LubeRDMA:
A Fail-safe Mechanism of RDMA

Shengkai Lin, Qinwei Yang,

Zengyin Yang, Yuchuan Wang, **Shizhen Zhao**

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

HUAWEI

Remote Direct Memory Access (RDMA)

LLM APP

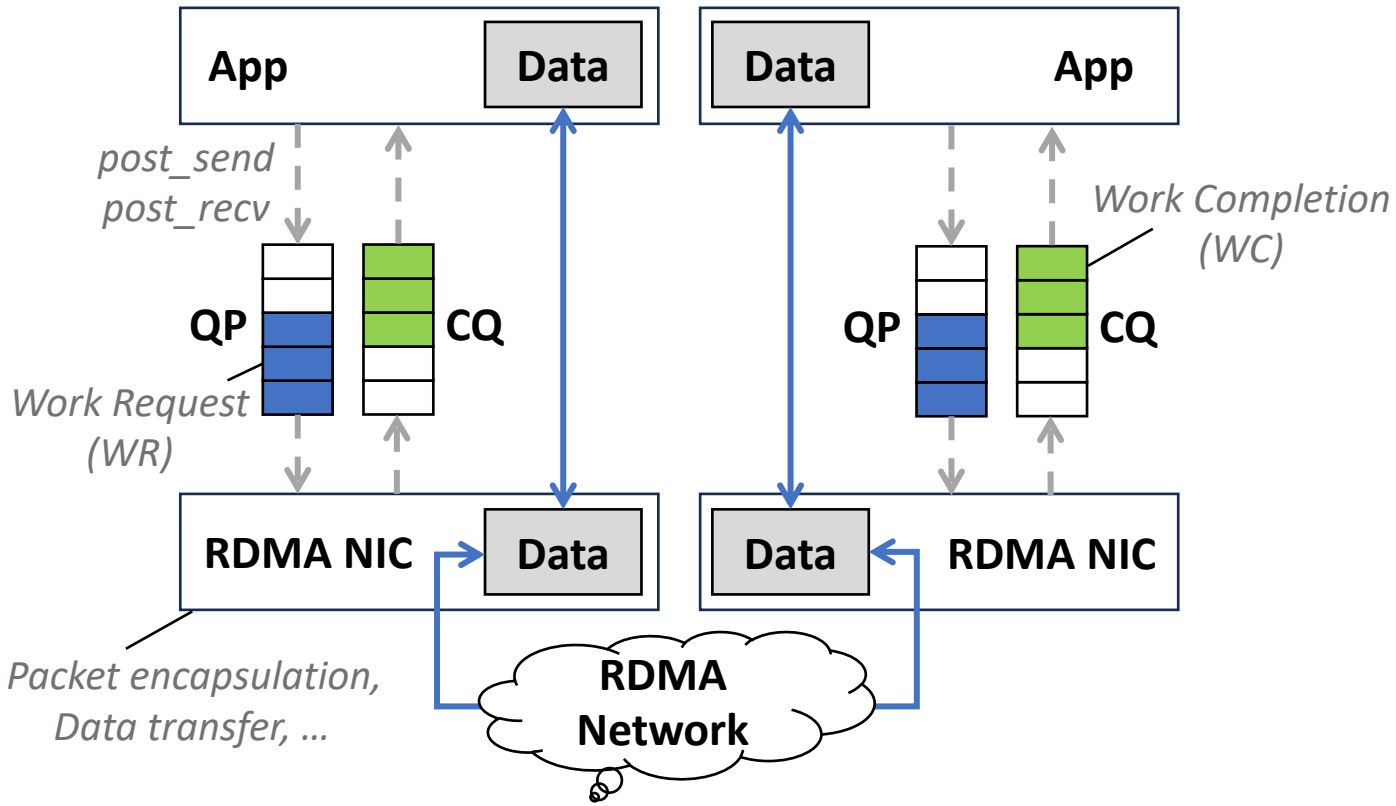
Gemini

Claude

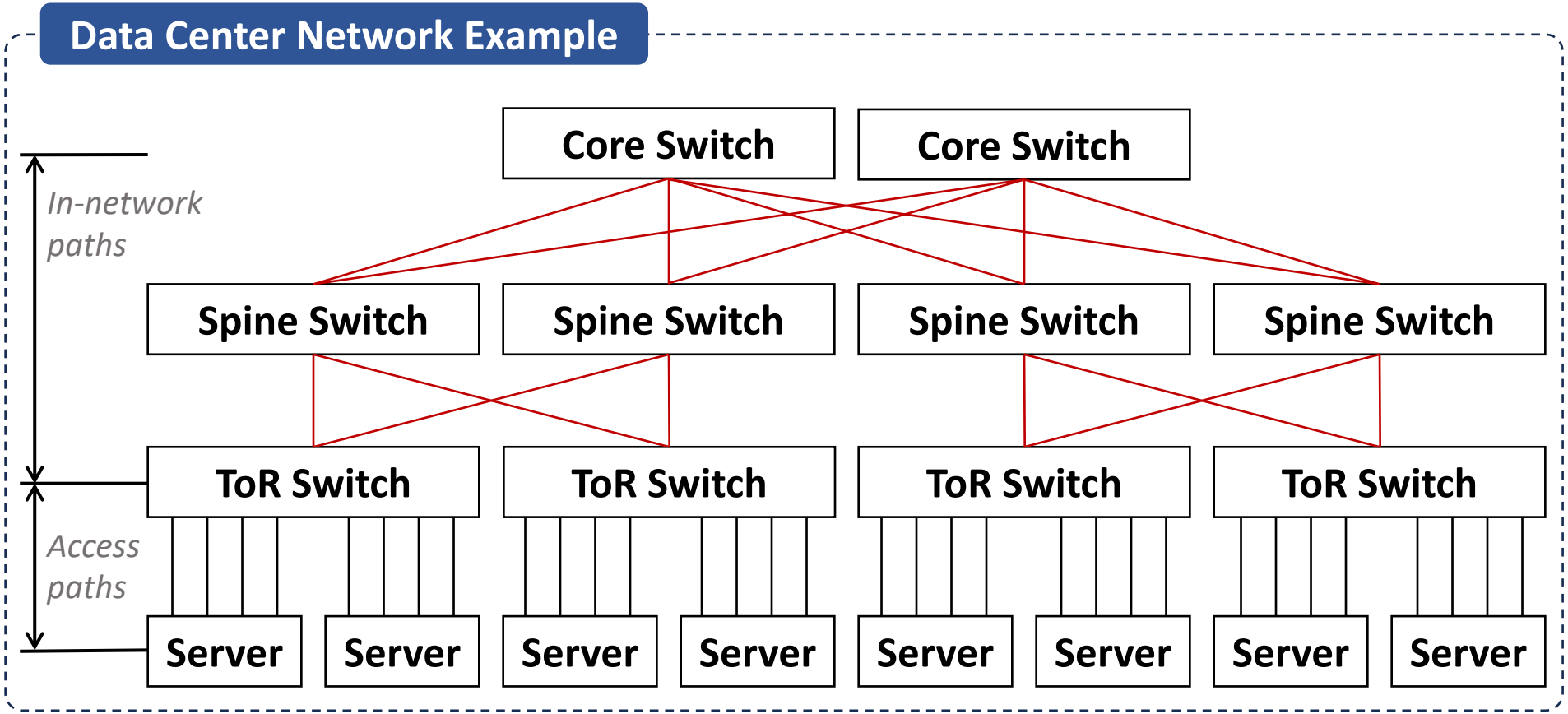
Requirements

- High throughput
- Low latency
- Low CPU overhead

Typical RDMA Workflow



