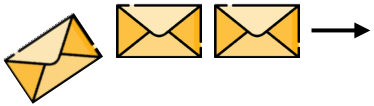
Masking Corruption Packet Losses in Datacenter Networks with Link-local Retransmission

Raj Joshi, Cha Hwan Song, Xin Zhe Khooi, Nishant Budhdev, Ayush Mishra, Mun Choon Chan, Ben Leong

NUS National University of Singapore

Image Credit: Google Datacenters

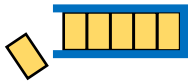
Packet Loss in Datacenter Networks



Packet Loss

Degrades Application Performance

Impact on Revenue

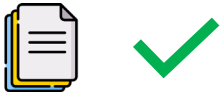


Congestion

Link-local Corruption

This Paper

Mitigating the Impact of Corruption Pkt Loss



Significant Prior Work

Little Attention

Why do we care about corruption packet loss?

Corruption Packet Loss – Significant

Comparable to Congestion Loss

Large-scale study by Microsoft
(350K links, 15 datacenters)
[Zhuo et al. SIGCOMM'17]

Packet drops affected customers

due to corruption – study by
Alibaba Cloud
[Zhou et al. SIGCOMM'20]

Increase in tail FCTs

+

Drop in throughput

How can we fix corruption packet loss?

