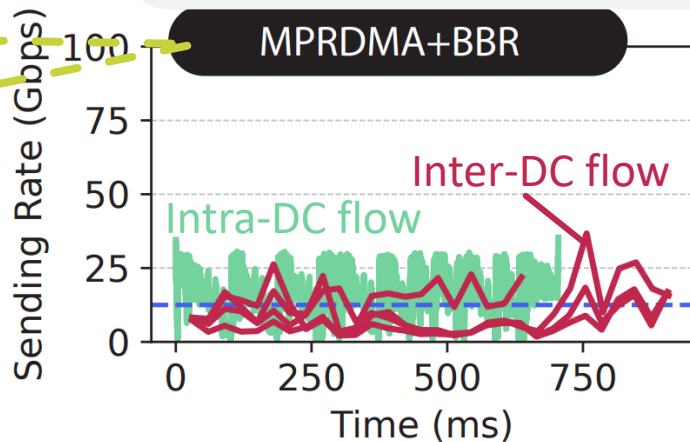
Huge gap in congestion notification and reaction granularity
Existing solutions fall short in **efficiently** converging to fairness

8:1 incast from senders both inside and across datacenters

Separate control loops

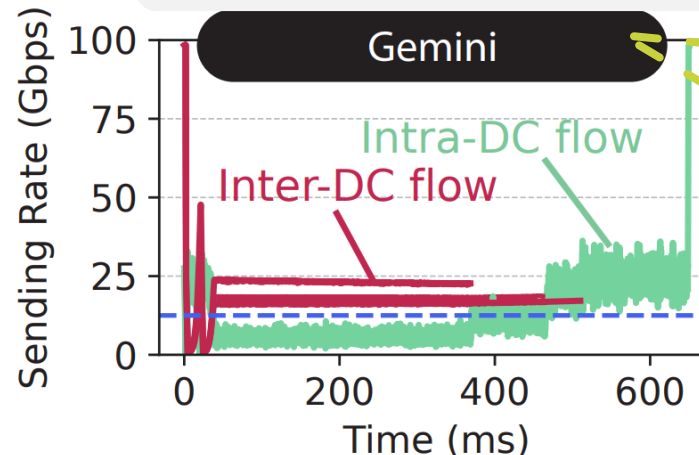
Unified control loops

Reaching
guaranteed
fairness is
challenging



[BBR: Congestion-based congestion control., Commun. ACM 2017]

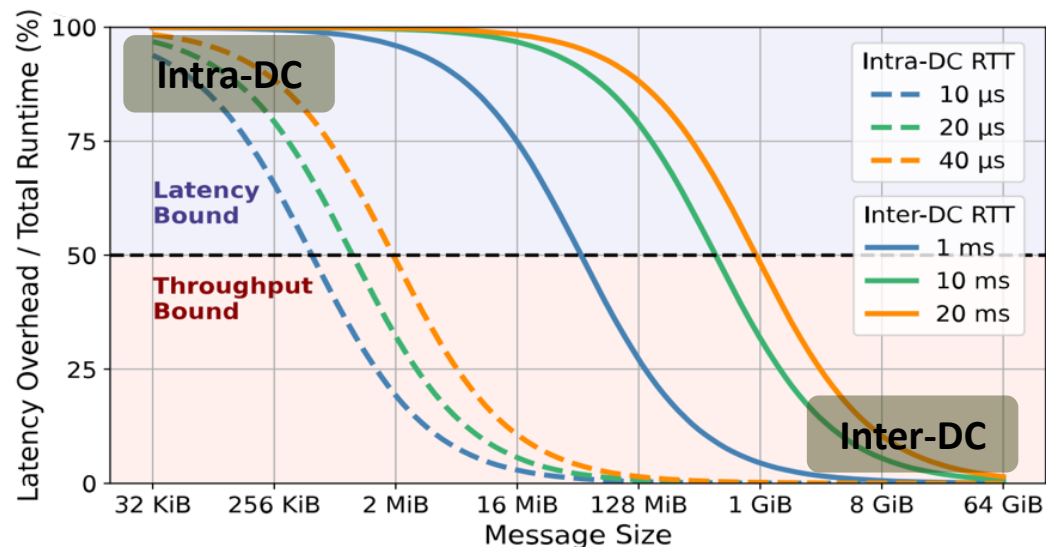
[Multi-Path transport for RDMA in datacenters, NSDI 2018]



[Congestion Control for Cross-Datacenter Networks, ToN 2022]

Reaches
fairness but
convergences
is **very slow**

Inter- and intra-DC RTTs have huge gaps



Slow loss detection and recovery

>1 ms RTT across datacenters

BDP <-> Buffer sizes mismatch

~ 100 MB buffering **per port** with 10 ms inter-DC RTT

Unfairness / slow convergence to fairness

Huge gap in congestion notification and reaction granularity
Existing solutions fall short in **efficiently** converging to fairness

Congestion Control is crucial

Uno

Throughput Bound

Vastly diverse RTTs and delayed feedback

Mismatch between BDP and buffer capacity.

UnoCC

Fair between different RTTs.
Uses phantom queues to mitigate BDP mismatch.

Latency Bound

Most messages are latency bound.

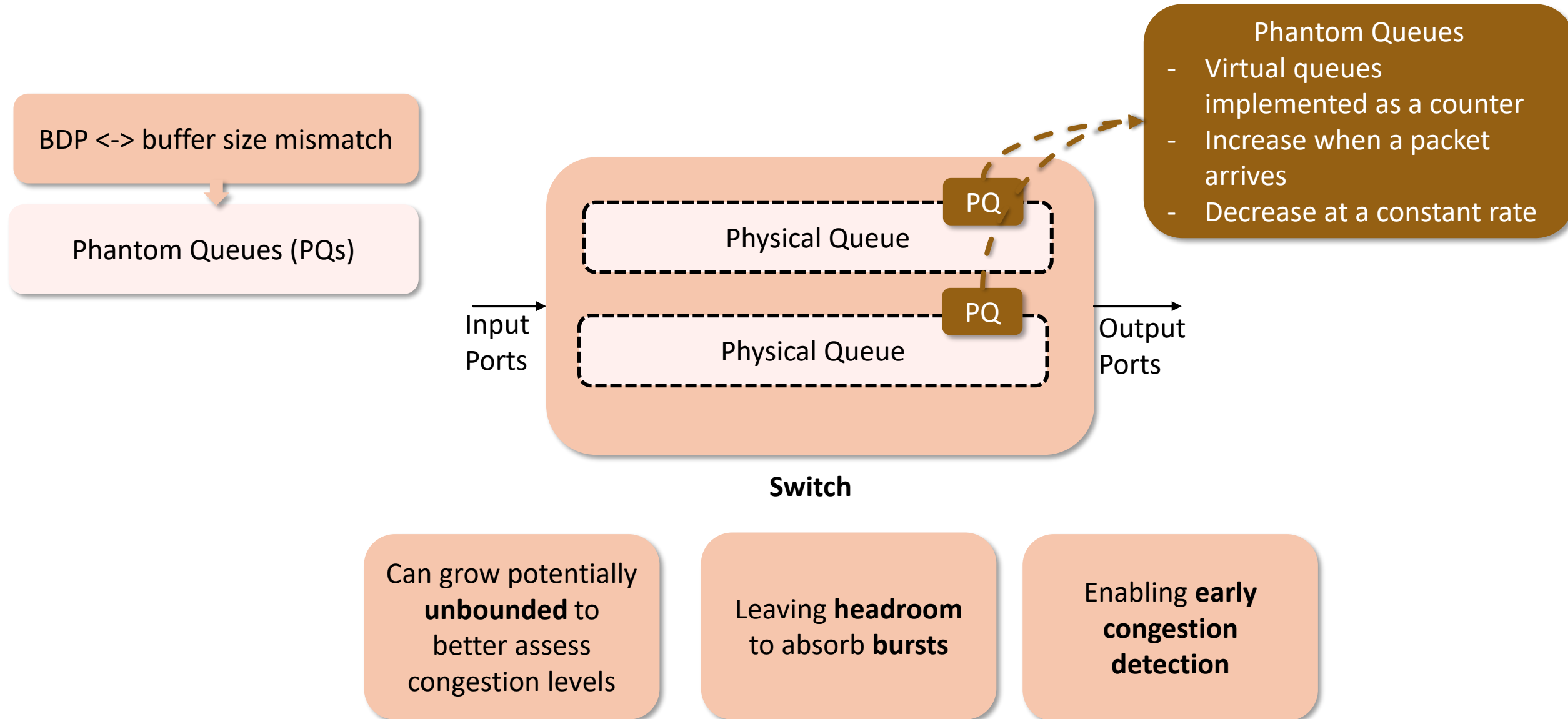
Single loss can have a dramatic impact.

UnoRC

Incorporates erasure coding with load balancing (UnoLB) to limit drops.

Reliability is the most important aspect

UnoCC: fast and fair congestion control for intra- and inter-DC traffic

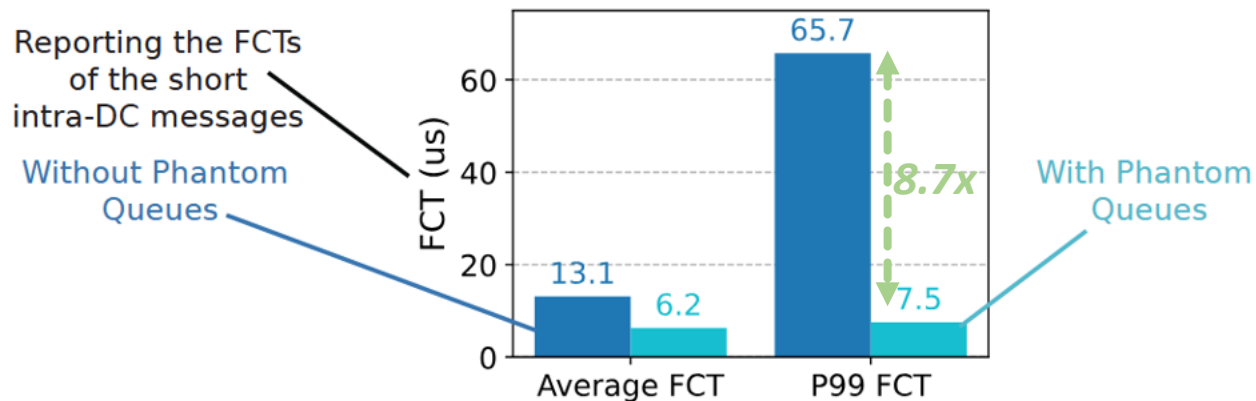
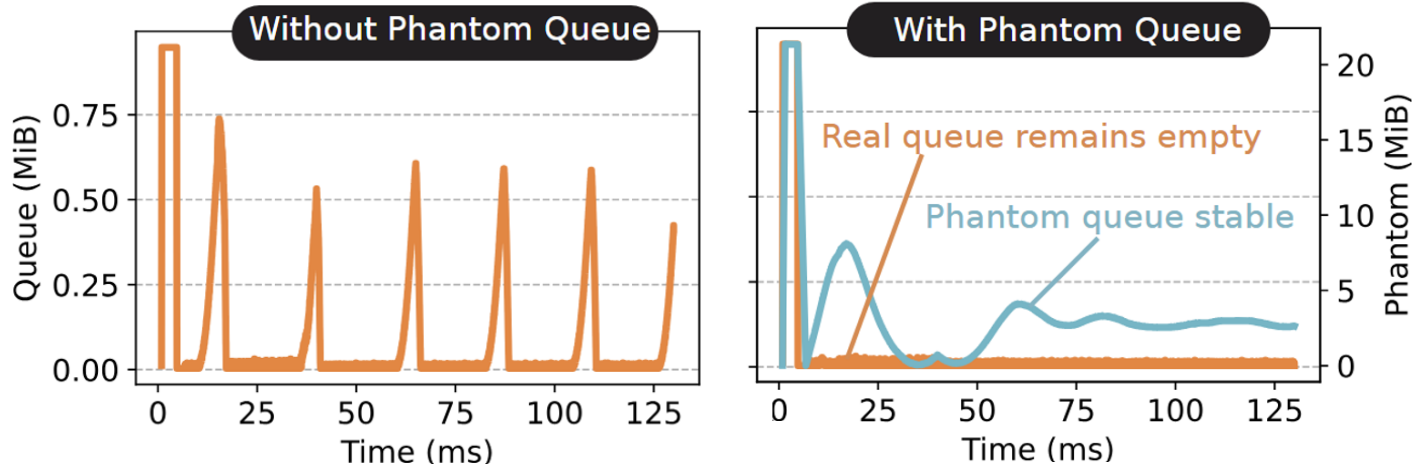


UnoCC: fast and fair congestion control for intra- and inter-DC traffic

BDP $\leftrightarrow$ buffer size mismatch

Phantom Queues (PQs)

Google RPCs beside eight long-lived cross-DC flows



UnoCC: fast and fair congestion control for intra- and inter-DC traffic

BDP <-> buffer size mismatch

Phantom Queues (PQs)

Unfairness/slow convergence

ECN
marked

Inter-DC
flow

Intra-DC
flow

We can potentially achieve **fairness** using **AIMD**.

[Congestion Control for Cross-Datacenter Networks, ToN 2022]

But, convergence happens **slowly**.



UnoCC: fast and fair congestion control for intra- and inter-DC traffic

BDP $\leftrightarrow$ buffer size mismatch

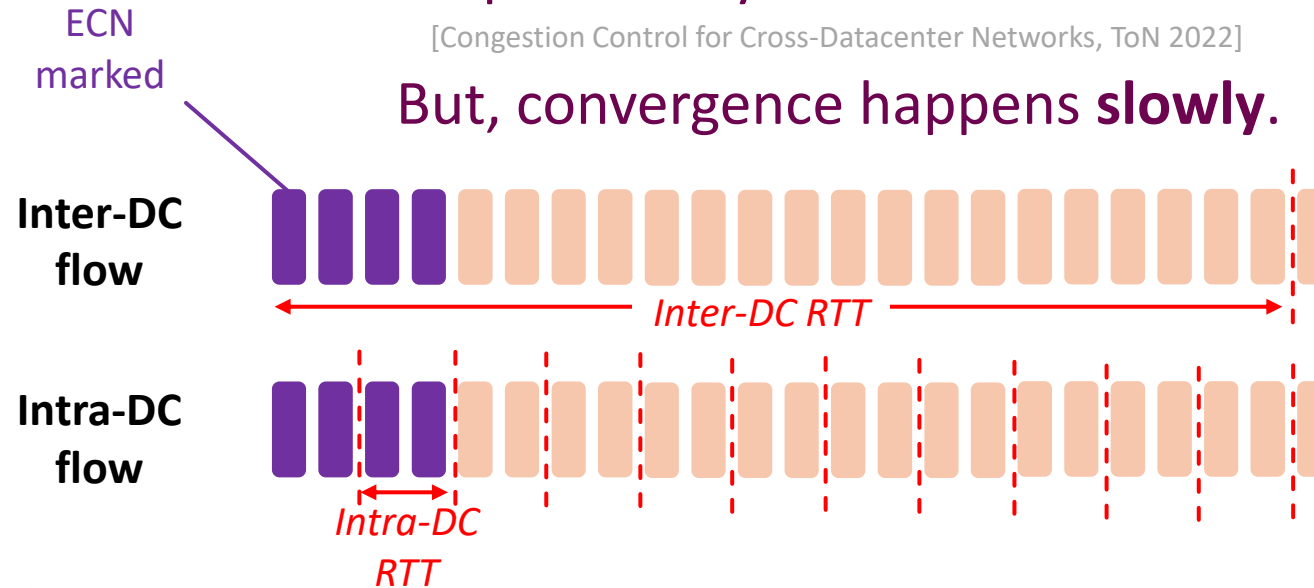
Phantom Queues (PQs)

Unfairness/slow convergence

We can potentially achieve **fairness** using AIMD.