Failures in RDMA Network



Larger application,

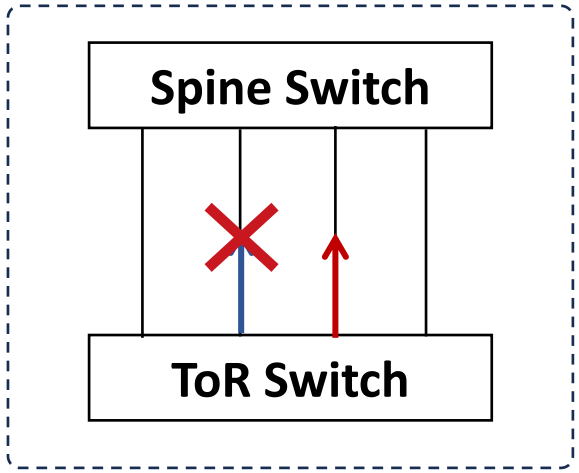
More failures.

Failure types:

- Optical Modules
- RDMA NICs
- Links
- Switches

Current RDMA Network is Insufficient

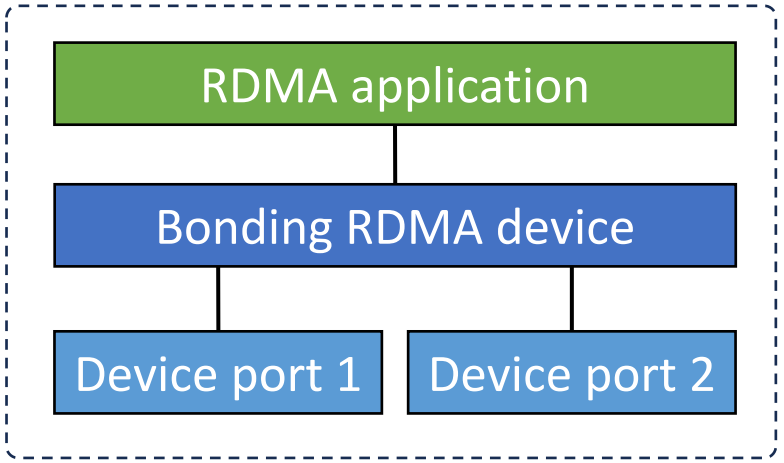
- ***In-network Rerouting***



Reroute packets to equivalent links

Can only tolerate **in-network** failures.

- ***RoCE LAG (Link Aggregation)*** [1]