Physical Repair

Several hours to days

Until then

Mitigate/Mask

Effects of corruption pkt loss

Prior Work – Mitigating Corruption Loss

Current state-of-the-

Disable corru

[e.g. NetPilot,

(* more related work

Wireless Networks

Not revisited in the context of datacenter networks

Challenges unique to DC context

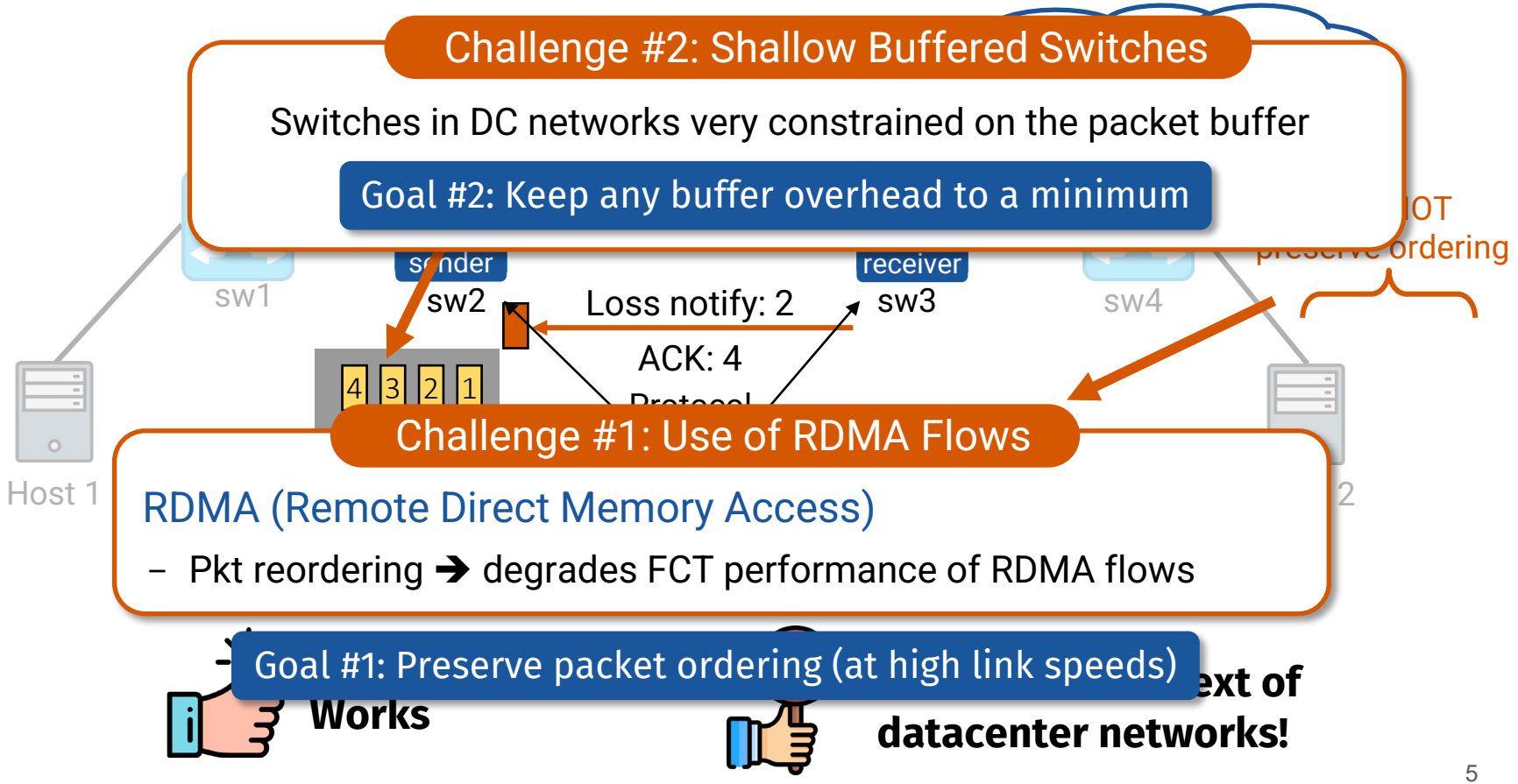
Classical loss-recovery strategy

Our Work

Link-local Retransmission

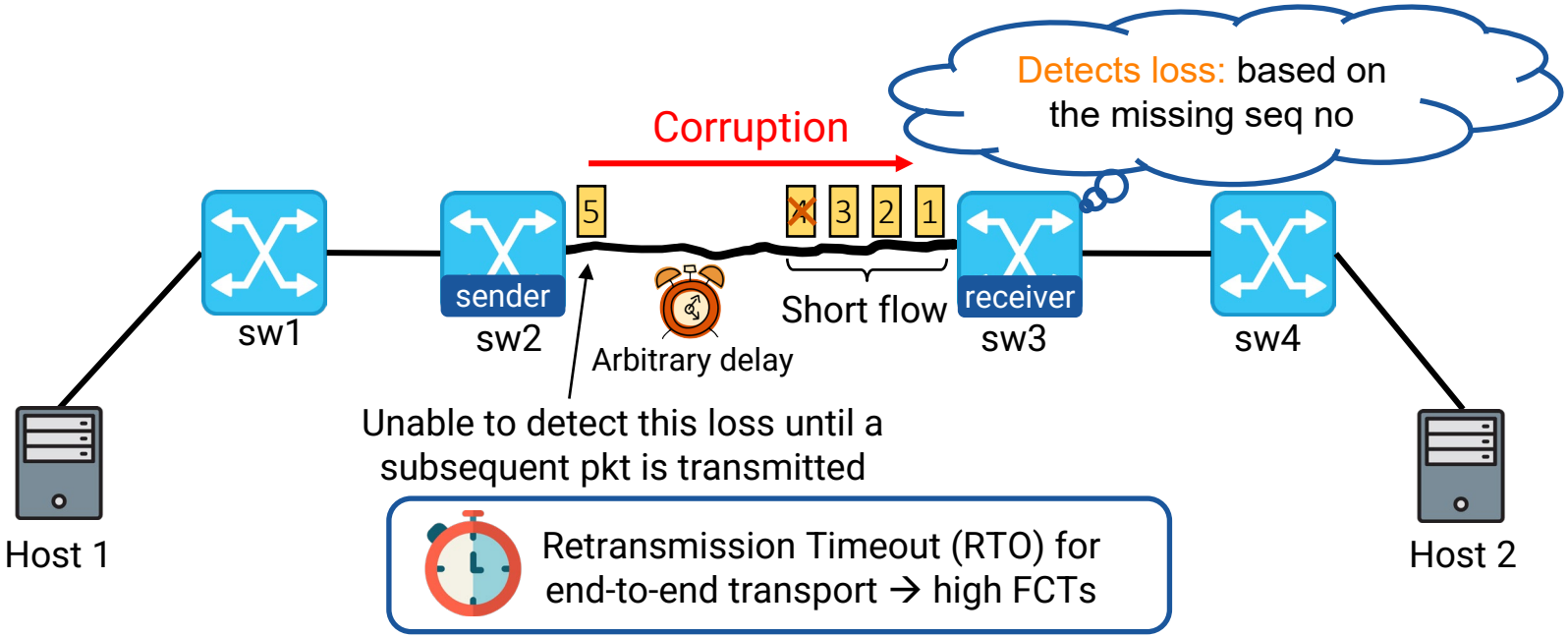
[LinkGuardian]

Basic Link-local Retransmission



Challenge #3: Flows in datacenter are very short!

Short flows + corruption → higher probability tail (last) packet is lost



Goal #3: Detect and recover tail packet losses (as quickly as possible)

LinkGuardian

LinkGuardian – 3 Key Techniques

Preserve Pkt Ordering

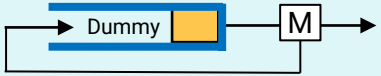
Ordering Buffer
At receiver switch

This talk!

Shallow Buffered Switches

Fast & Efficient ACK Mechanism
At receiver switch

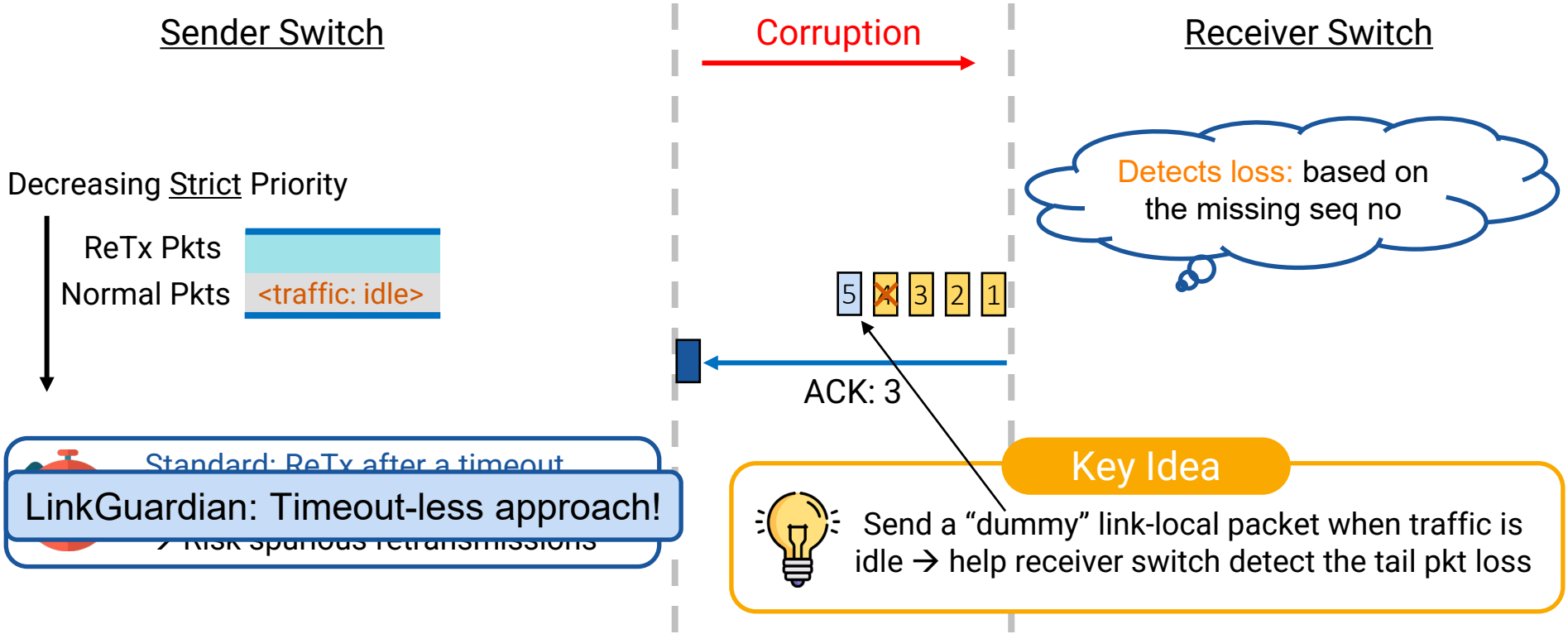
Tail Pkt Loss



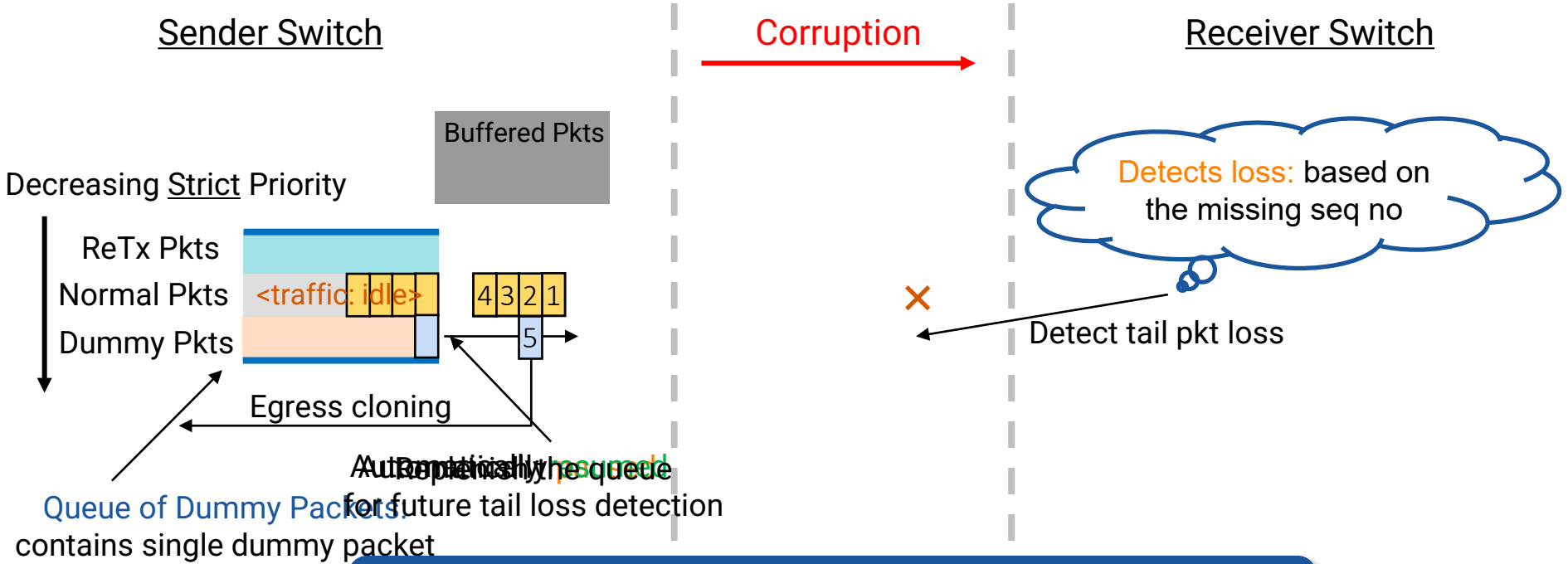
Self-Replenishing Queue of Dummy Pkts
At sender switch

Basic Link-Local Retransmission Scheme

Handle Tail Packet Loss



Handle Tail Packet Loss



Tail losses are detected near instantaneously
No extra overhead: dummy pkts sent only when link is idle

LinkGuardian Implementation and Evaluation

Intel Tofino programmable switch (~1800 lines of P4)

Source code available!

Hardware Testbed

- Intel Tofino switches, Intel Xeon servers w/ 100 Gbps NICs

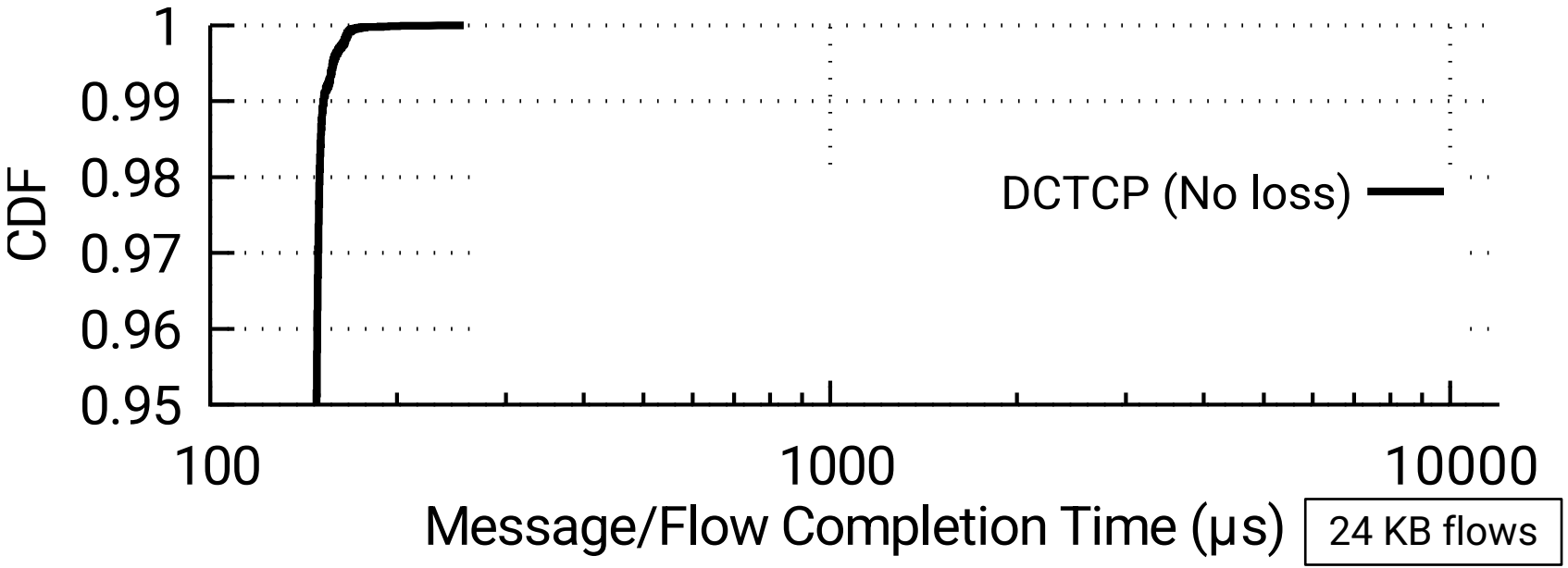
LinkGuardian Variants

- LinkGuardian (default): preserves packet ordering
- LinkGuardianNB (non-blocking): “lite” version → everything, except preserving ordering

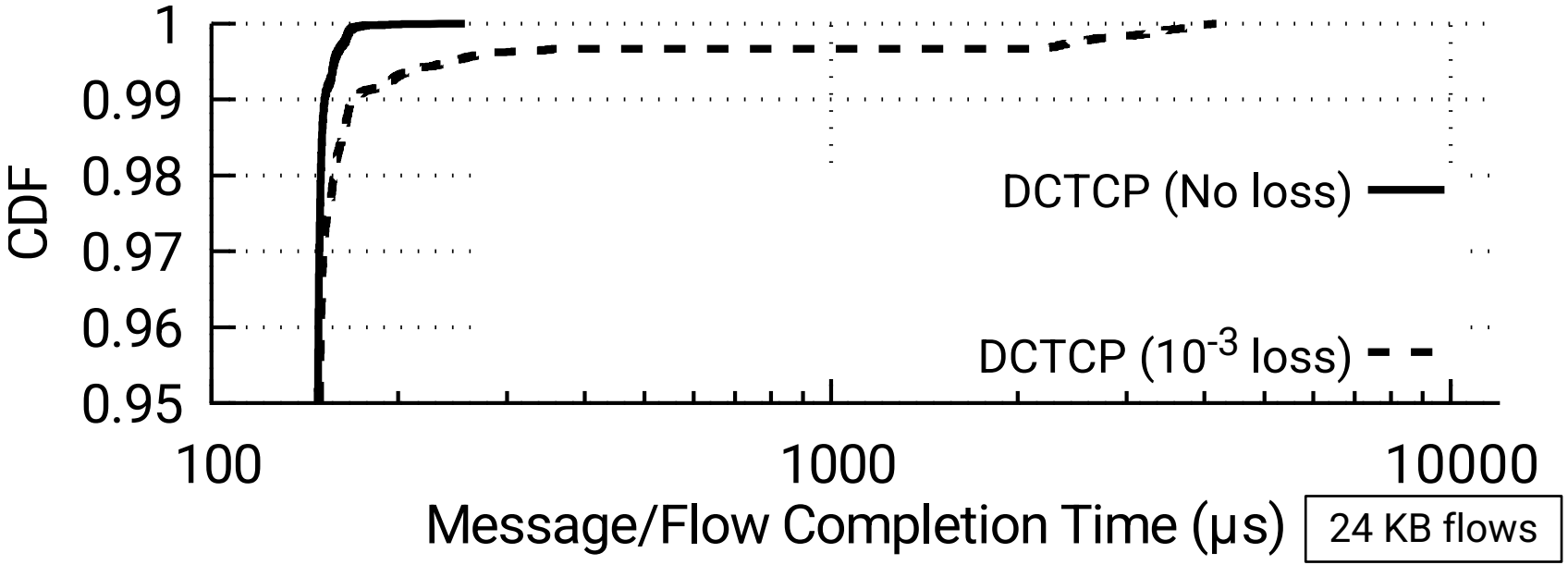
Variable Optical Attenuator

Corruption Loss Rate: 10^{-3}
(more loss rates in the paper)

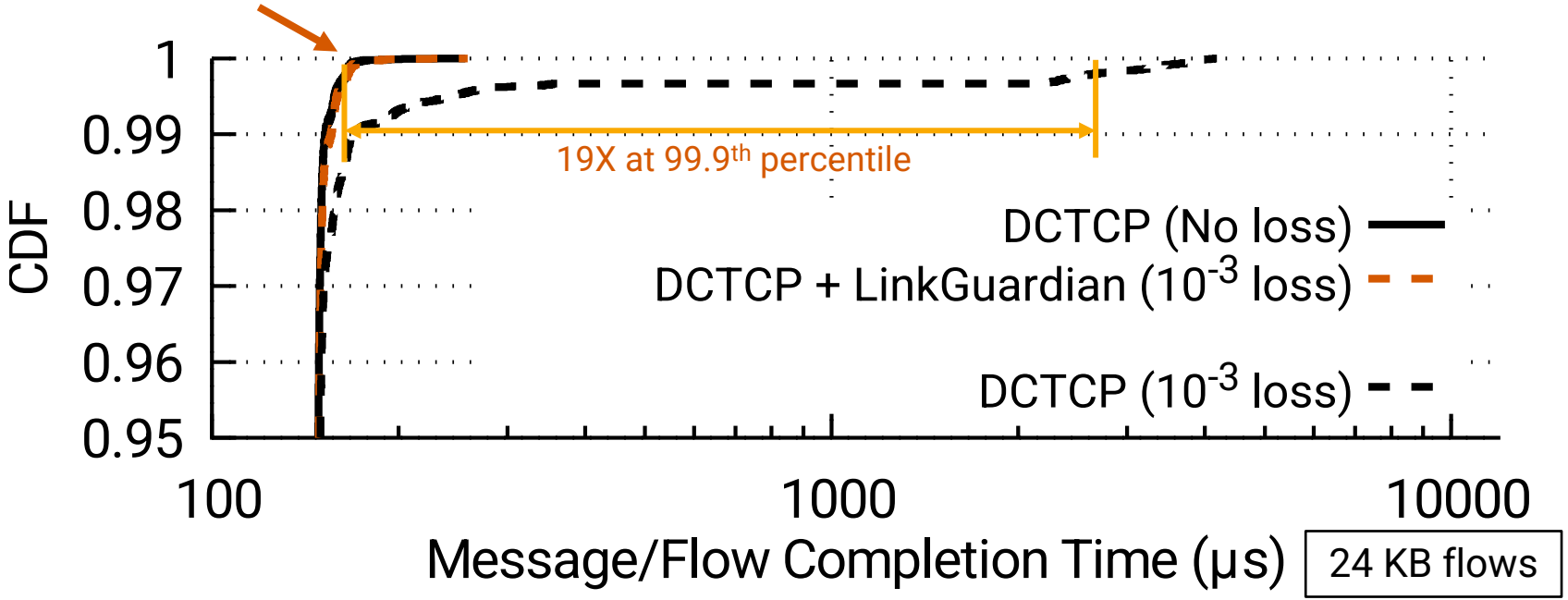
LinkGuardian improves tail FCTs for TCP (DCTCP)



LinkGuardian improves tail FCTs for TCP (DCTCP)

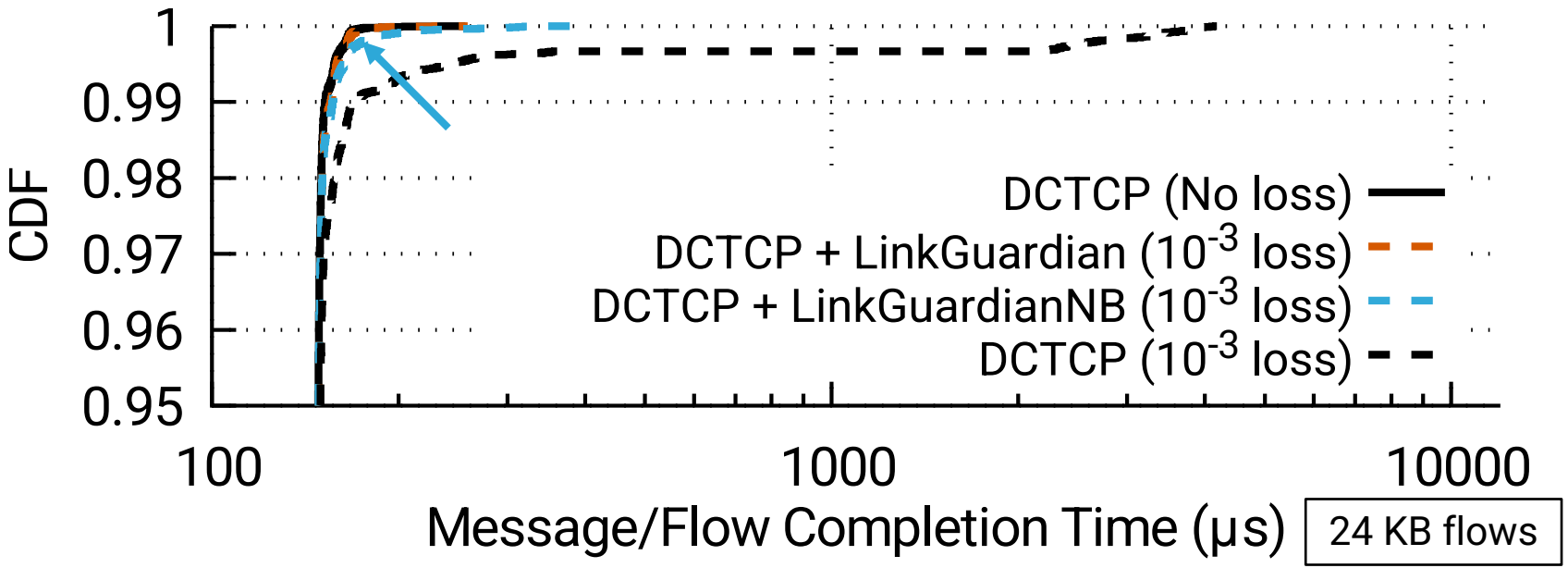


LinkGuardian improves tail FCTs for TCP (DCTCP)



LinkGuardian: improves 99.9th percentile FCT by 19 times. Matches no loss case!

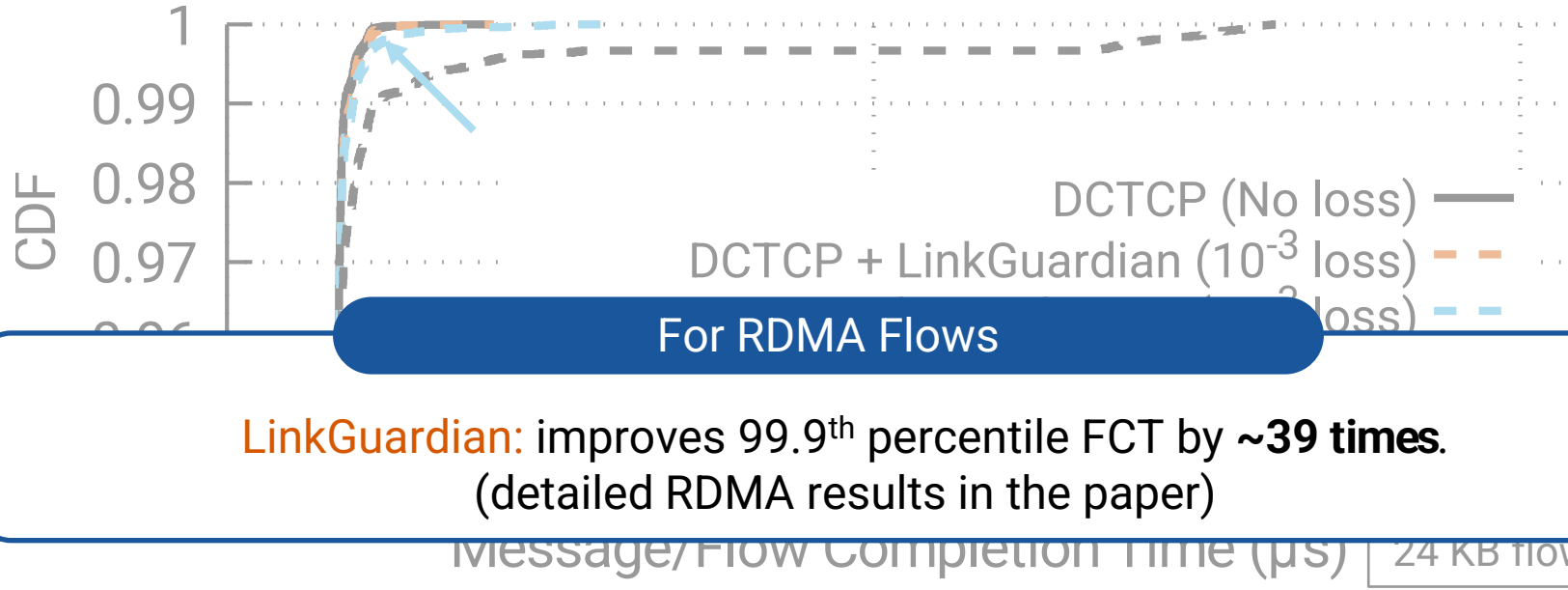
LinkGuardian improves tail FCTs for TCP (DCTCP)



LinkGuardianNB: nearly the same improvement, even without maintaining pkt order!

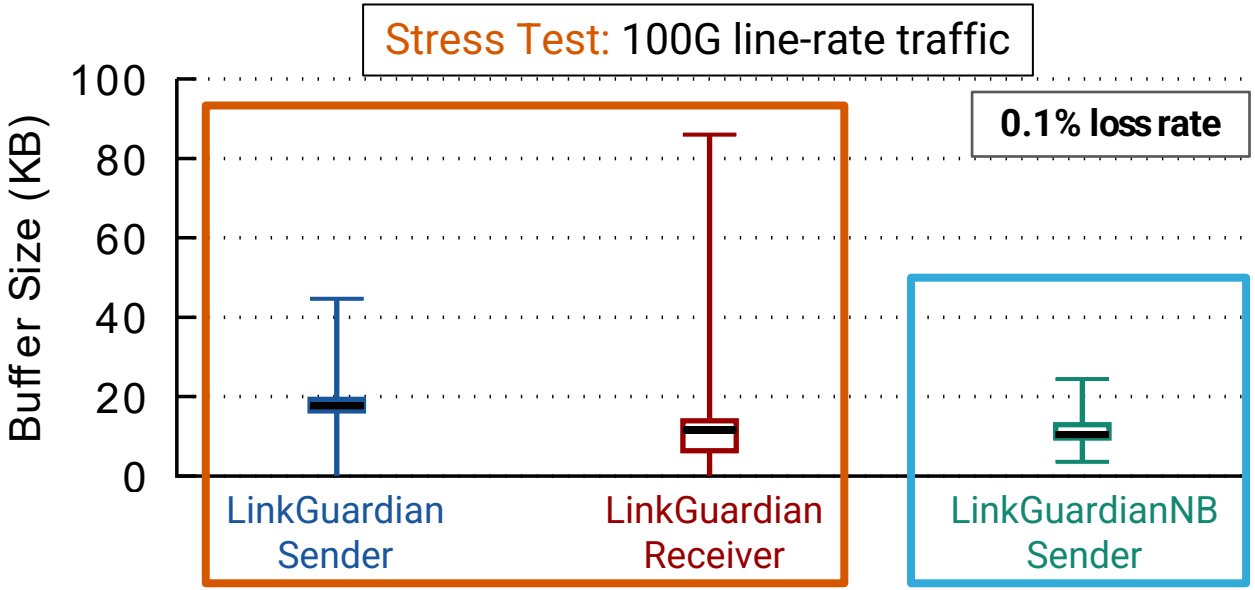
TCP flows are short: reduction in cwnd (due to out-of-order) has minimal impact!
(more detailed analysis in the paper)

LinkGuardian improves tail FCTs for TCP (DCTCP)



LinkGuardianNB: nearly the same improvement, even without maintaining pkt order!
TCP flows are short: reduction in cwnd (due to out-of-order) has minimal impact!
(more detailed analysis in the paper)

LinkGuardian's Packet Buffer Overhead



LinkGuardian: < 90KB for both sender and receiver switches

LinkGuardianNB: < 25KB for sender switch. No ordering buffer at receiver!

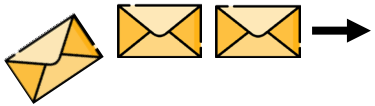
Put this result in context: Switches have 10's of MBs of packet buffer

Summary

LinkGuardian ➔ mitigate corruption pkt loss

Selective In-network Recovery (sub-RTT)

pkt caching in dataplane



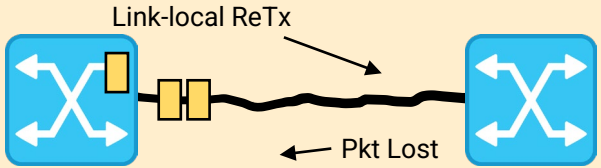
Handles Tail Pkt Loss

Without employing any timeouts

Preserves Pkt Ordering

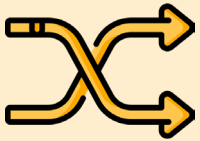
For reordering sensitive RDMA flows

Key Takeaways



Link-local ReTx

Both *practical* and *effective* in today's DC networks



Out-of-Order ReTx

Sufficient for short TCP flows

Thank you!

https://github.com/NUS-SNL/linkguardian

Credits: This presentation template was adapted from Slidesgo, including icons by Flaticon.

Flaticon icons by: Becris, Vectors Market, Flat Icons, Maxim Basinski Premium, Darius Dan, noomtah, iconixar, Kiranshastry, Bombasticon Studio, fjstudio, Vectorslab, Chattapat, Freepik, Andy Horvath, itim2101, Pavel Kozlov, Pixel perfect, Cuputo, Icon home, Fajri rama, Good Ware, Vitaly Gorbachev, Eucalyp, pongsakornRed, and Octopocto.