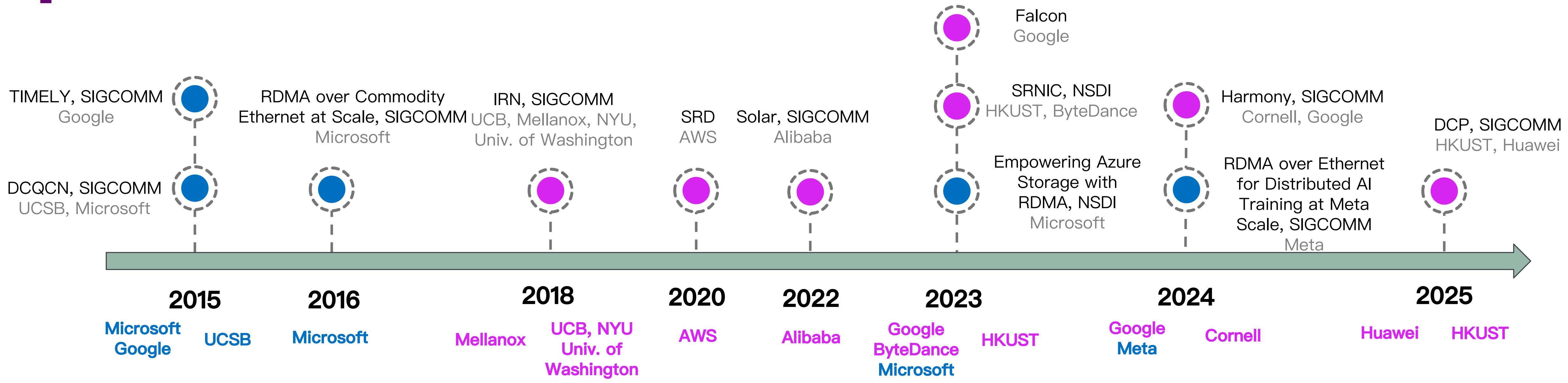
Revisiting RDMA Reliability for Lossy Fabrics

Wenxue Li[*#]**,** Xiangzhou Liu[*], Yunxuan Zhang[*], Zihao Wang[*], Wei Gu[#], Tao Qian[#], Gaoxiong Zeng[#], Shoushou Ren[#], Xinyang Huang[*], Zhenghang Ren[*], Bowen Liu[*], Junxue Zhang[*], Kai Chen[*], Bingyang Liu[#]

[*]*iSING Lab, Hong Kong University of Science and Technology* [#]*Huawei*

This work was done while Wenxue Li was an intern at Huawei

Landscape: From Lossless to Lossy



- Restricted Deployment Scale of Lossless RDMA Network:

Go-Back-N retransmission

↓

Require PFC/CBFC to ensure lossless fabric

↗

PFC/CBFC causes several **performance issues** (HoL blocking, congestion spreading, deadlock) and significant **switch buffer overhead**

↓

Restricted Deployment Scale

Issues of RNIC–SR: (#1) Incompatibility with Packet–level LB

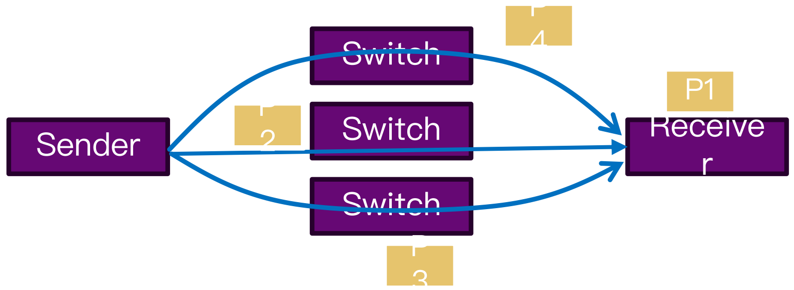
SOTA lossy RDMA solutions

Implements a simplified **selective repeat (SR)** mechanism in RNICs to enhance loss recovery efficiency

- However, even with RNIC–SR, **performance issues persist** in lossy fabrics

1 RNIC–SR is natively incompatible with packet–level load balancing (LB)

- RNIC–SR assumes single–path transmission, with ECMP as the default LB scheme.
- ECMP hashing collisions cause significant throughput degradation. (especially for AI workload!)
- Packet–level LBs are promising alternatives to ECMP.
- However, combining packet–level LB with RNIC–SR leads to excessive spurious retransmissions.

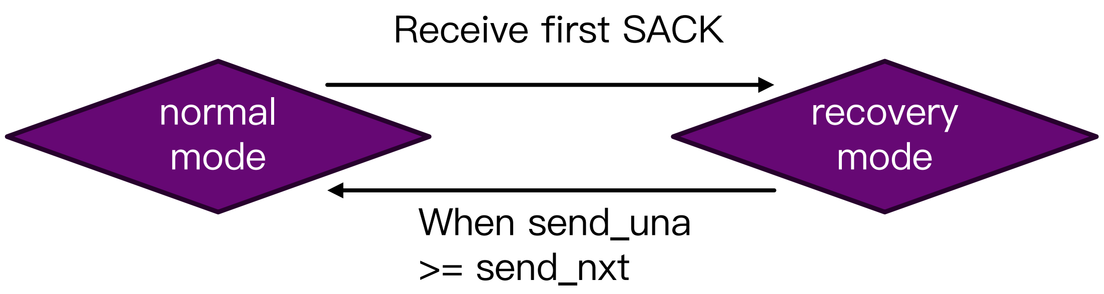


P4 arrives first. The receiver will require P2 & P3 retransmissions

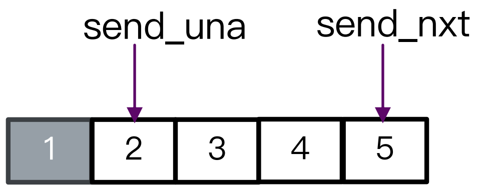
Issues of RNIC–SR: (#2) Excessive RTOs

2 Certain lost packets cannot be recovered through fast retransmission in RNIC–SR

- RNIC–SR **requires a SACK to trigger the loss recovery mode**. If the tail packet of a flow is lost, no SACK is generated → **Reliance on RTO**
- To avoid retransmission ambiguities, the sender **enters the loss recovery mode only once and remains in this state until it exits**. If the retransmitted packets are dropped again → **Reliance on RTO**



During recovery mode, packets are only retransmitted once



Could significantly degrade performance!

Our Proposal: DCP

- We aim to revisit RDMA reliability to fully meet the following objectives:

❶ **Independence from PFC**

❷ **Compatibility with packet-level LB**

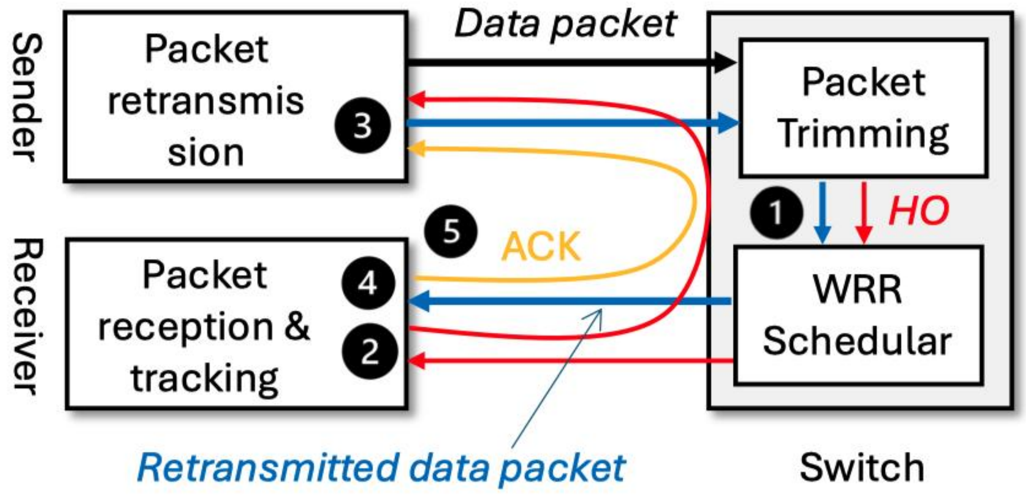
❸ **Ability to quickly retransmit any lost packet**

❹ **A hardware-oriented design***

*With the feasibility of RNIC offloading (i.e., low memory and processing overhead)

Key Idea of DCP

- DCP-Switch ensures a lossless **Control Plane** (for header transfer) while allowing the **Data Plane** (for payload transfer) to operate in a lossy manner.
- DCP's key idea: leverage the lossless CP feature to enhance RNICs



RNIC's reliability: packet retransmission, reception, & tracking

Comparison of DCP and related works

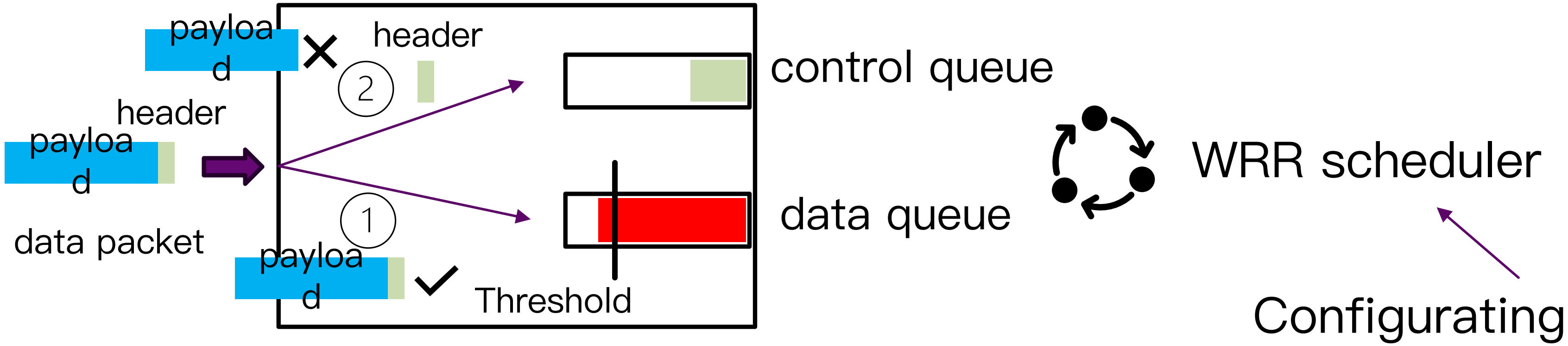
Requirements	R1	R2	R3	R4
RNIC-GBN [8]	×	×	×	✓
RNIC-SR [9, 10, 40, 51]	✓	×	×	✓
MPTCP [45]	✓	✓	×	×
NDP [26]	✓	✓	✓	×
CP [18]	✓	✓	✓	×
MP-RDMA [36]	×	✓	×	✓
DCP	✓	✓	✓	✓

[18] Catch the Whole Lot in an Action: Rapid Precise Packet Loss Notification in Data Centers, NSDI 2014
[26] Re-architecting datacenter networks and stacks for low latency and high performance, SIGCOMM 2017
[36] Multi-Path Transport for RDMA in Datacenters, NSDI 2018

DCP Design: Lossless Control Plane

Step 1:

- When there is no congestion (low queue length): the whole data packet is enqueued into the data queue
- When the data queue's length *exceeds* a threshold: the payload is trimmed, and goes to Step 2



Configurating the WRR weight to ensure lossless control queue

Step 2:
The DCP tag in the remaining header is modified, and the header-only (HO) packet is enqueued into the control queue

DCP Design: Efficient HO-based Retransmission

- **Straightforward HO-based retransmission**

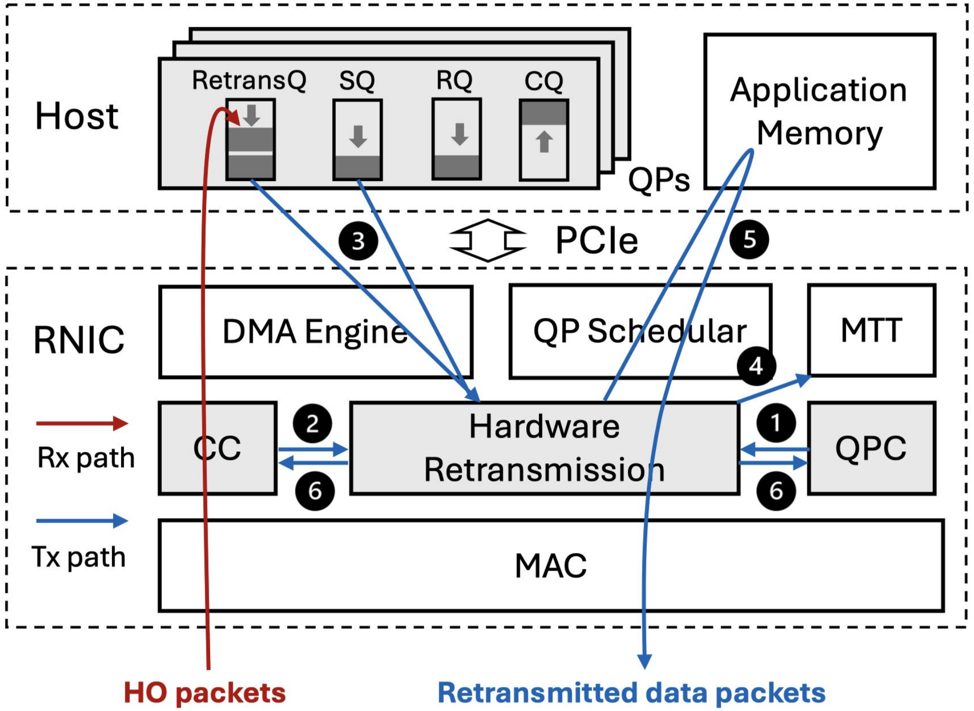
Upon receiving an HO, the sender-side RNIC (1) fetches the corresponding WQE and processes it; (2) fetches the data; (3) encapsulates the data to a packet.

Issues #1: Inefficient retransmission

Two PCIe RTTs, one packet. 1KB/2us = 4Gbps

Issue #2: Incompatible with the CC module

Since HO packets are stateless, the retransmission rate is tied to the receiving rate of the HO packets



1. Extracts metadata from HO, packages it into a retrans. entry, writes it into the corresponding QP's **RetransQ**

1. Check if RetransQ is empty
2. Get the rate/win value from CC
3. Fetch **multiple retrans. entries and WQEs** from RetransQ and SQ, respectively
4. Virtual to physical
5. Fetch and encapsulate **multiple packet payloads**
6. Update CC states

DCP Design: Order-tolerant Packet Reception

- How to handle out-of-order (OOO) packets?

Reorder Buffer?

The standard RDMA header format must be extended to allow the RNIC to **write all packets, whether in-order or OOO, directly to the correct locations** in application memory

Our header extension approach*:

<table>
<tr><th colspan="2">One-sided operation (e.g., Write)</th><th colspan="2">Two-sided operation (e.g., Send)</th></tr>
<tr>
<td rowspan="2">RDMA Extended Transport Header (RETH)
<table><tr><th>Field Name</th><th>Field Abbreviation</th><th>Field Size (in bits)</th></tr><tr><td>Virtual Address</td><td>VA</td><td>64</td></tr><tr><td>Remote Key</td><td>R_Key</td><td>32</td></tr><tr><td>DMA Length</td><td>DMALen</td><td>32</td></tr></table></td>
<td rowspan="2">Note:
sender RNIC fills VA individually for each packet</td>
<td>Send Sequence Number (SSN)</td>
<td>The posting order of two-sided operations (i.e., send, write-with-immediate)</td>
</tr>
<tr>
<td>PSN in BTH
→ first PSN & PSN offset</td>
<td>Note:
the PSN space is reduced</td>
</tr>
</table>

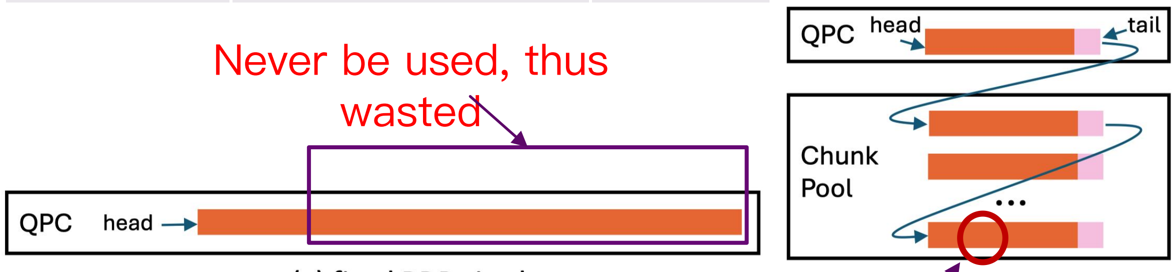
*Common approach adopted by many protocols, such as Falcon, xxx

DCP Design: Bitmap–free Packet Tracking

- How to track which packets have been received or lost?

Usually using packet–level bitmap, which, however, faces tradeoffs.

Approach	(a) Fixed BDP–sized	(b) Linked chunk	DCP
Packet rate	Good (constant packet processing latency)	Bad (linear latency with OOO degree)	Good
Memory overhead	High	Low	Low



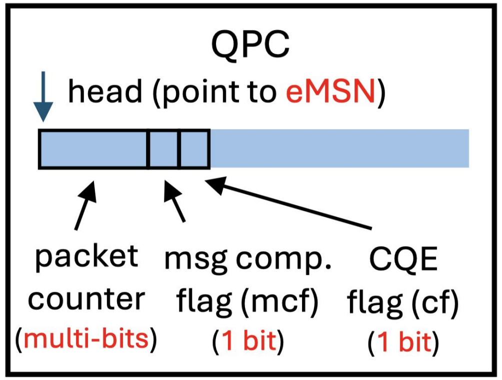
(a) fixed BDP-sized

(b) Linked chunk

High processing cycles

The retransmission module ensure that only truly lost packets are retransmitted → **Exactly–Once**

For any given packet, exactly one copy arrives at the receiver



Bitmap–free Packet Tracking

- A multi–bit counter for each message
  - Memory requirement: $n \rightarrow \log_2(n)$ bits
- Need a customized timeout as Fallback

Combination of Design Modules

Traditional packet–level bitmaps + the remaining components of DCP–RNIC also function correctly

Commonly–used packet–level bitmap tracking

Lossless control plane

HO–based retransmission

Order–tolerant packet reception

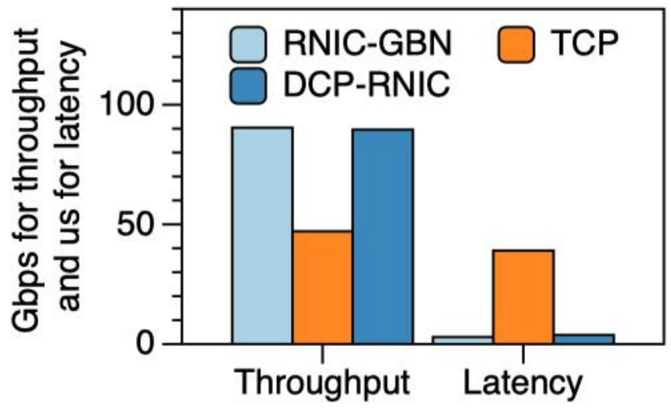
Bitmap–free packet tracking*

Highly relies on Exactly–once feature and uses timeout to handle extreme cases, such as control plane loss and switch/link failures.

*The bitmap–free design is **orthogonal** to the rest of DCP–RNIC's architecture.

Implementation

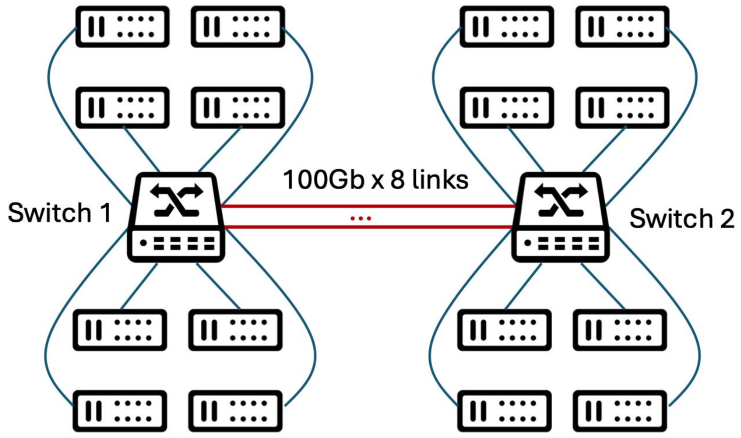
- DCP Switch:
  - We implement the lossless control plane using Tofino2 switch
- DCP RNIC:
  - We build a fully functional prototype of DCP-RNIC using an FPGA board
  - We implement DCP-RNIC by modifying specific modules based on an FPGA-based RNIC-GBN baseline prototype.



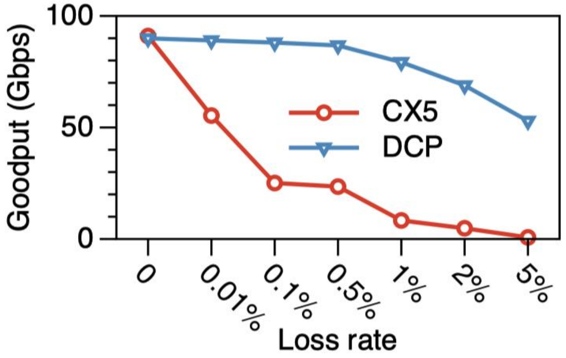
DCP-RNIC successfully maintains hardware offloading capabilities

Evaluation

- Realistic testbed evaluations

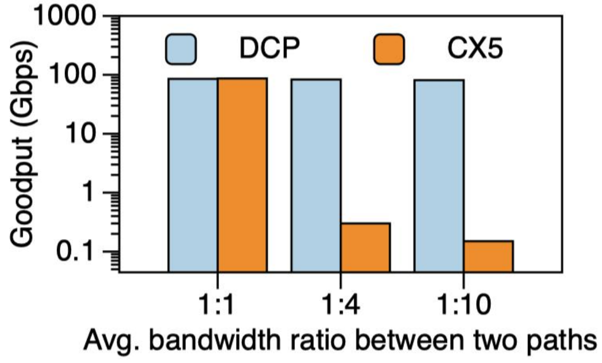


Testbed topology:
- Two switches and 16 servers



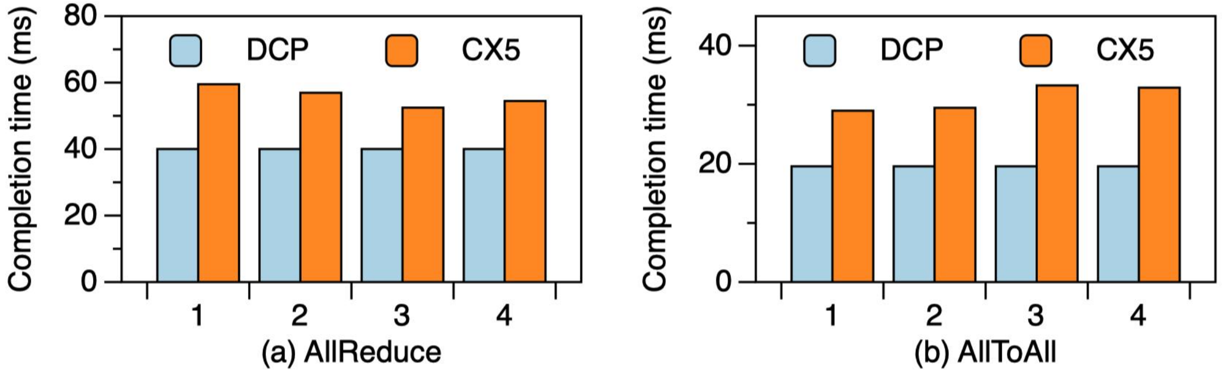
#1: DCP achieves **superior loss recovery efficiency**

1.6× ~ 72×



#2: DCP is natively compatible with AR

DCP maintains stable goodput under all capacity ratios



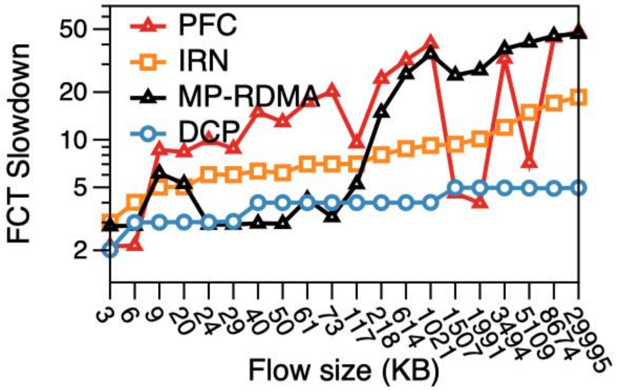
#3: DCP benefits AI workloads

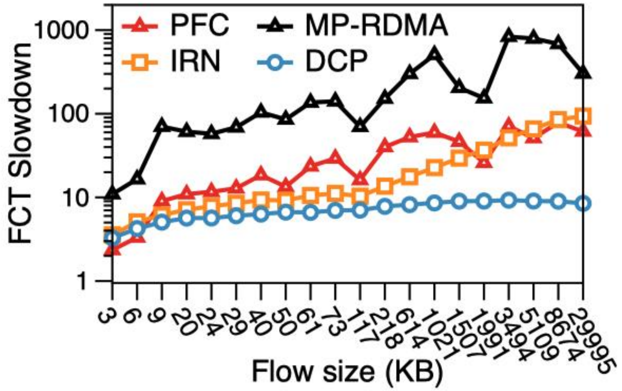
DCP reduces the JCT of AllReduce and AllToAll by up to 33% and 42%, respectively.

Evaluation (Cont.)

- Simulations:
  - Two-layer CLOS network
  - 256 servers (16 per rack)
  - All links operate at 100 Gbps

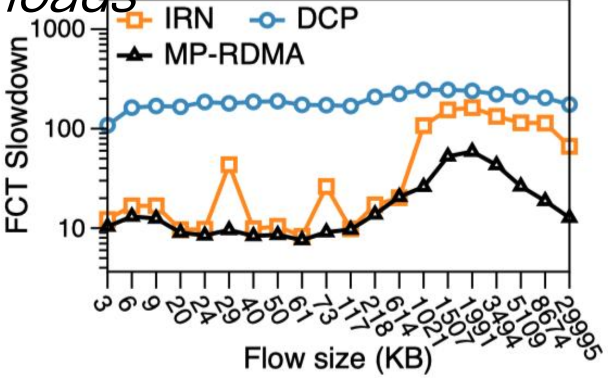
#1: Cross-DC scenarios



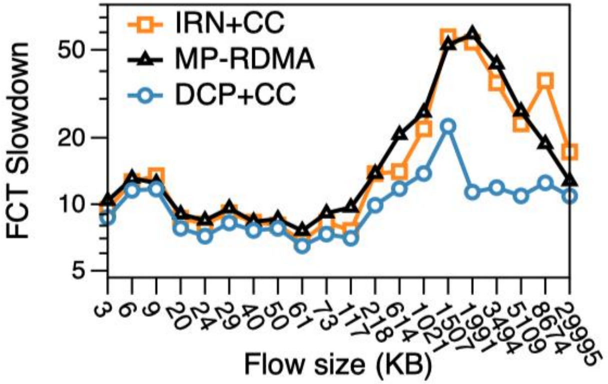


Lossless solutions: **600MB/6GB** switch buffer for 100/1000 km distances
Lossy solutions; (IRN and DCP): **32 MB** switch buffer

#2: DCP needs CC under high loads

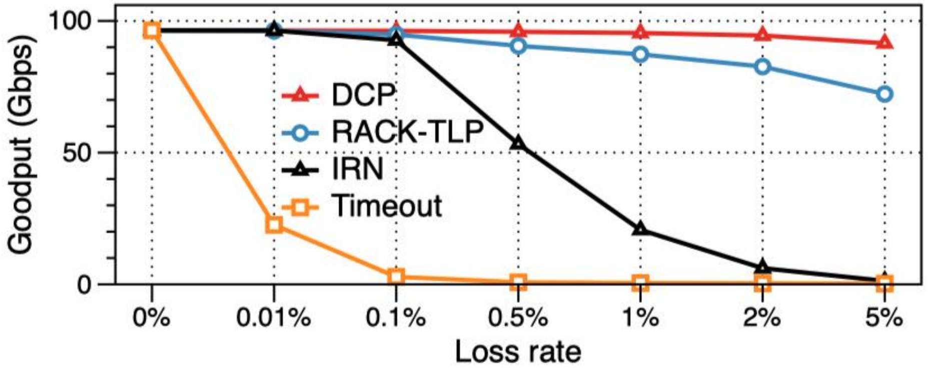


Without any CC: many retransmitted packets further exacerbate congestion



DCP+CC achieves the best performance under high loads

#3: Comparison with Timeout and RACK-TL



RACK-TLP performs better than IRN, but this comes at the cost of overhead from maintaining timestamps

Conclusion

- We present DCP, a transport architecture that rethinks RDMA reliability for lossy networks.
- By leveraging a lightweight lossless control plane in switches, DCP enhances the RNICs' reliability, enabling compatibility with packet-level LB, precise retransmission, and minimal memory and processing overhead
- Our prototype and evaluation show that DCP significantly outperforms existing RDMA solutions, advancing the practicality of high-performance RDMA over lossy fabrics.

Thank You!

Contact: wlicv@connect.ust.hk