Bonds the two ports of an RNIC together

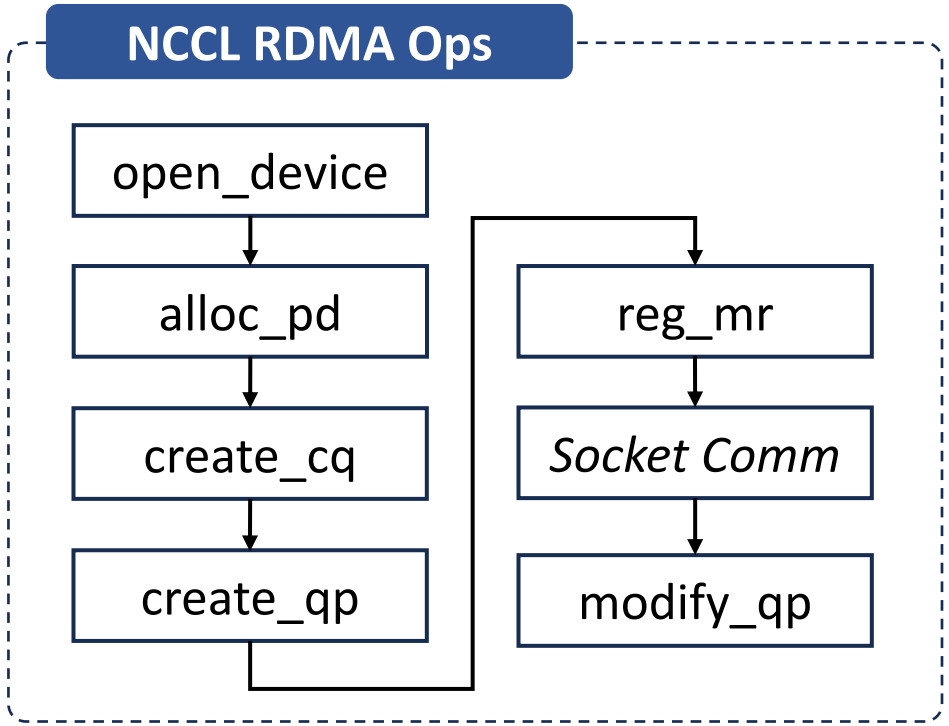
Double cost of access links.

Cannot tolerate the **whole RNIC** failure.

Access layer failures are *Single Point of Failures (SPOF)*

[1] https://enterprise-support.nvidia.com/s/article/How-to-Configure-RoCE-over-LAG-ConnectX-4-ConnectX-5-ConnectX-6

LubeRDMA: Intuition & Challenges



Intuition

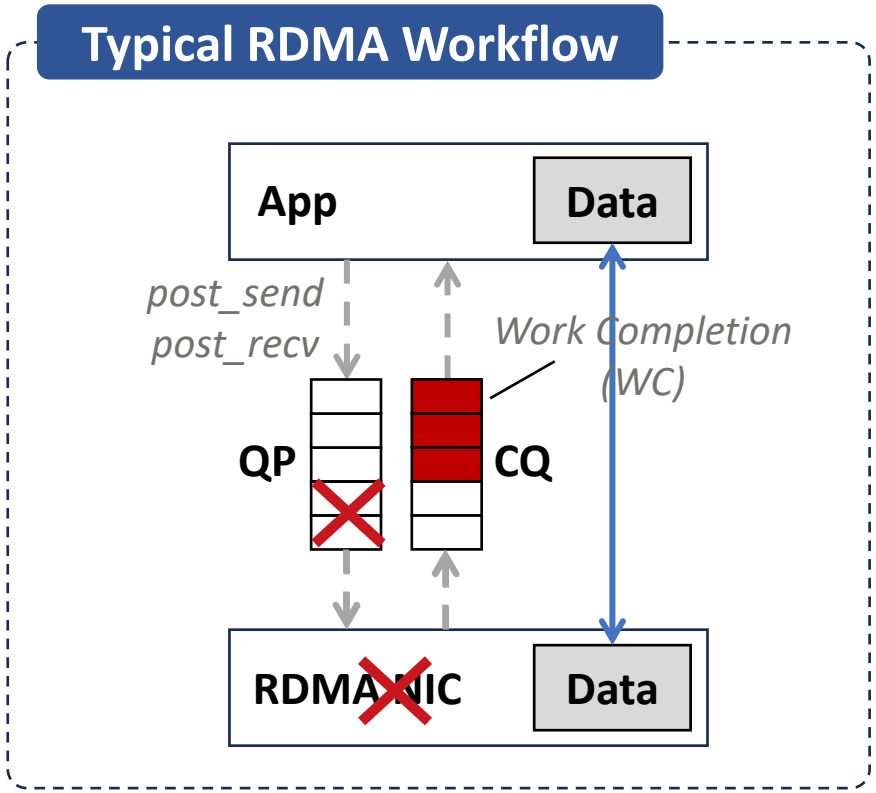
Take multiple RNICs as *backup* of each other
Handle RDMA failures in *RDMA layer*.

Challenges

① Should manage backup RNICs independently

1. *Control verbs.*
2. *Match local & remote QPs and exchange some attributes between servers through a socket, such as GID, QPN, rkey.*

LubeRDMA: Intuition & Challenges



Intuition

Take multiple RNICs as *backup* of each other

Handle RDMA failures in *RDMA layer*.

Challenges

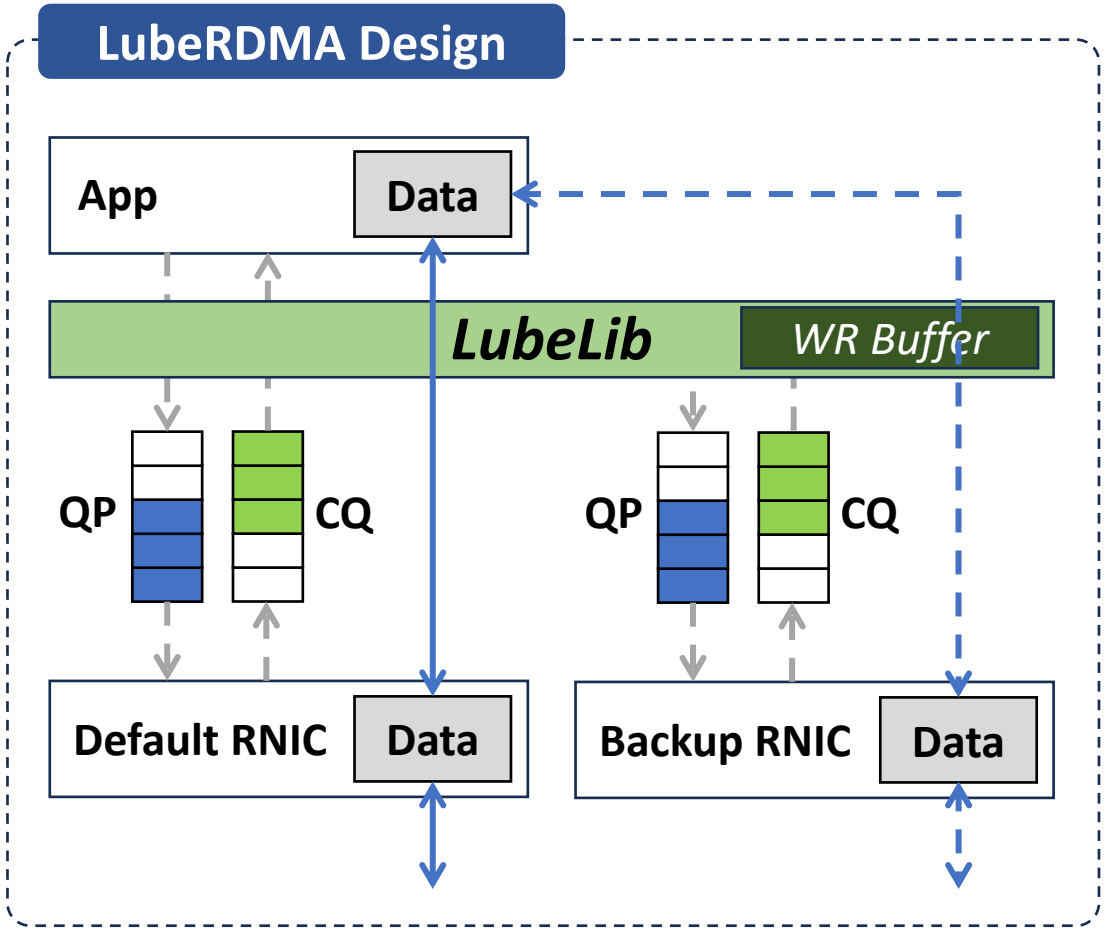
① Should manage backup RNICs independently

1. *Control verbs.*
2. *Match local & remote QPs and exchange some attributes between servers through a socket, such as GID, QPN, rkey.*

② Ensure RDMA traffic consistency during failures.

③ Performance & overhead are still primary concern.

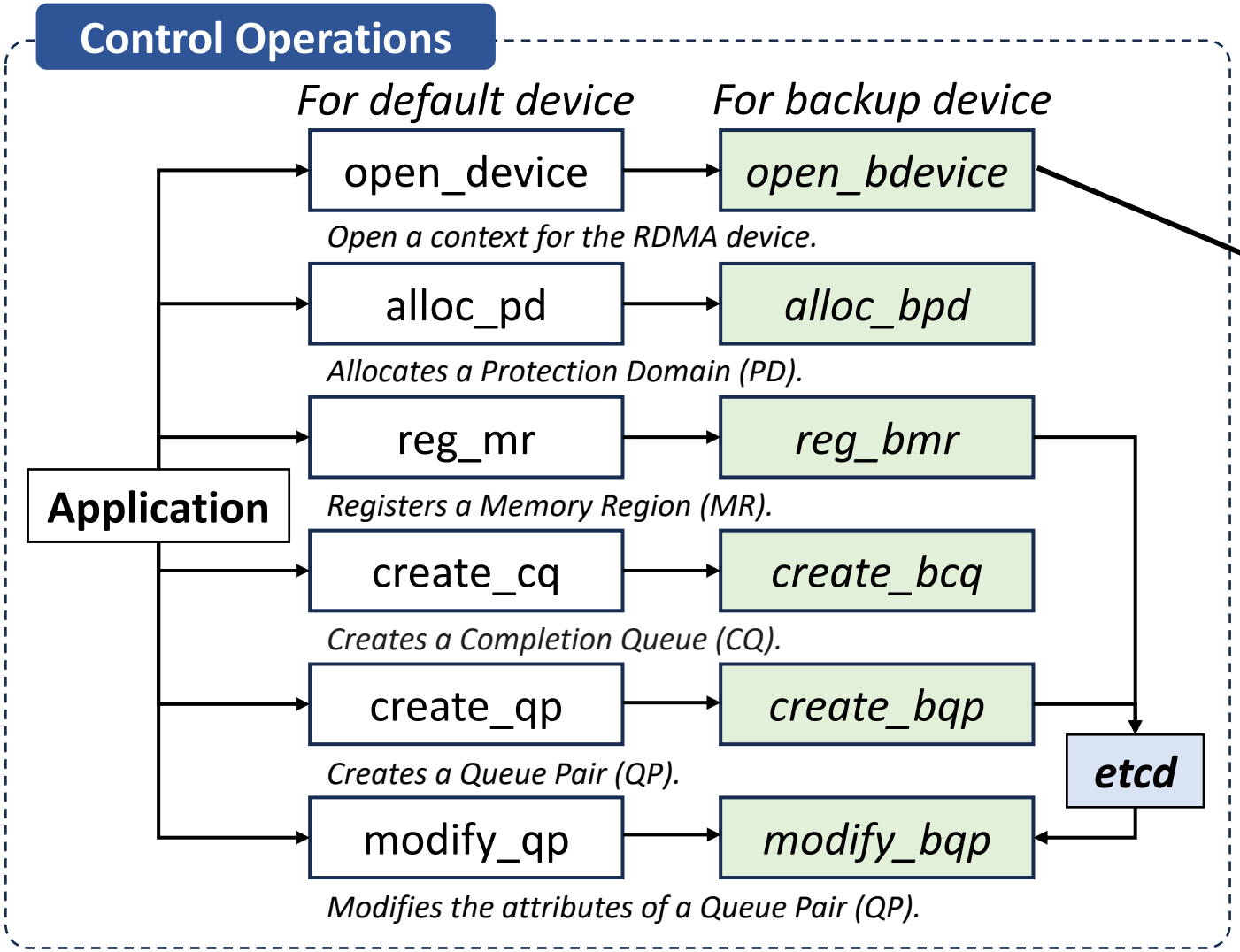
LubeRDMA:
A Fail-safe Mechanism of RDMA



Main Design

1. Shadow control verbs & resources
2. Failure-resilient data verbs

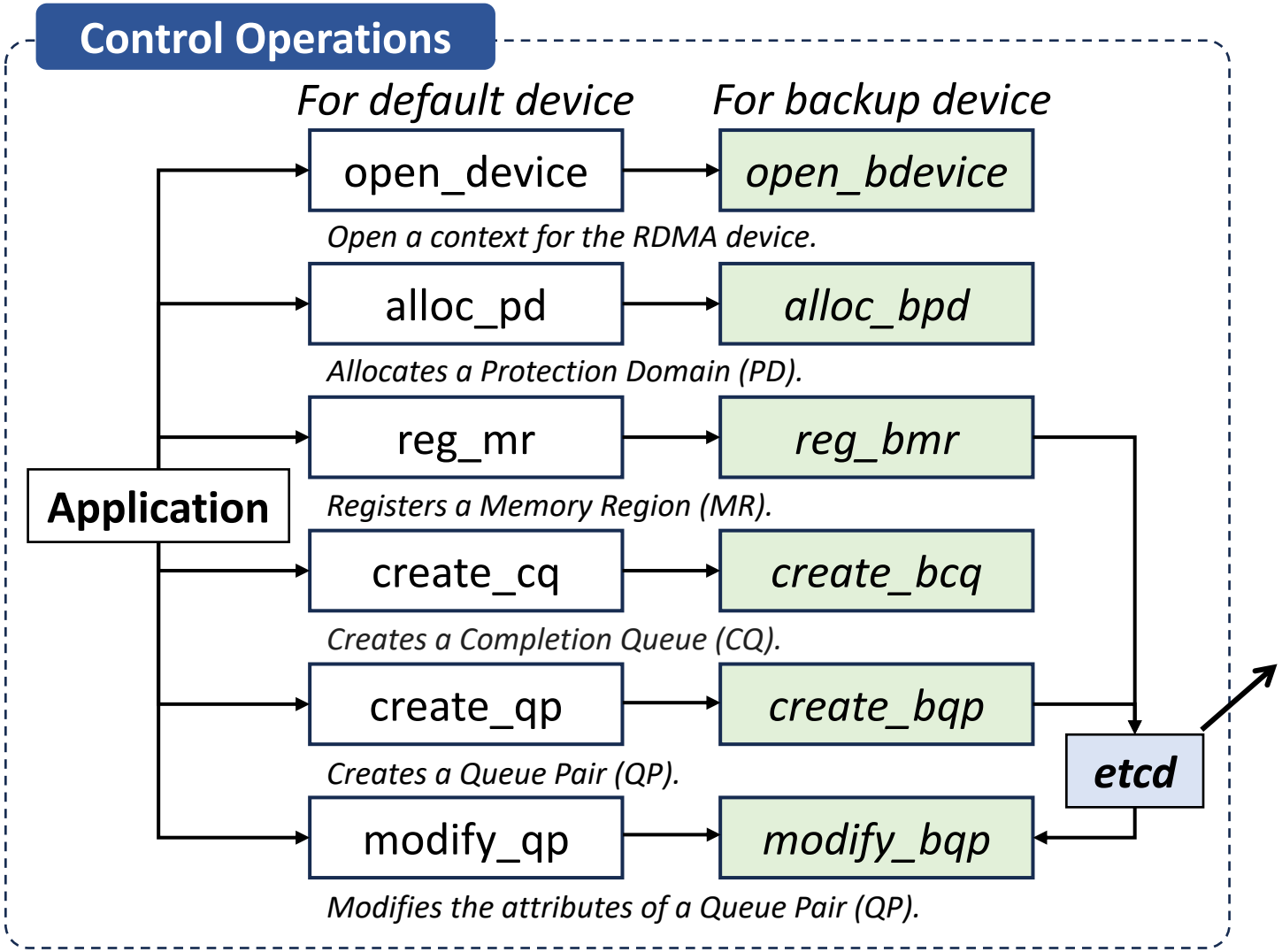
LubeRDMA:
Shadow control verbs & resources



① Shadow Control Verbs

- Initialize and configure the backup RNIC with the same control verbs.
- Call control verbs for backup RNIC within control verbs for default RNIC.

LubeRDMA:
Shadow control verbs & resources

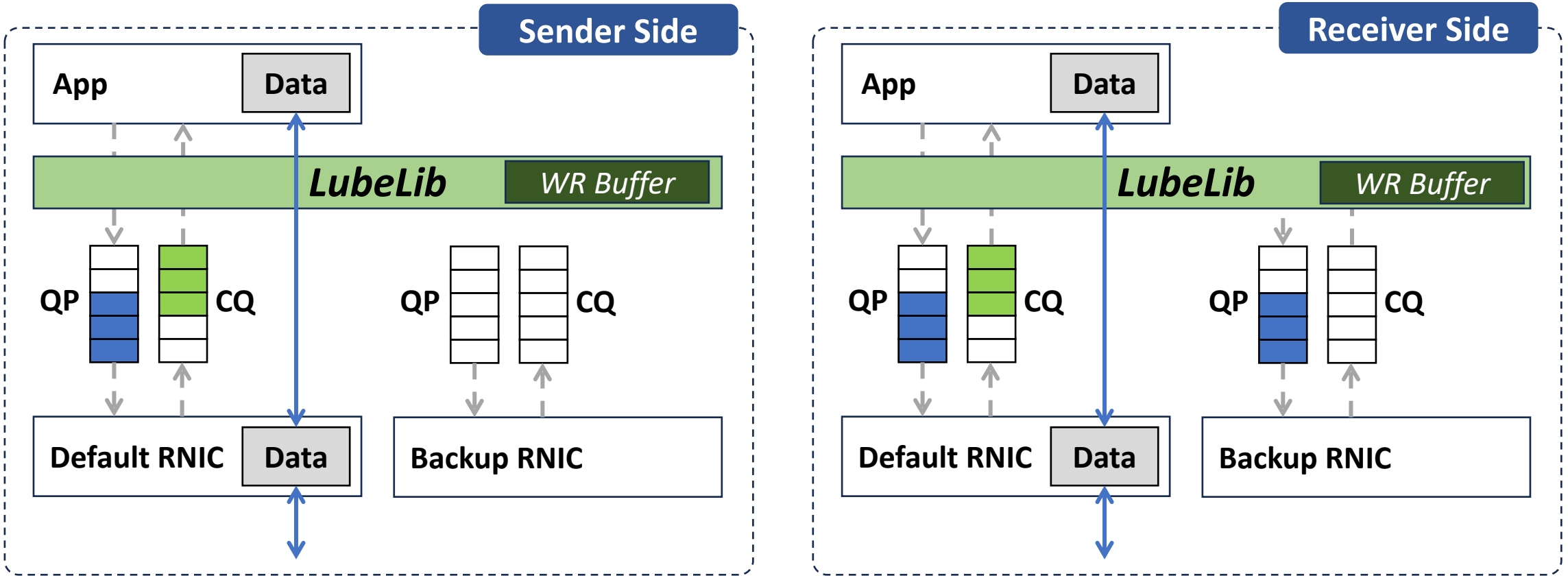


② Shadow QP Configuration Attributes

- Match & exchange configuration attributes for backup QP by KV storage.
- Store & query backup attributes by default ones

Key	Value
Default GID & QPN	*Backup* GID & QPN
Default MR rkey	*Backup* MR rkey

LubeRDMA: Failure-resilient data verbs

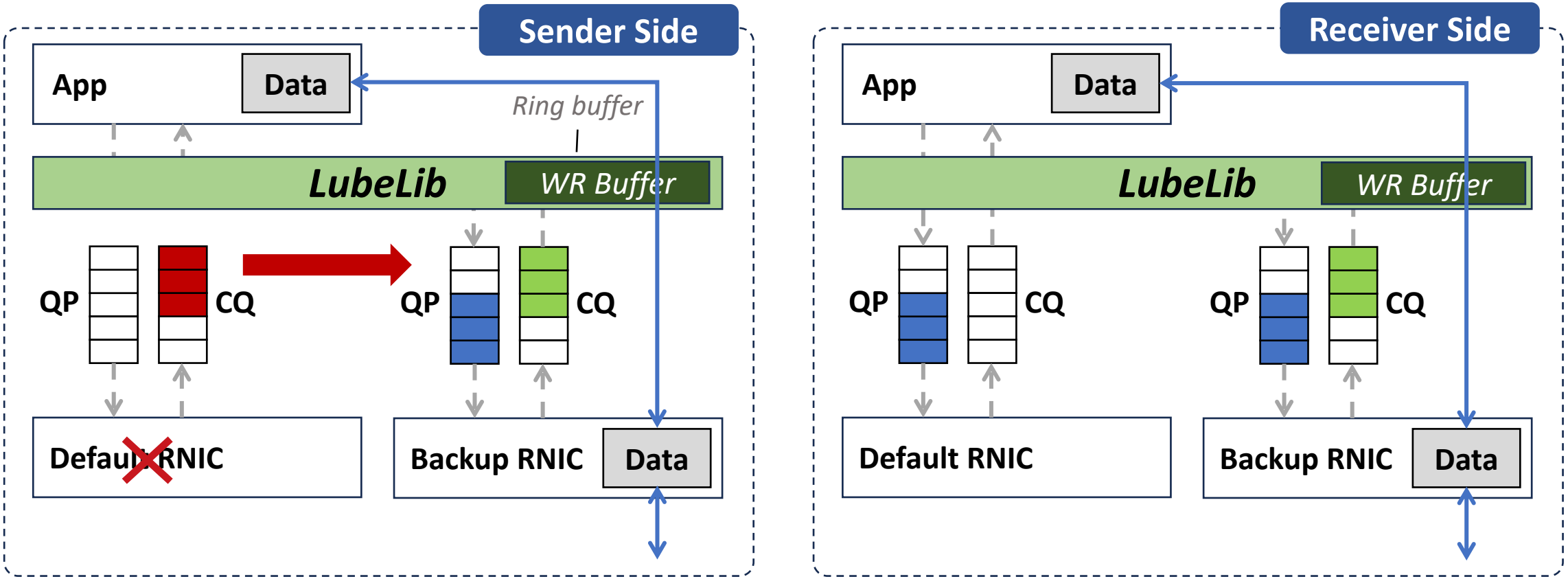


1. Post WRs to default QP;
2. **Buffer outstanding WRs**

1. Usual Status

1. **Post WRs to ALL QPs**;
2. Buffer outstanding WRs

LubeRDMA: Failure-resilient data verbs

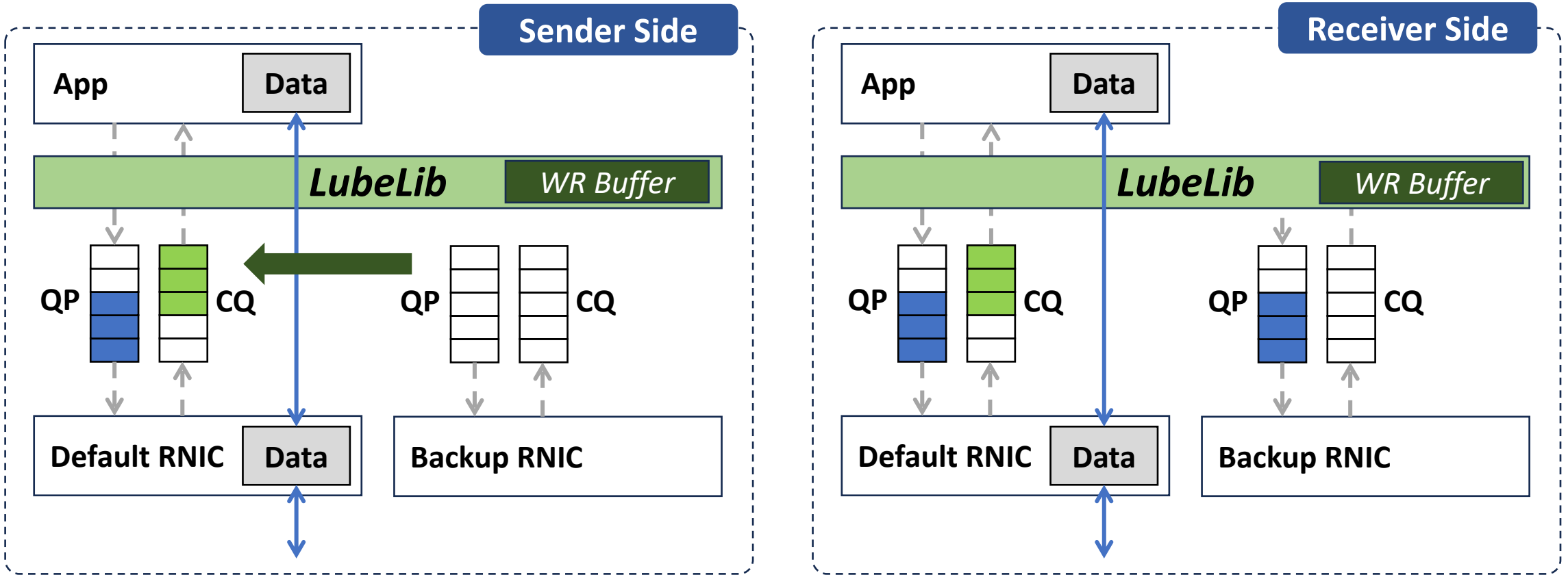


1. Repost WRs to backup QP;
2. Buffer outstanding WRs;
3. Switch traffic to backup RNIC.

2. Failover Process

1. Post WR to ALL QPs;
2. Buffer outstanding WRs

LubeRDMA: Failure-resilient data verbs

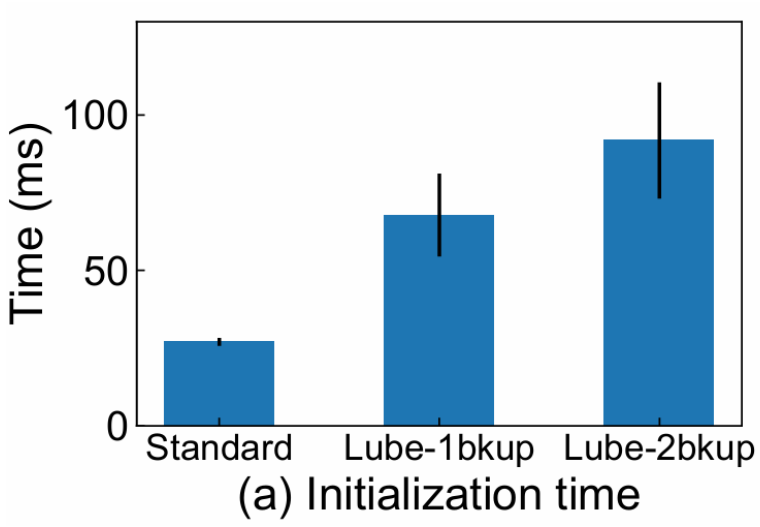


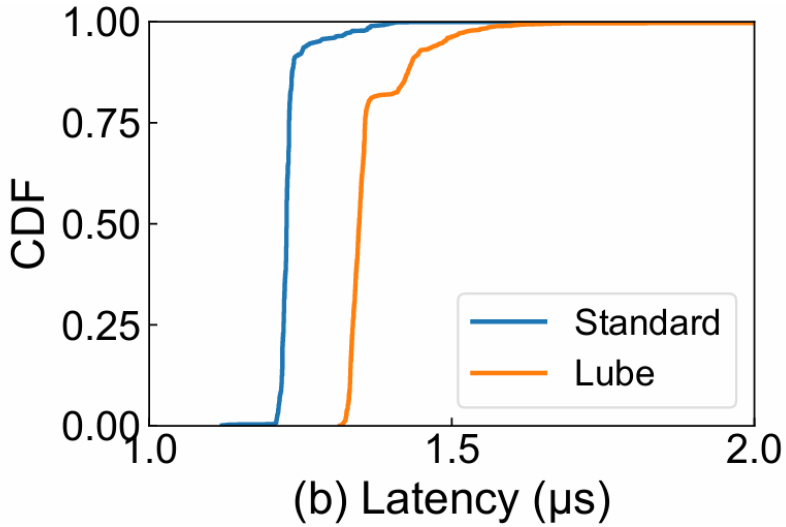
1. Pullup the default QP;
2. Switch traffic back to default QP at intervals for probing.