[Congestion Control for Cross-Datacenter Networks, ToN 2022]

But, convergence happens **slowly**.



Rate reduced at different granularities, **prolonging convergence to fairness**

Same degree of congestion captured differently, **causing configuration sensitivity**

UnoCC: fast and fair congestion control for intra- and inter-DC traffic

BDP $\leftrightarrow$ buffer size mismatch

Phantom Queues (PQs)

Unfairness/slow convergence

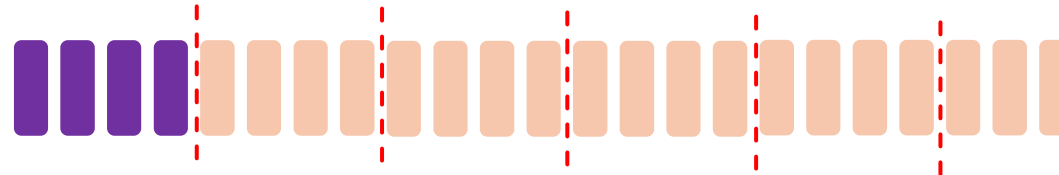
Unified reaction granularity to congestion

We can potentially achieve **fairness** using AIMD.

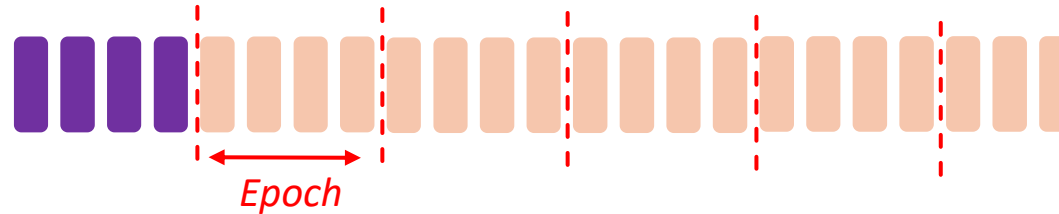
[Congestion Control for Cross-Datacenter Networks, ToN 2022]

But, convergence happens **slowly**.

Inter-DC
flow



Intra-DC
flow



Rate reduced at different granularities, **prolonging convergence to fairness**

Same degree of congestion captured differently, **causing configuration sensitivity**

UnoCC: fast and fair congestion control for intra- and inter-DC traffic

8:1 incast from senders both inside and across datacenters

BDP $\leftrightarrow$ buffer size mismatch

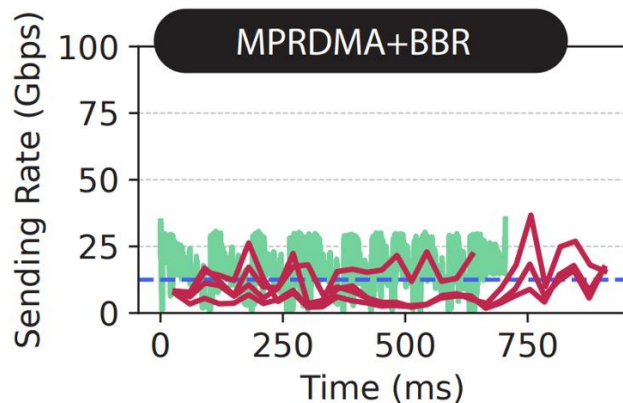
Phantom Queues (PQs)

Unfairness/slow convergence

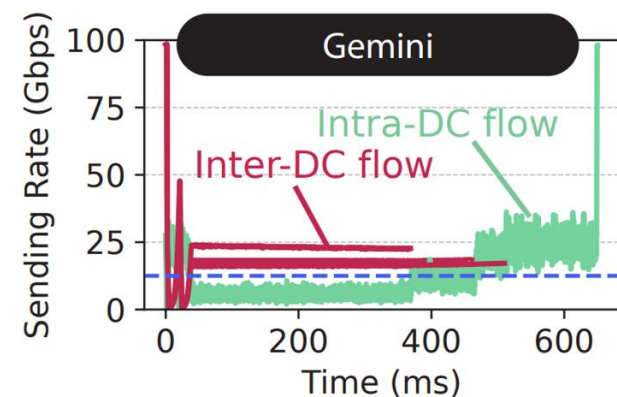
Unified reaction granularity to congestion

Drastic rate reduction under extreme congestion

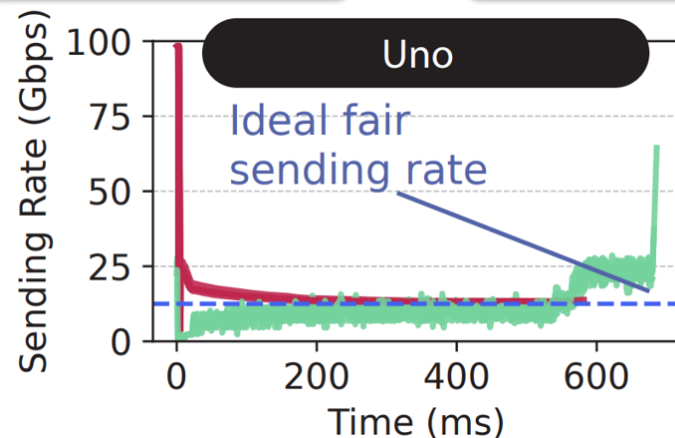
Smoother rate reduction upon phantom queue congestion



Unfairness



Slow convergence to fairness

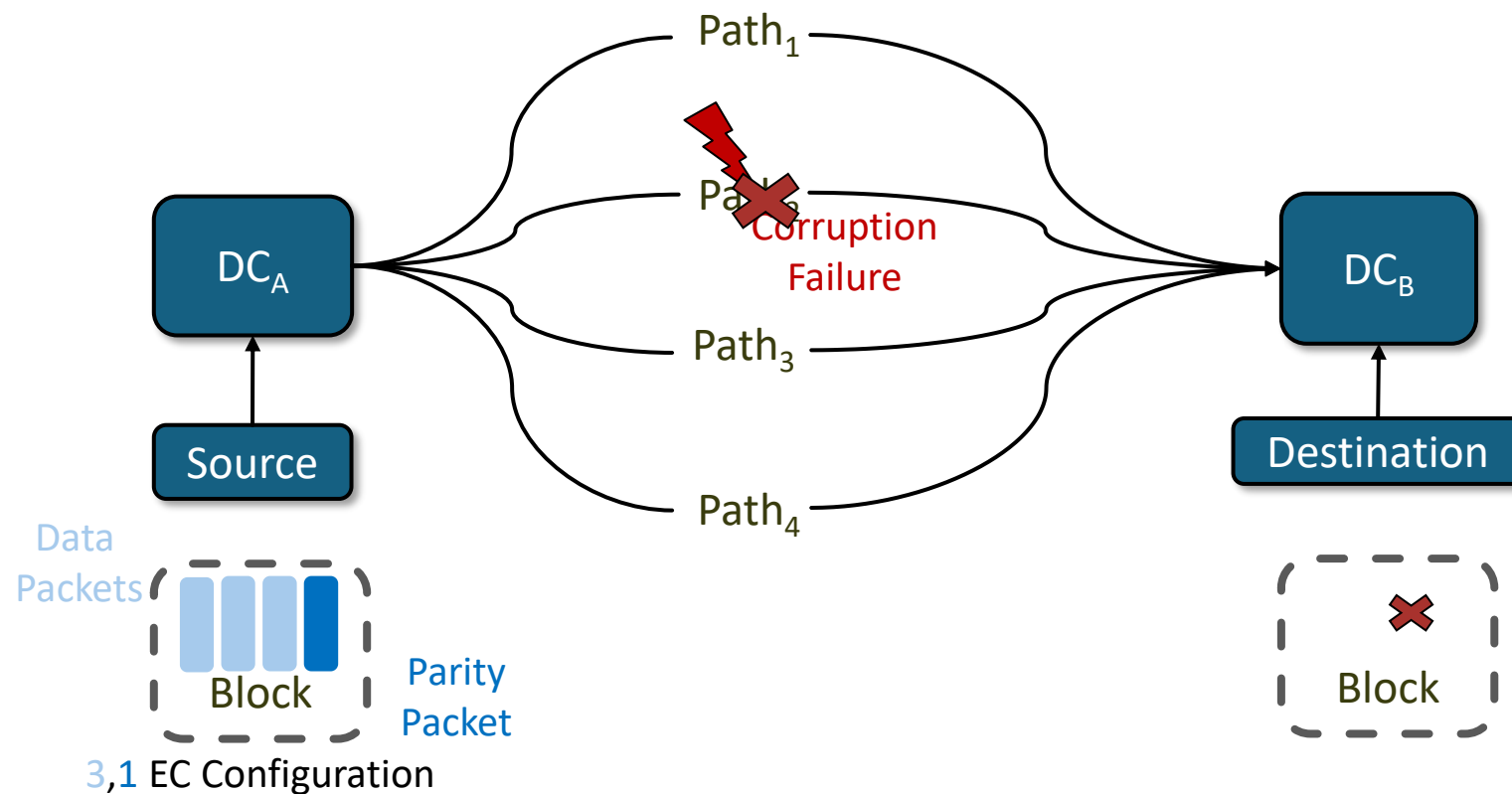


Fast convergence to fairness

UnoRC: Reliable connectivity through load balancing and erasure coding

Packet losses are destructive when latency-bound

Losses are recovered through erasure coding



Data Packet is lost but receiver can still reconstruct the block

However, a link failure would still prevent the block from being delivered

UnoRC: Reliable connectivity through load balancing and erasure coding

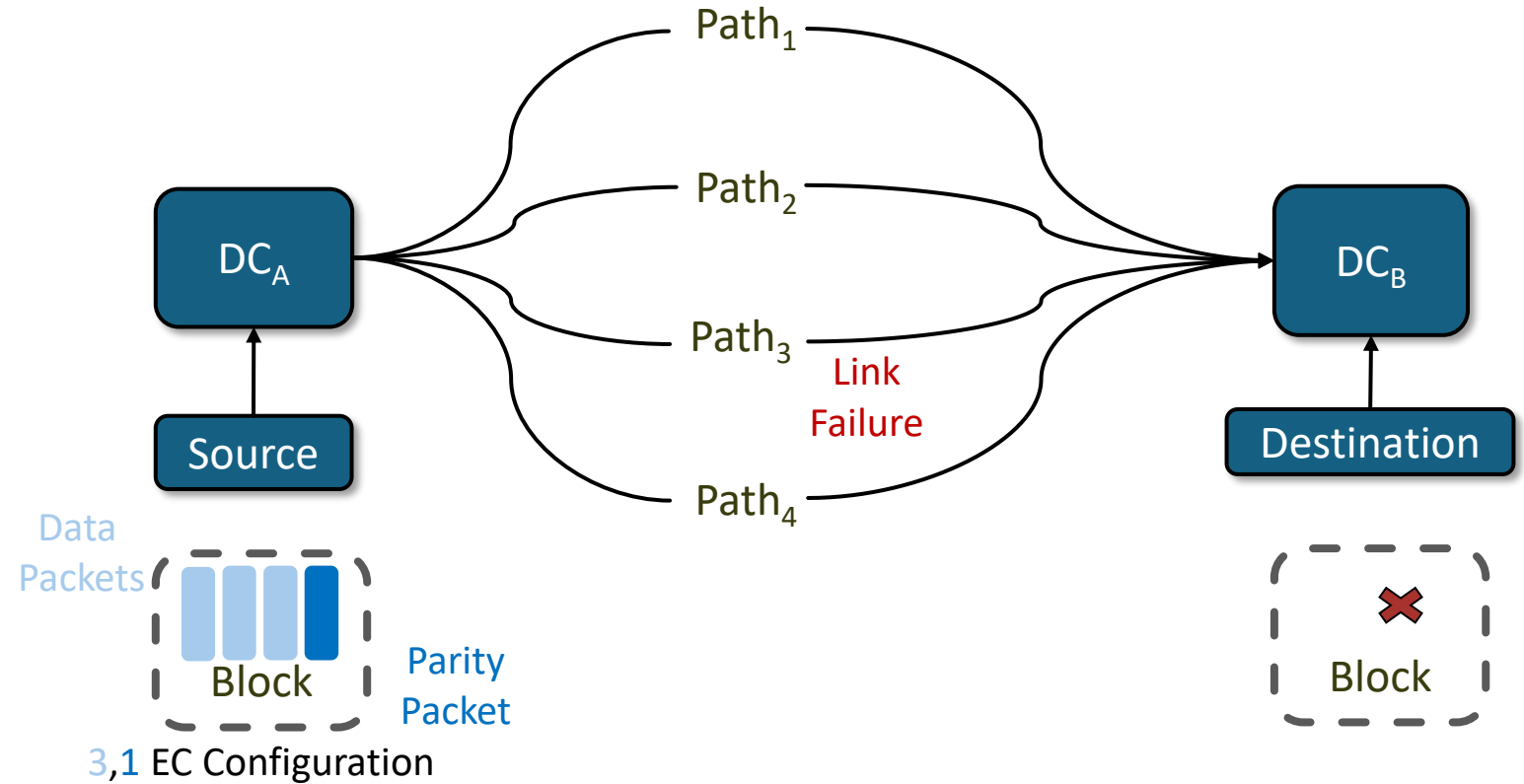
Packet losses are destructive when latency-bound

Losses are recovered through erasure coding

Erasure coding helps recovery from drops but not from failures

Adaptive load balancing to utilize multipathing

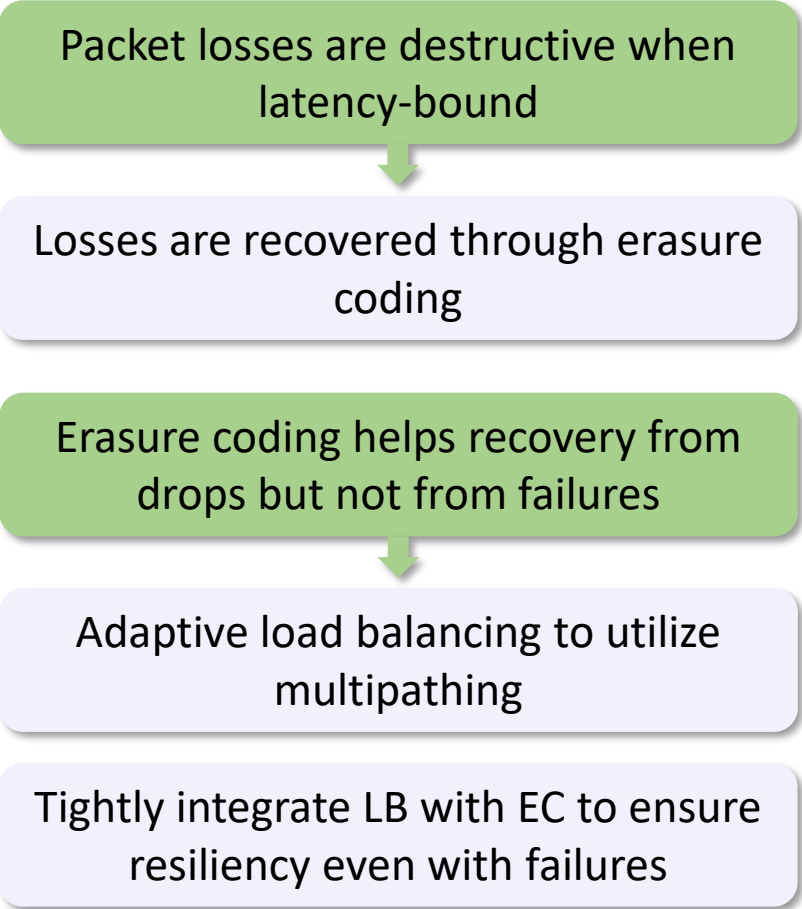
Tightly integrate LB with EC to ensure resiliency even with failures



The block is reconstructed even with a link failure thanks to multipathing

UnoRC quickly learn to avoid the failing path by looking at when it last was active

UnoRC: Reliable connectivity through load balancing and erasure coding

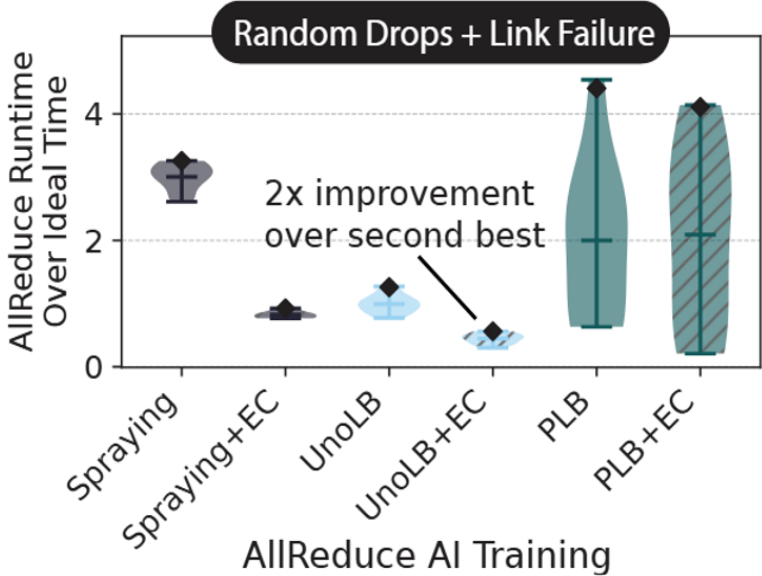


Measured Drop Rates

Losses Within a Block	Setup 1 (65ms RTT)		Setup 2 (33ms RTT)	
	Drops	Loss Rate	Drops	Loss Rate
1	97 403	3.0×10^{-4}	12 785	4.0×10^{-5}
2	23 984	7.5×10^{-5}	7 262	2.3×10^{-5}
3	5 007	1.6×10^{-5}	1 560	4.9×10^{-6}

Table 1: Packet loss information for two datacenter configurations.

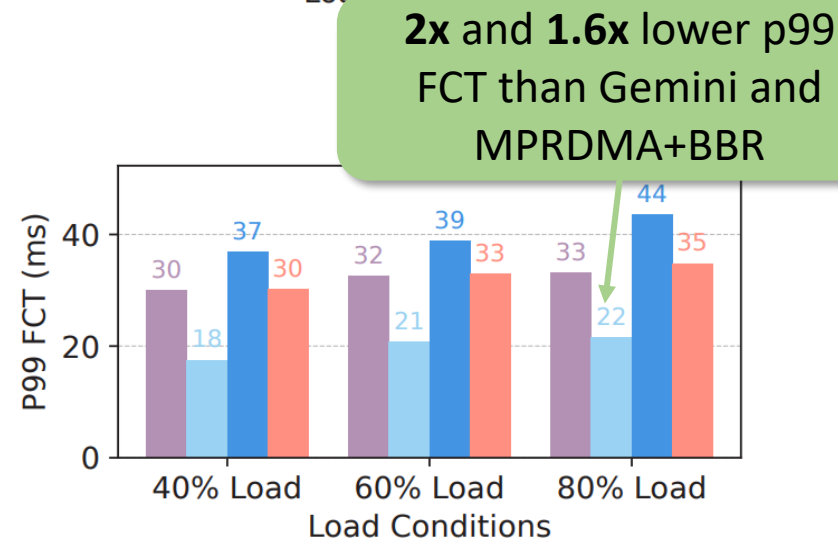
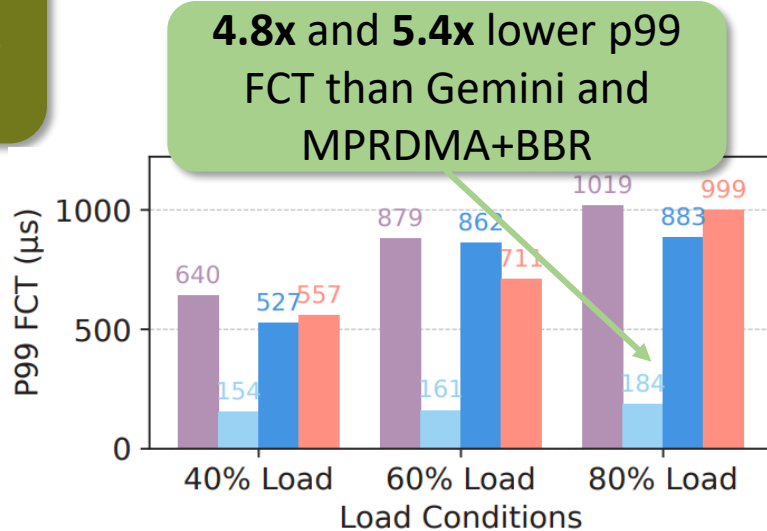
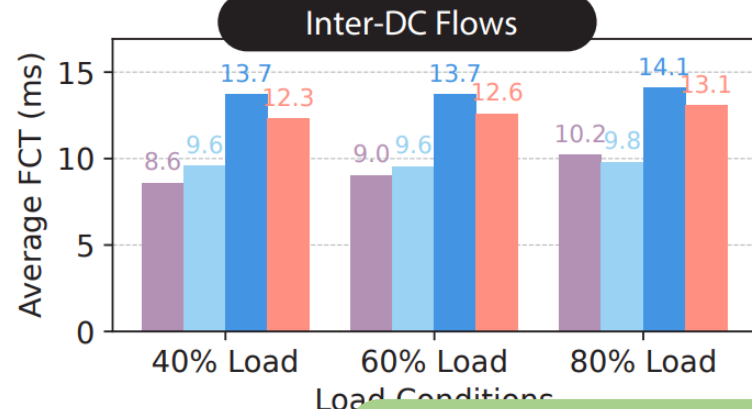
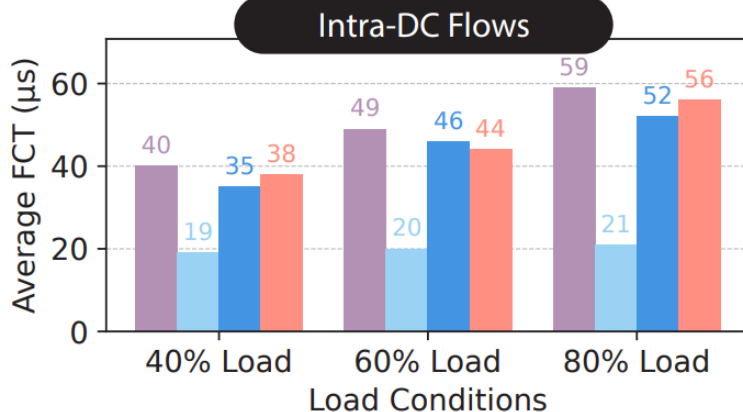
CrossPipe Schedule



Uno is performant under realistic workload

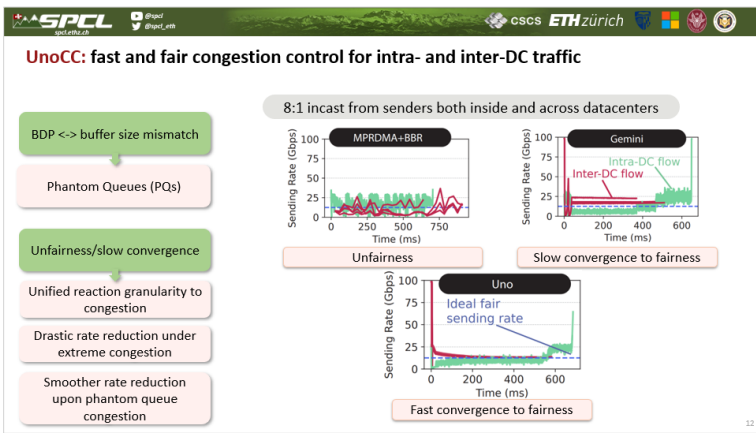
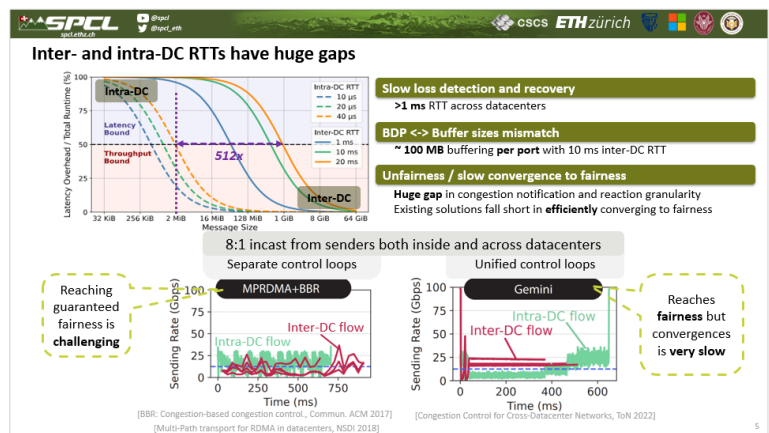
Large-scale simulations with **100 Gbps** links and an RTT difference of **128x** between intra-DC and inter-DC

Running Datacenter Traces
Google web search + Alibaba WAN



Uno+ECMP Uno Gemini MPRDMA+BBR

Conclusions



Paper:



Open-source code and artifact:

