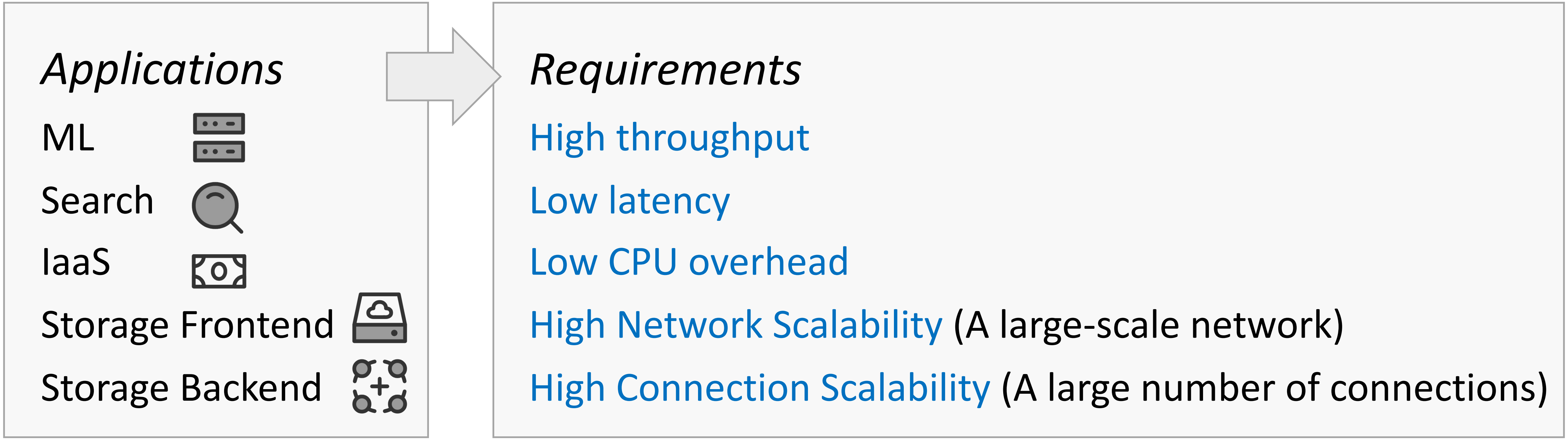
SRNIC: A Scalable Architecture for RDMA NICs

Zilong Wang, Layong Luo, Qingsong Ning, Chaoliang Zeng, Wenxue Li,
Xinchen Wan, Peng Xie, Tao Feng, Ke Cheng, Xiongfei Geng,
Tianhao Wang, Weicheng Ling, Kejia Huo, Pingbo An, Kui Ji, Shideng Zhang,
Bin Xu, Ruiqing Feng, Tao Ding, Kai Chen, Chuanxiong Guo

iSingLab HKUST

ByteDance

Datacenter applications’ requirements to the high-speed networks



Applications

ML

Search

IaaS

Storage Frontend

Storage Backend

Requirements

High throughput

Low latency

Low CPU overhead

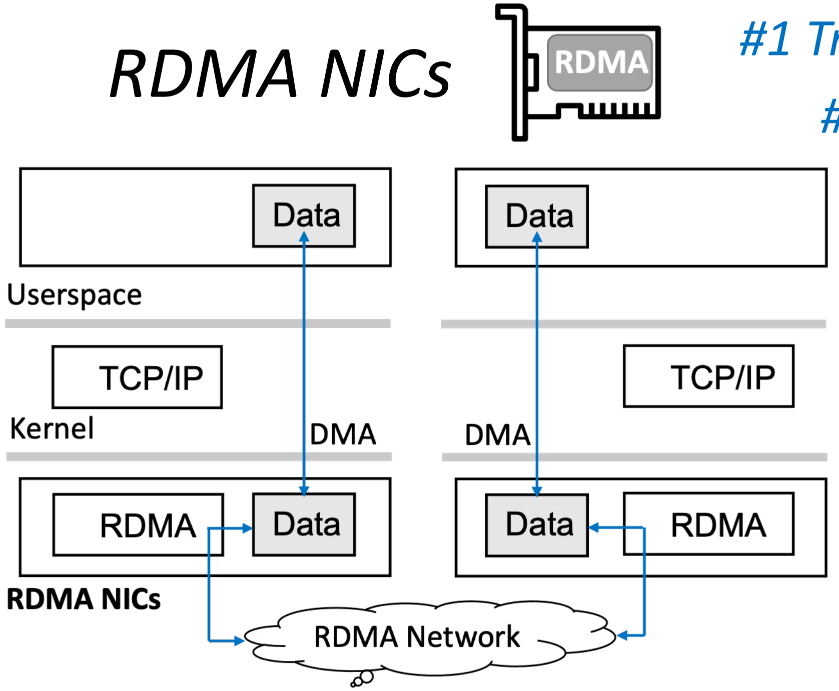
High Network Scalability (A large-scale network)

High Connection Scalability (A large number of connections)

RDMA becomes the de-facto standard for high-speed networks in modern datacenters

- High performance and low CPU overhead are inherent properties of RDMA

RDMA NICs



#1 Transport offload

#2 Kernel bypass

Requirements

- High throughput ✓
- Low latency ✓
- Low CPU overhead ✓
- High Network Scalability
- High Connection Scalability

Network Scalability Issue

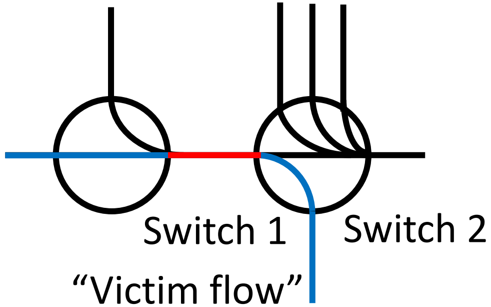
Lossless network requirement

PFC

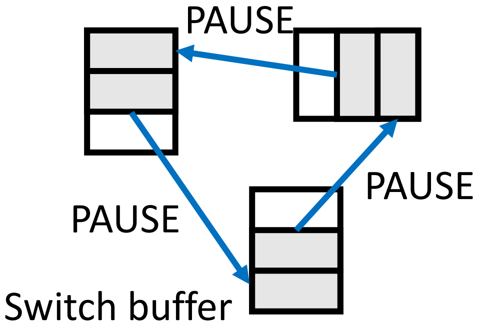
Management risks & network anomalies

Limited network scale

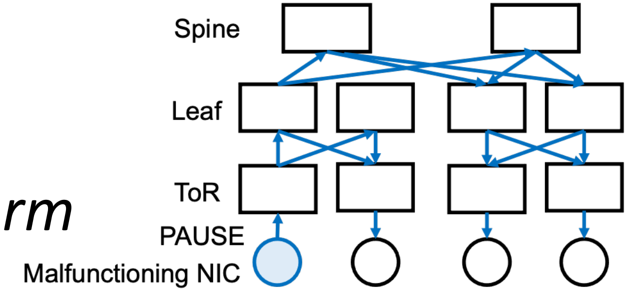
1. HoL Blocking



2. Dead Lock

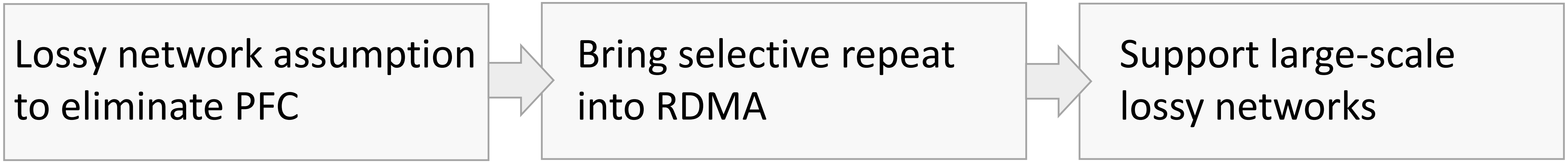


3. PFC Storm



Network Scalability Issue

- Exsiting work IRN addresses the network scalability issue



Lossy network assumption to eliminate PFC

Bring selective repeat into RDMA

Support large-scale lossy networks

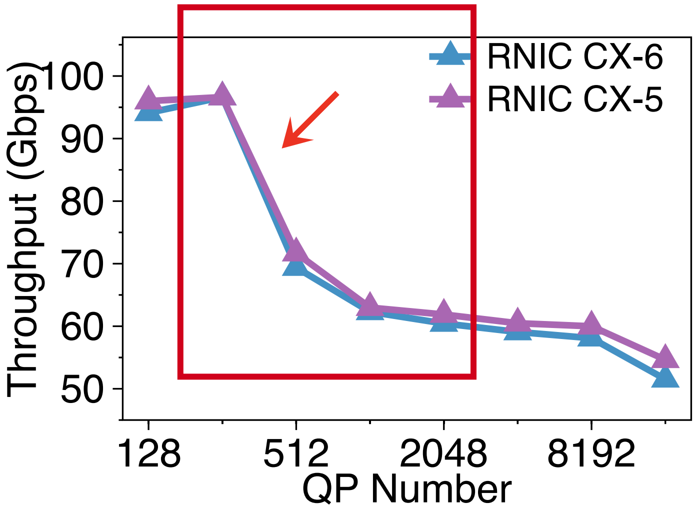
- However, IRN leaves connection scalability issue unsolved

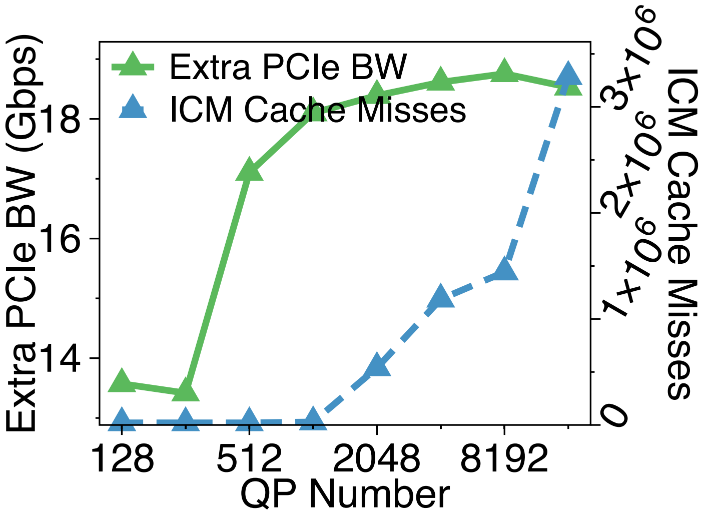
Connection Scalability Issue

Number of connections exceeds a certain threshold

Cache misses

RDMA performance collapse





SRNIC Design Goal

- Maximize the connection scalability, while preserving high performance and low CPU overhead as commercial RNICs, and maintaining high network scalability as IRN

Requirements	*Commercial RNICs*	*IRN*	*SRNIC*
High throughput	✓	✓	✓
Low latency	✓	✓	✓
Low CPU overhead	✓	✓	✓
High Network Scalability	✗	✓	✓
High Connection Scalability	✗	✗	✓

Idea & Design Principles

Idea Redesign the protocol and architecture to minimize the on-chip data structures and their memory requirements

Design Principles

1) Keep as many RDMA functionalities as possible in hardware
2) Assume lossy fabric, and handle packet loss as efficient as possible
3) Reduce the on-chip memory requirements as much as possible

Design Goals

- Preserve high performance and low CPU overhead
- Maintain high network scalability
- Maximize connection scalability

Strategies to Minimize On-chip Data Structures

Three Steps

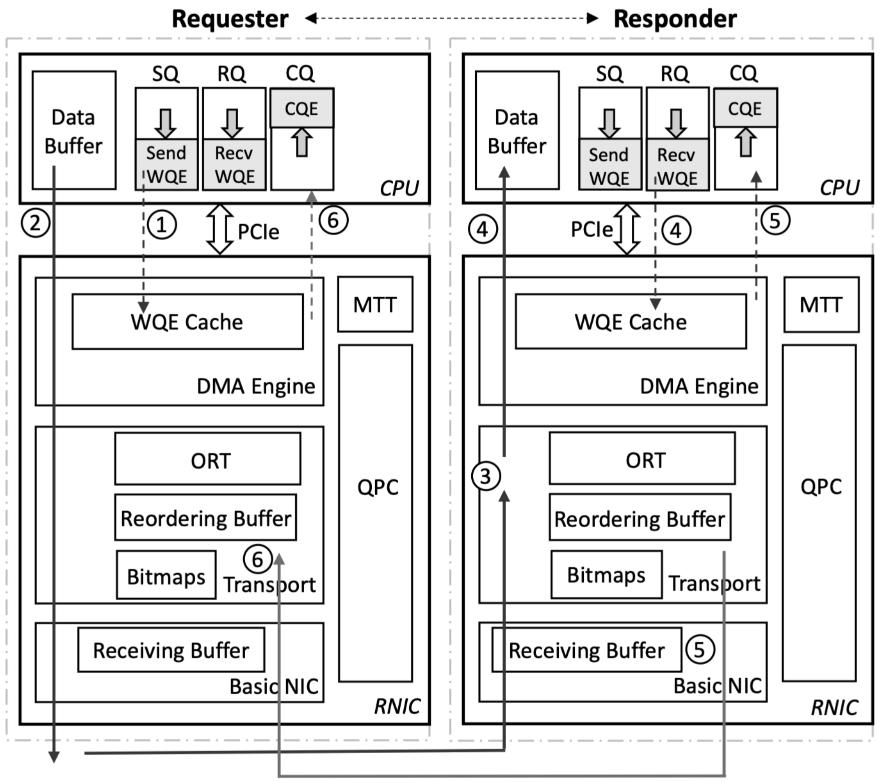
Analyze the data structures involved in typical RDMA data flows using an RDMA conceptual model

Analyze data structures for on-chip storage or removal

Co-design protocol and architecture for minimal on-chip memory requirements

Data Structures Involved

RDMA Conceptual Model
& RDMA Data Flow



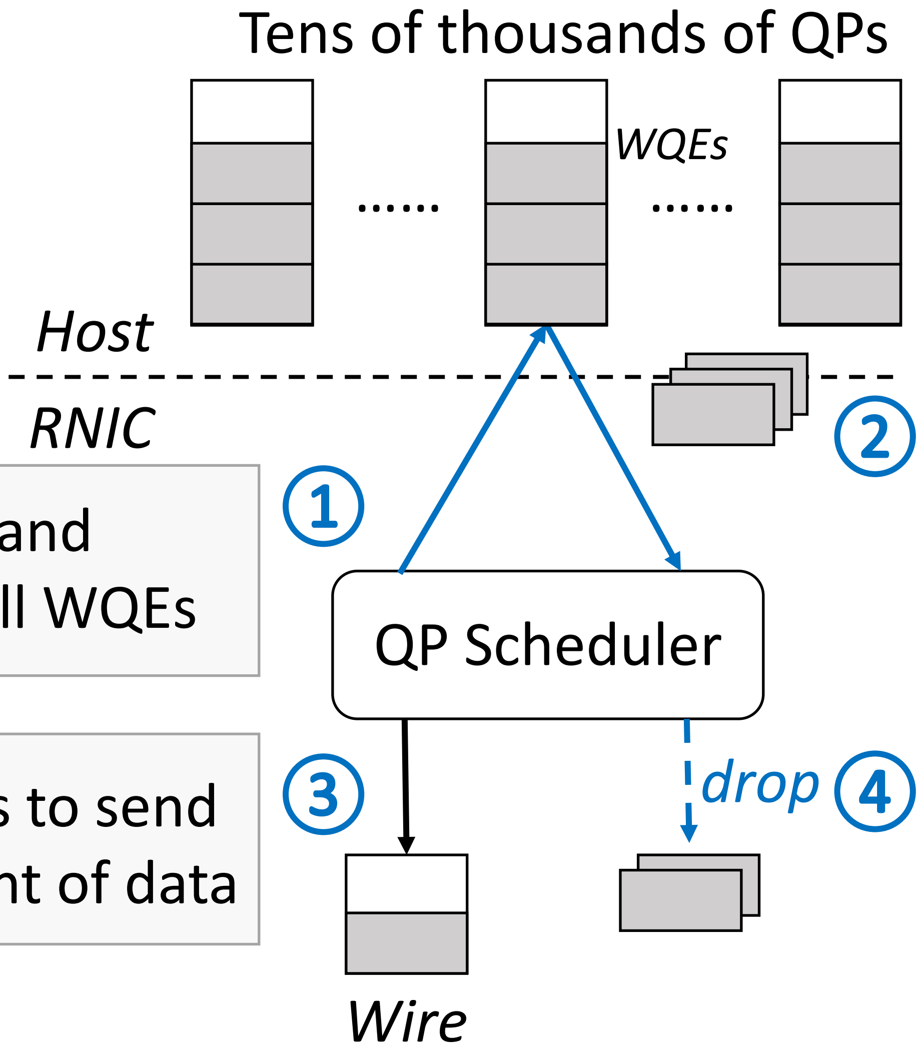
Data Structures

1. Receiving Buffer
2. QP Context (QPC)
3. Memory Translation Table (MTT)
4. WQE Cache
5. Bitmap
6. Outstanding Request Table (ORT)
7. Reordering Buffer

Analysis & Optimization Ideas

Data Structures	*Status*	*Optimization Ideas*	
• Receiving Buffer	On-chip	Must be stored on-chip	
• QPC	On-chip		
• MTT	On-chip		
• WQE Cache	Removed	Cache-free QP Scheduler	(§1)
• Bitmap	Removed	Bitmap onloading	(§2)
• ORT	Removed	Header extensions	(§3)
• Reordering Buffer	Removed		

§1 Cache-free QP Scheduler



1 Schedule QPs and proactively pull WQEs

2 Fetch a few WQEs for batch transactions

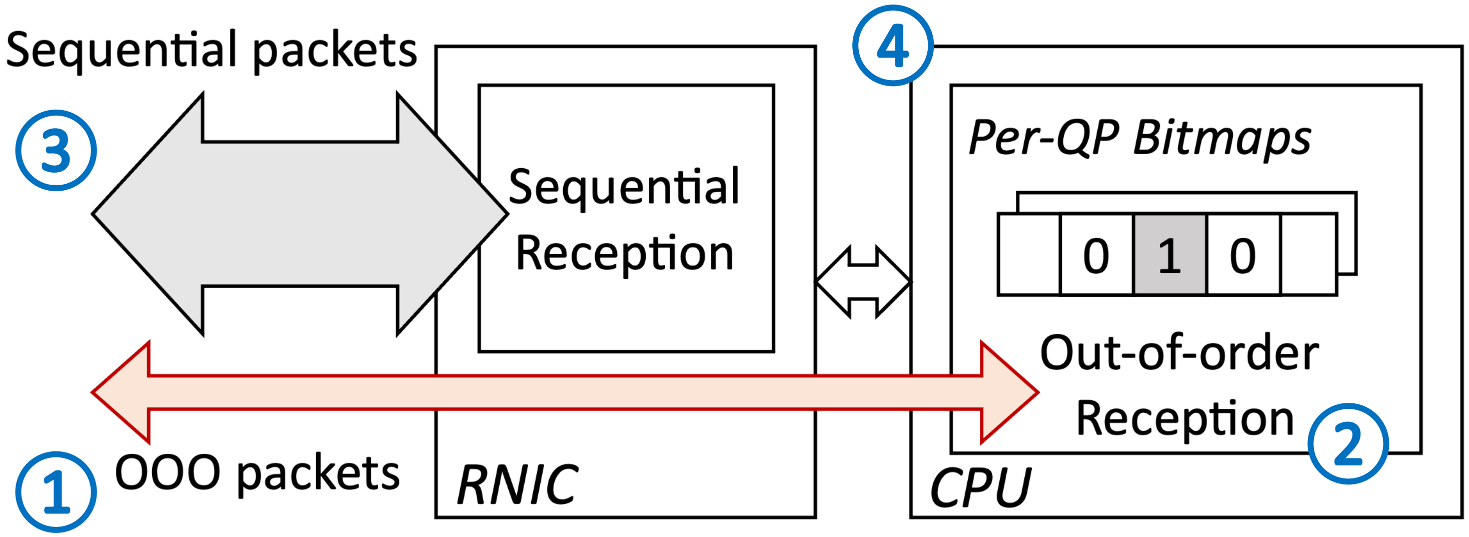
3 Consume WQEs to send a certain amount of data

4 Drop unused WQEs instead of cache them

§2 Bitmap Onloading

Sequential packets are the majority

Implement sequential reception logic on-chip and onload bitmaps for out-of-order reception



Balance the performance and memory requirements

Packet loss results in out-of-order packets

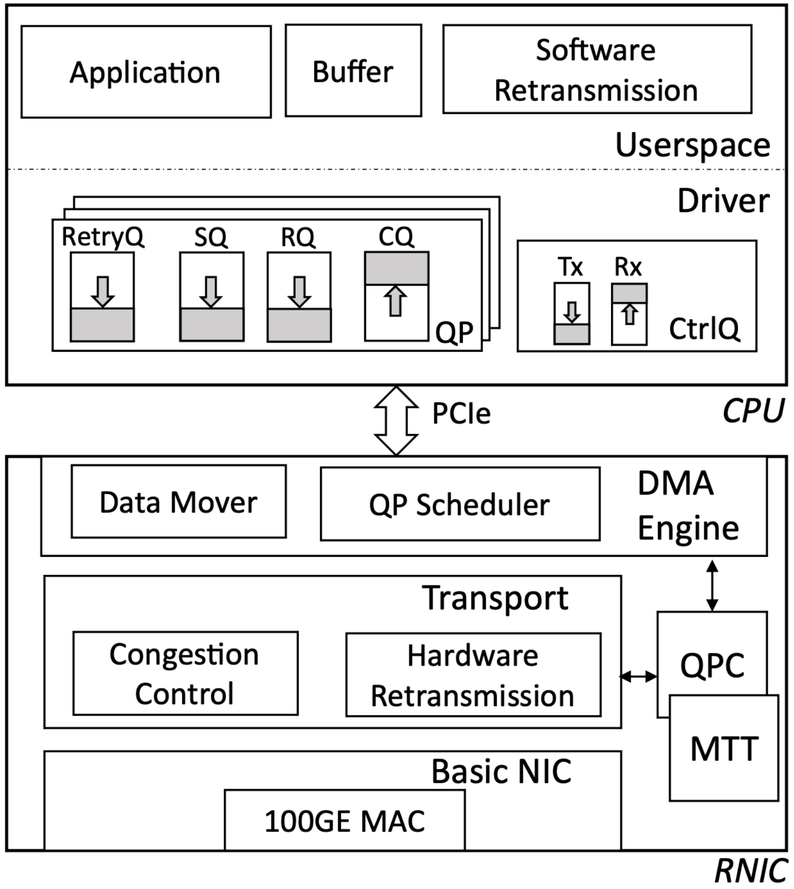
Bitmaps are only needed for out-of-order reception

§3 Header Extensions

- "RDMA does not support selective packet retransmission nor the out-of-order reception of packets", from IB specification
  - Sticking to the RDMA format in lossy RDMA hurts connection scalability
  - Need to revise the RDMA protocol and extend the packet header for a lossy network

- Carry the metadata of loss recovery in request headers and let responses echo back
  - Requester can recover the lost requests with these metadata in the response header without the need of outstanding request table
- Carry the information of data buffer address in packet headers
  - Receiver can directly place all incoming packets into the user buffer at correct addresses and remove reordering buffer

SRNIC Design



- SRNIC, a scalable RDMA NIC architecture with protocol and architecture co-design

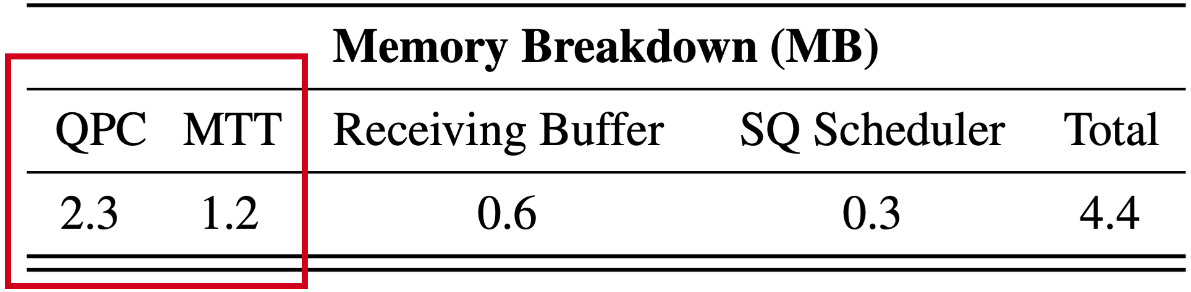
- Minimized on-chip memory requirements:
  1) QPC, and 2) MTT

- Function modules:
  1) DMA Engine, 2) Transport, and 3) Basic NIC

Prototype Implementation

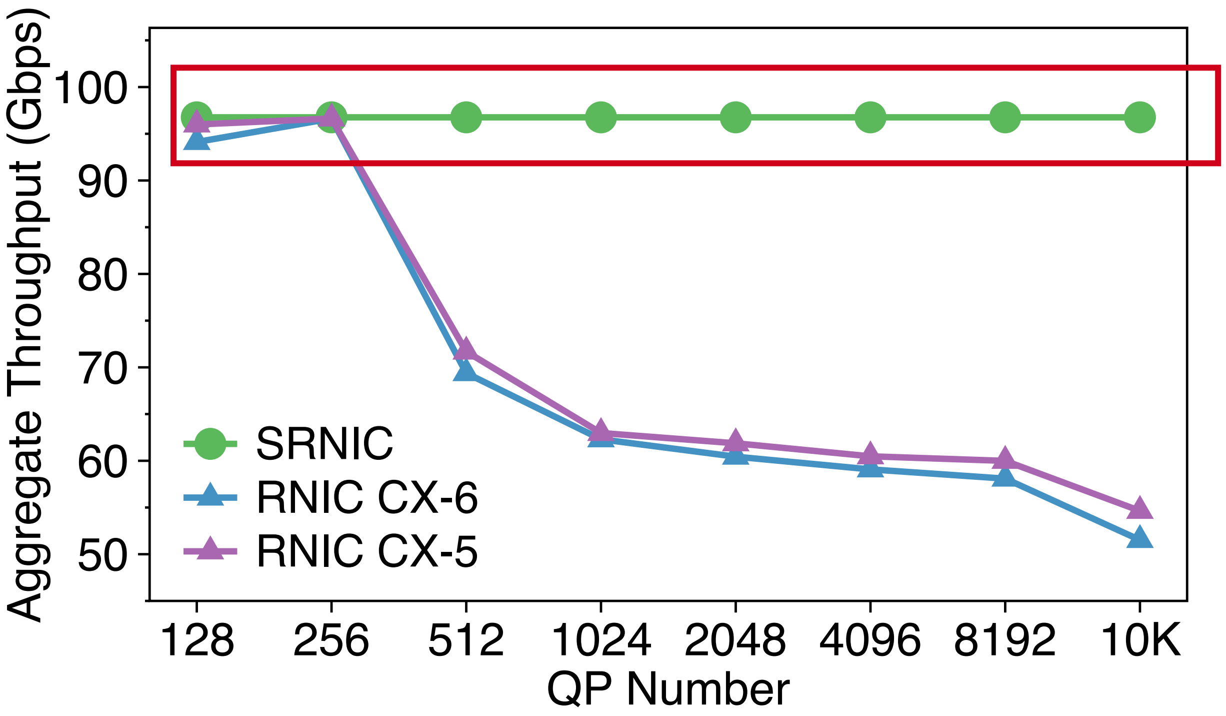
- Prototype setting
  - Xilinx FPGA with a clock frequency of 300 MHz

- Resource usage
  - 4.4 MB in total
  - 2.3 MB QPC for 10K QPs, 1.2 MB MTT cache



Memory Breakdown (MB)				
QPC	MTT	Receiving Buffer	SQ Scheduler	Total
2.3	1.2	0.6	0.3	4.4

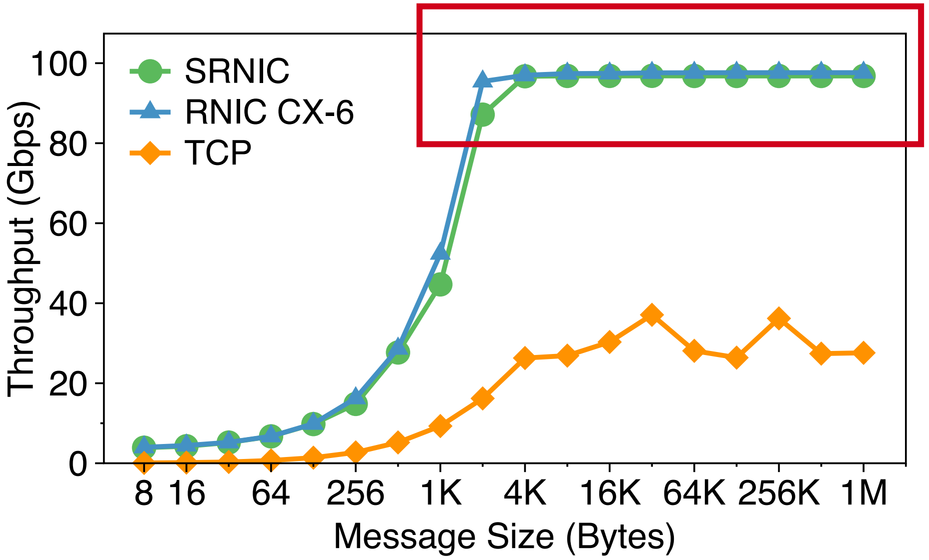
Evaluation – Connection Scalability



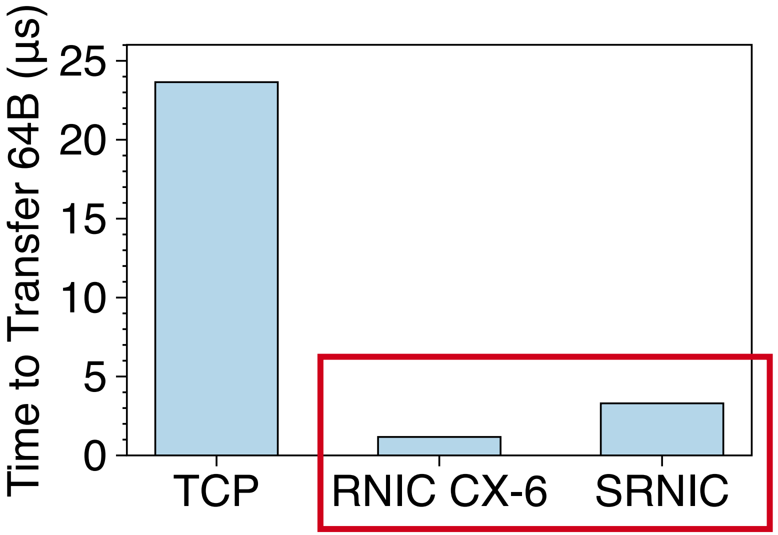
SRNIC achieves high connection scalability

- 10K performant connections, 18x higher than CX-5 in terms of normalized connection scalability

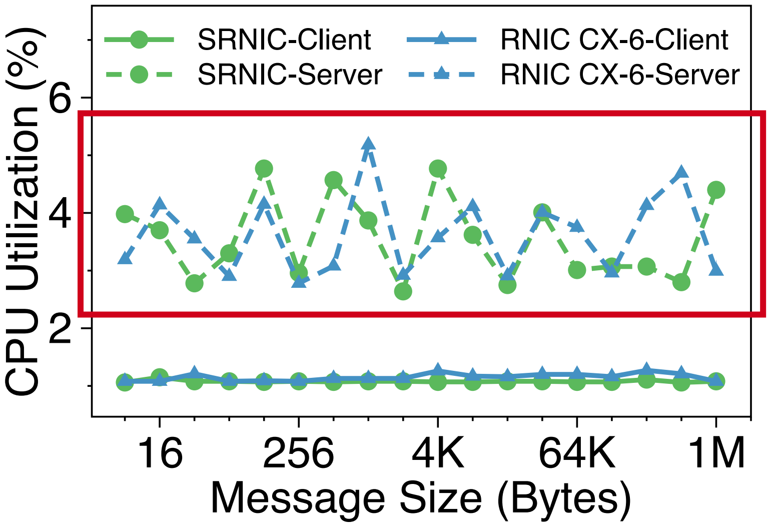
Evaluation – Performance & CPU Overhead



Throughput



Latency

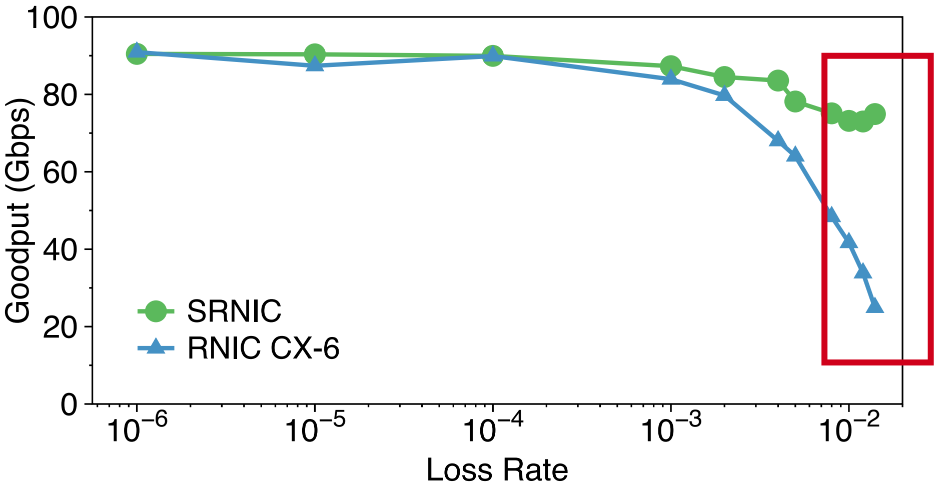


CPU Overhead

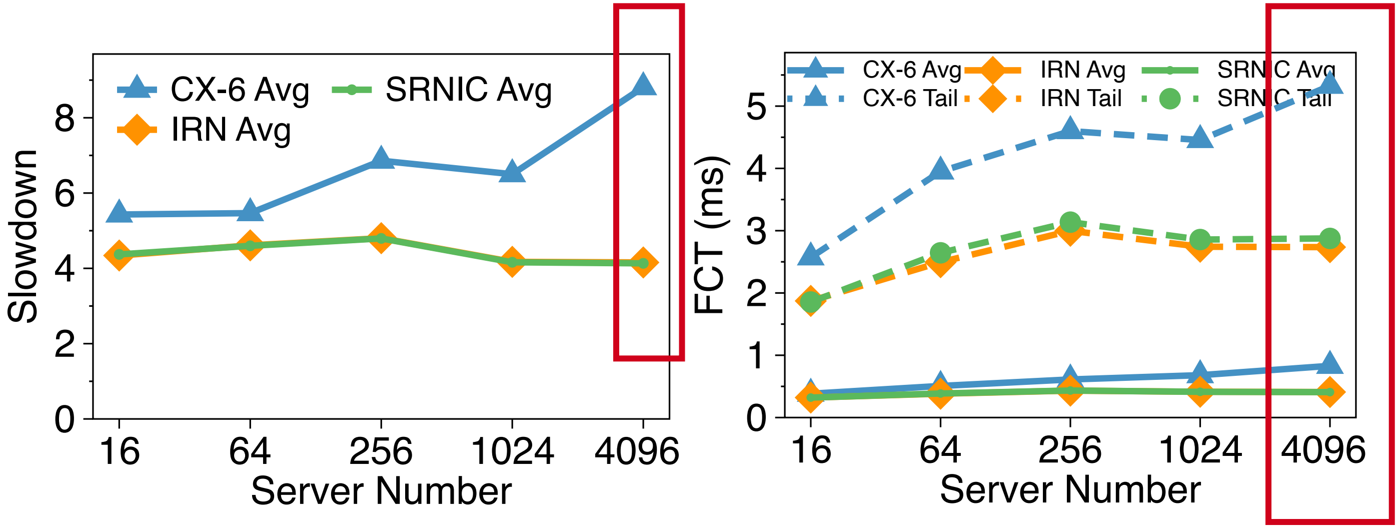
SRNIC preserves high performance and low CPU overhead as commercial RNICs

- 97 Gbps throughput
- 3.3 us latency
- 5% CPU overhead

Evaluation – Network Scalability



Loss Tolerance



Performance over Lossy Networks

SRNIC maintains high network scalability as IRN

- High loss tolerance, 3x higher than CX-6
- Stable performance over large-scale lossy networks

Summary

- We design SRNIC, a scalable RDMA NIC architecture, addresses connection scalability challenges while achieving high performance, low CPU overhead, and high network scalability. SRNIC satisfies all five requirements to high-speed networks.
- We hope our early attempts can inspire the re-design of a new RDMA specification for lossy network

Thank You!

Contact email: zwangfb@cse.ust.hk