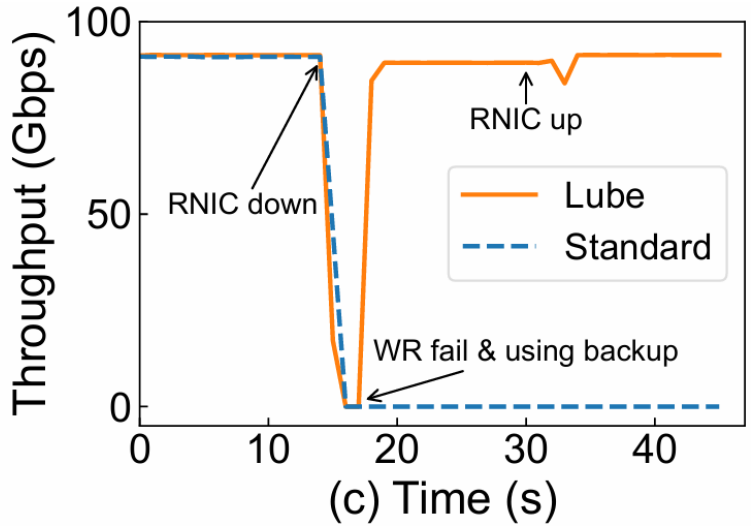
3. Recovery Process

1. Post WRs to ALL QPs;
2. Buffer outstanding WRs

LubeRDMA: Evaluation

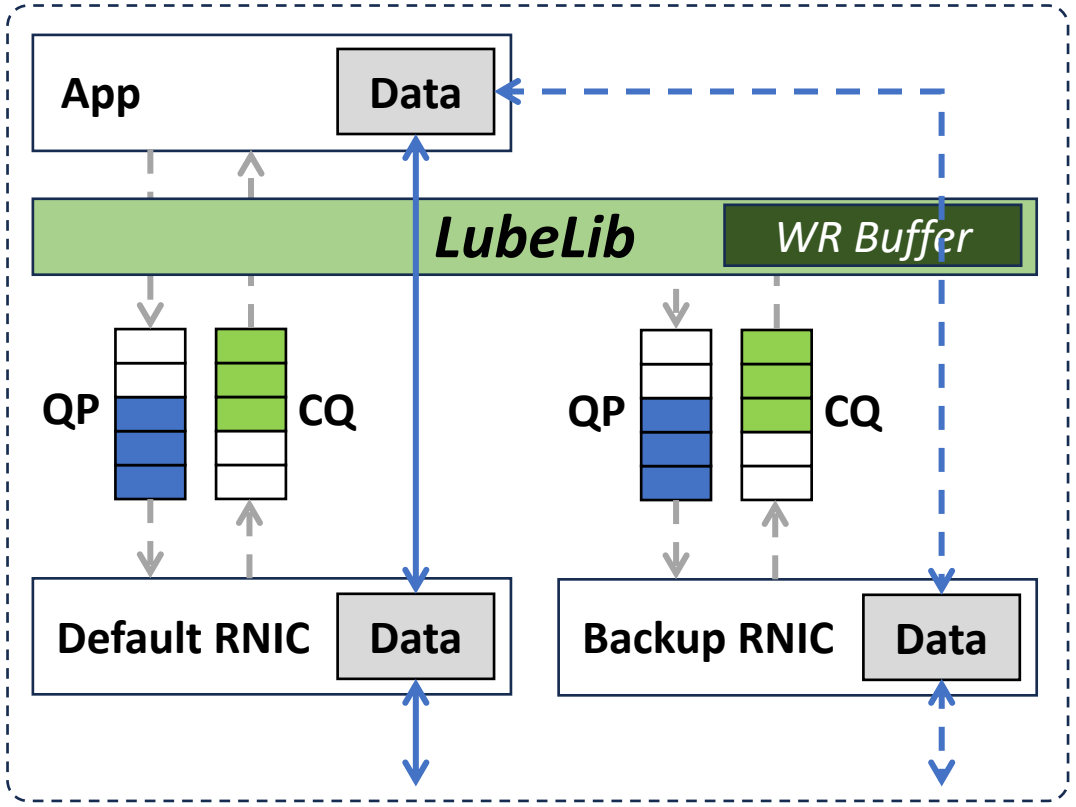






- Initialization time increases with more backup RNICs.
- ib_write_lat slightly increases.
- Effectively handle RDMA failures and prevent them from disturbing applications.

LubeRDMA:
A Fail-safe Mechanism of RDMA



Intuition

Take multiple RNICs as *backup* of each other.

Handle RDMA failures in *RDMA layer*.

Contribution

1. **Application-agnostic** backup RNIC management & failure handling;
2. **Low-overhead & high-performance** failure handling.

Contact us:

Shengkai Lin **jefflin@sjtu.edu.cn**

Shizhen Zhao shizhenzhao@sjtu.edu.cn

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY