SRC: A Scalable Reliable Connection for RDMA with Decoupled QPs and Connections

Yiren Zhao, Ran Shu, Yongqiang Xiong

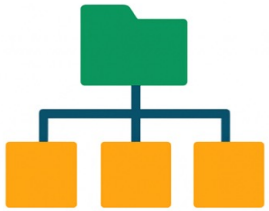
Microsoft Research

RDMA is Widely Used

- RDMA offers high throughput, low latency, and low CPU overhead
- Widely used in OLTP, KV Stores, Distributed File Systems, and ML

OLTP

KV Stores



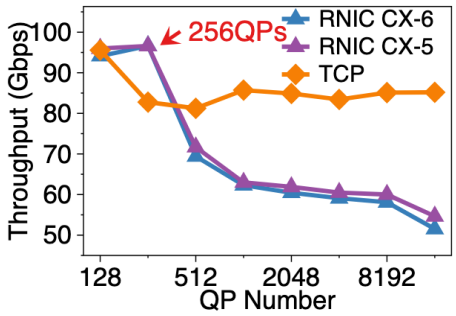
Distributed File Systems

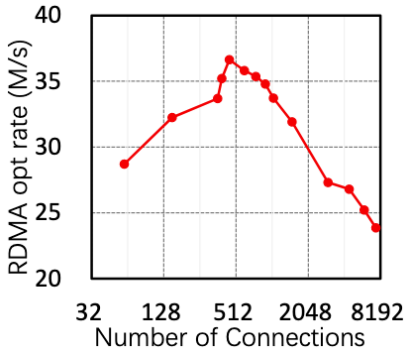
ML

RNIC Scalability Issue

- RNIC states can be **GB-scale**
  - **Hundreds of bytes per QP** state size, plus other states (MR state, transport state)
- Only **a few MBs** RNIC on-chip **SRAM**, used as cache

→ **Cache miss happens**





Performance drops at scale due to cache miss [1, 2]

[1] SRNIC: A Scalable Architecture for RDMA NICs, NSDI'23
[2] StaR: Breaking the Scalability Limit for RDMA, ICNP'21

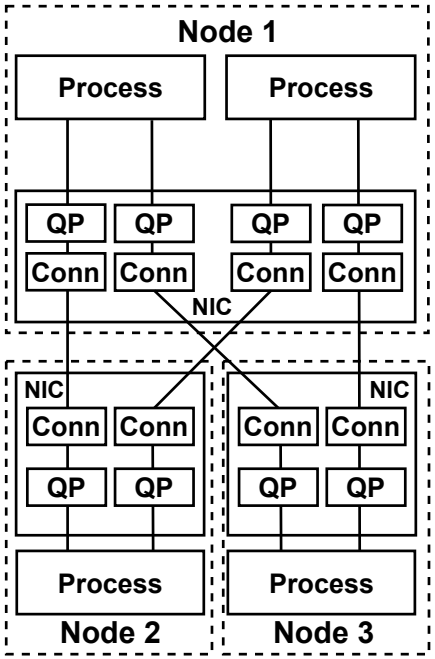
Reliable Connection

- Reliable Connection (RC) has strong guarantees of reliability, in-order message delivery, support for a full set of RDMA operations
- RC is the most desired and widely used
- Use RC requires a lot of connections

Assuming a cluster with N nodes, each node runs P processes

Reliable Connection (RC)

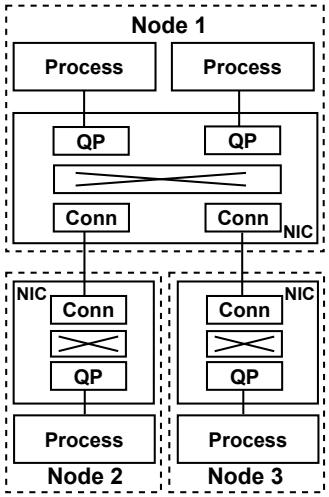
- 1 QP per connection
- **Total QPs: P x P x (N-1)**
- **Poor scalability**



RC

Existing Solutions

- The states are reduced by reducing the number of QPs

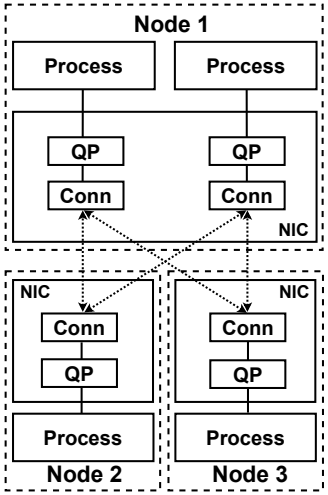


Software QP Sharing

Multiplexed in user space

Total: P x (N-1)

Scalable, but adds CPU overhead



DCT

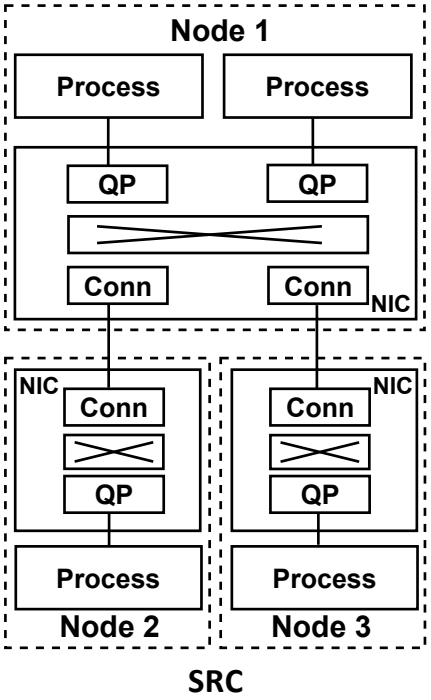
Dynamic HW-based sharing

slightly more than P

frequent connection switching

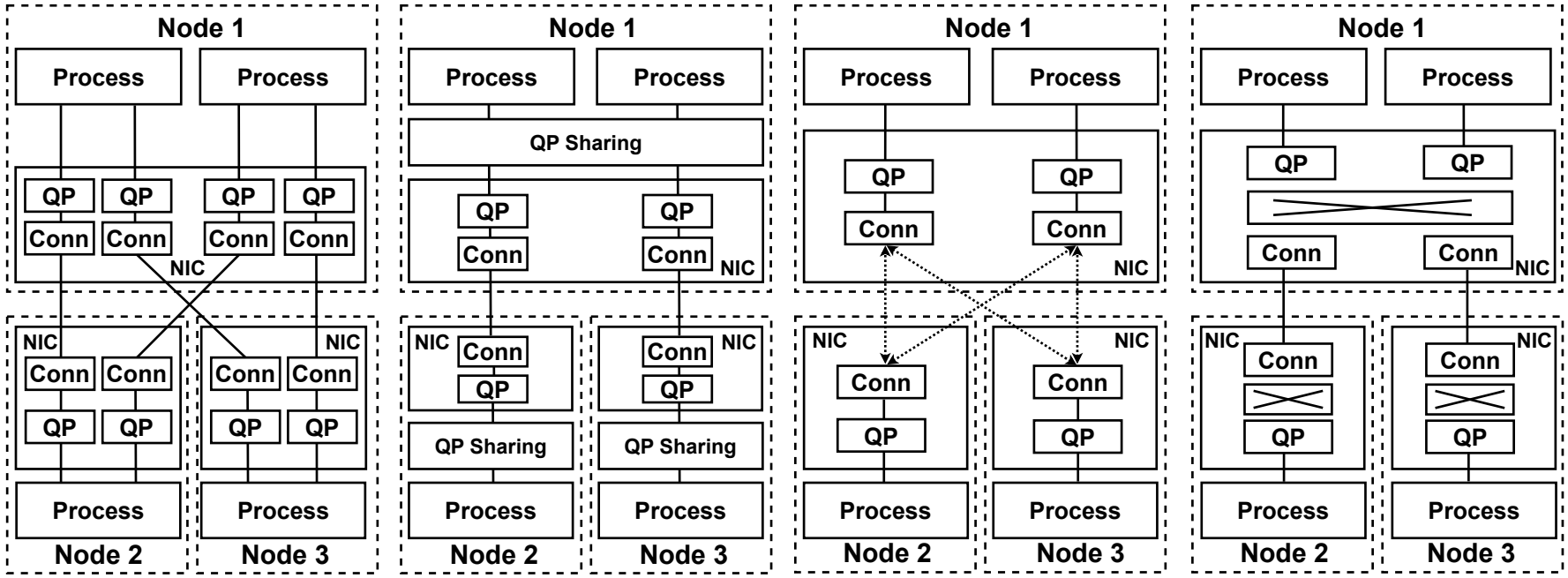
Our Solution: SRC

- Scalable Reliable Connection (SRC) decouples Queue Pairs (QPs) from network connections
  - **QPs** act as host-to-RNIC channels
  - **Connections** are maintained only between RNICs



SRC: decoupling QPs and connections
(HW-based connection sharing)

Our Solution: SRC



(a) RC (b) Software-based sharing (c) DCT (d) SRC

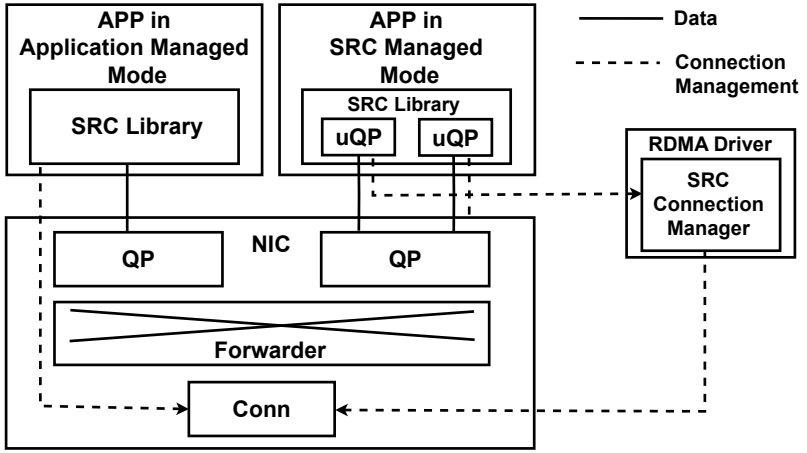
SRC: decoupling QPs and connections
(HW-based connection sharing)

Challenges

- Challenge #1: Minimize Additional States
- Challenge #2: Efficient State Lookup
- Challenge #3: Efficient and Friendly Abstraction

SRC Design

- Software managed mapping between QP and connection
- Software selects the connection for the request, to eliminate the lookup overhead on hardware
- Two operation modes
  - SRC-managed mode:
    almost compatible with RC
  - App-managed mode:
    advanced management capability



Preliminary Results

Nodes (N)	RC (MB)	XRC (MB)	DCT (MB)	SRC (MB)
32	8.869	0.443	0.014	0.019
128	36.335	1.817	0.014	0.053
512	146.198	7.310	0.014	0.190

- SRC reduces RDMA state size from 146.198 MB to 0.190 MB in a 512-server cluster running RDMA applications.
- Comparable to DCT but without its performance penalty.

Conclusion

- SRC decouples QPs and connections to improve scalability
- SRC provides an efficient software and hardware architecture
  - Software managed mapping between QP and connection
  - A lightweight mapping scheme for efficient forwarding between QPs and connections on the RNIC
  - Two operation modes for either compatibility or capability

SRC: A Scalable Reliable Connection for RDMA with Decoupled QPs and Connections

Thanks for Your Listening 🎉