One-Sided RDMA-Conscious Extendible Hashing for Disaggregated Memory

Pengfei Zuo, Jiazhao Sun, Liu Yang, Shuangwu Zhang, Yu Hua*

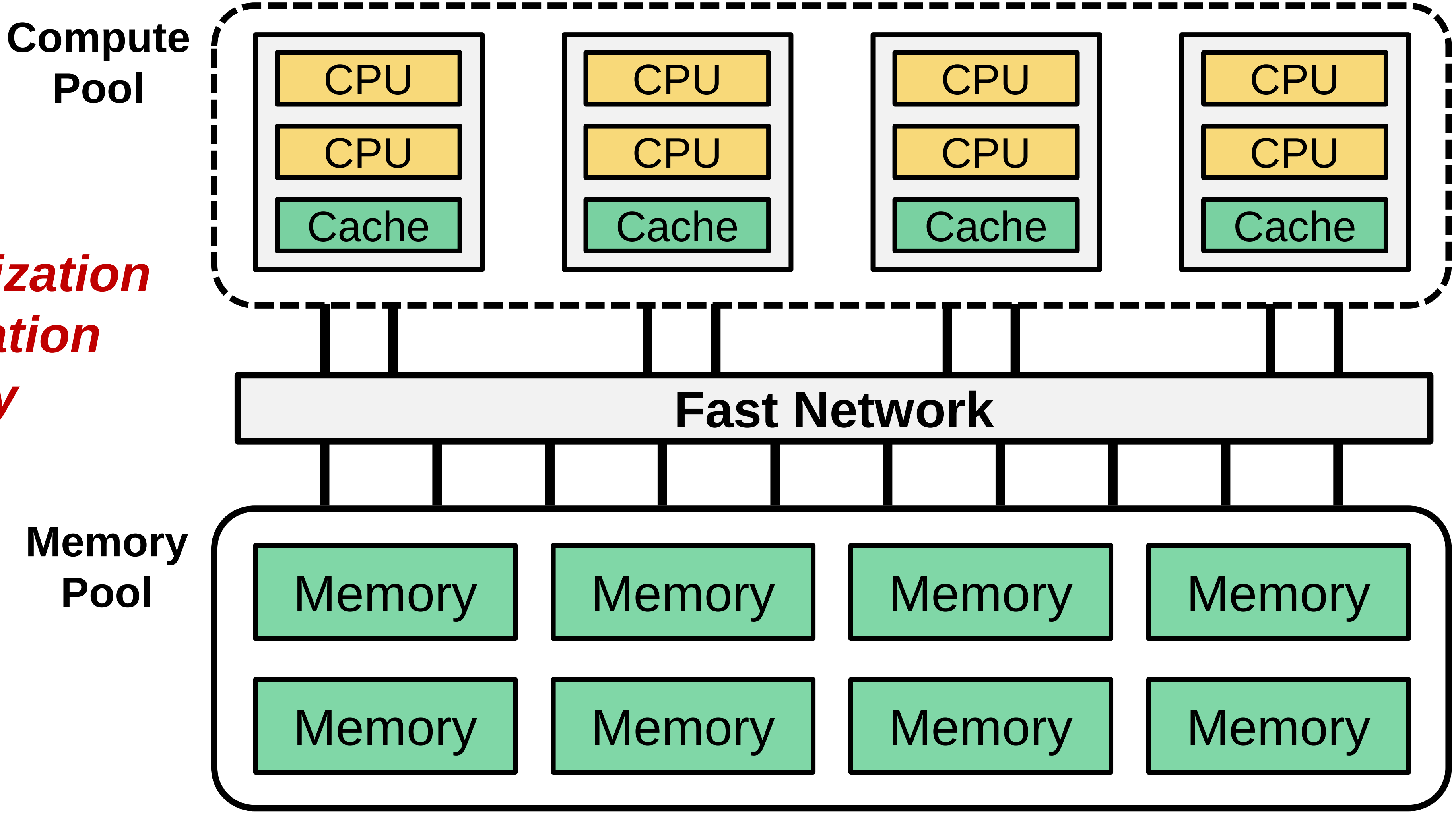
Huawei Cloud

**Huazhong University of Science and Technology*

USENIX Annual Technical Conference (ATC), 2021

Disaggregated Memory

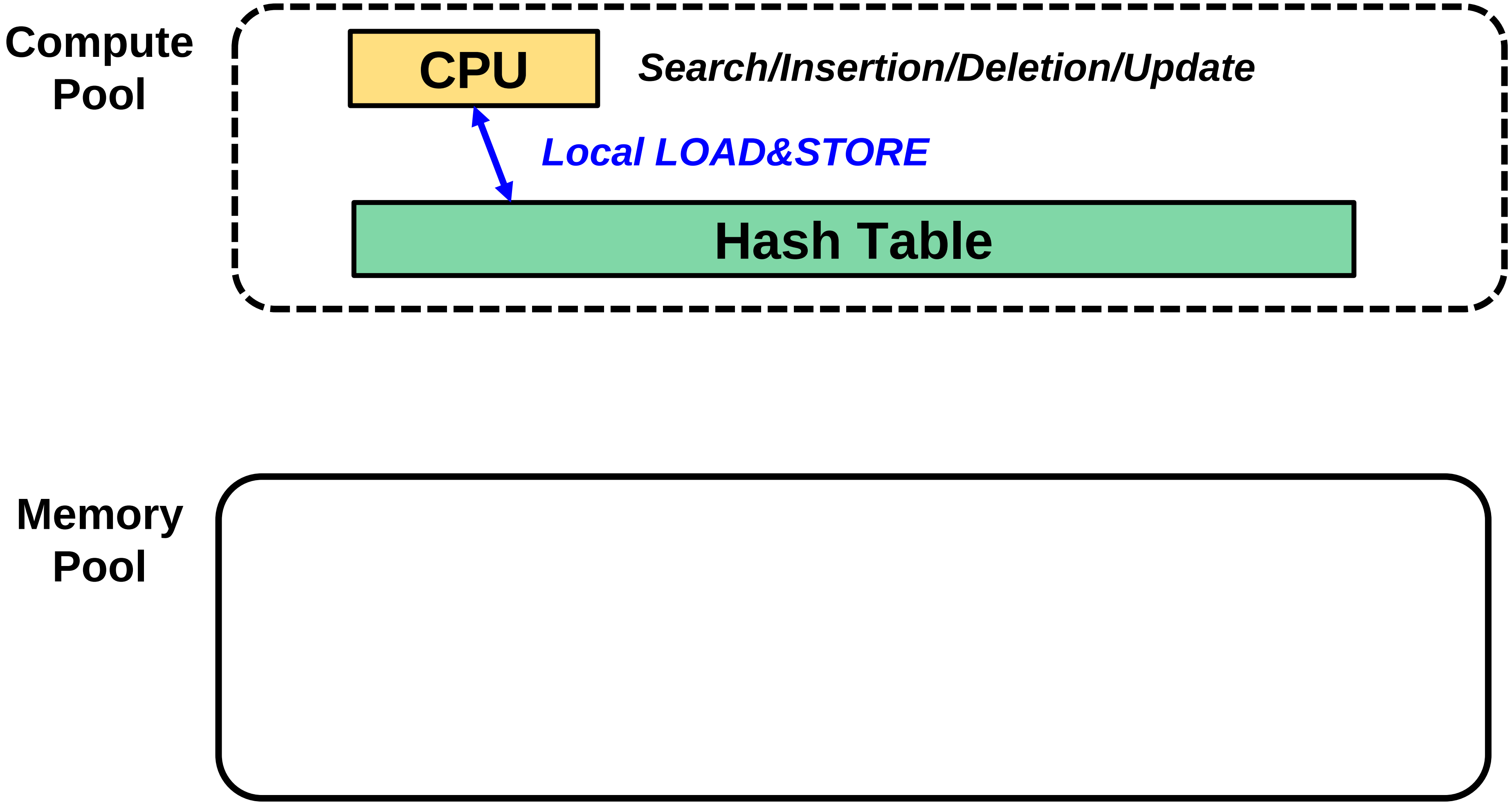
Resource Utilization
Failure Isolation
Elasticity



HUAWEI

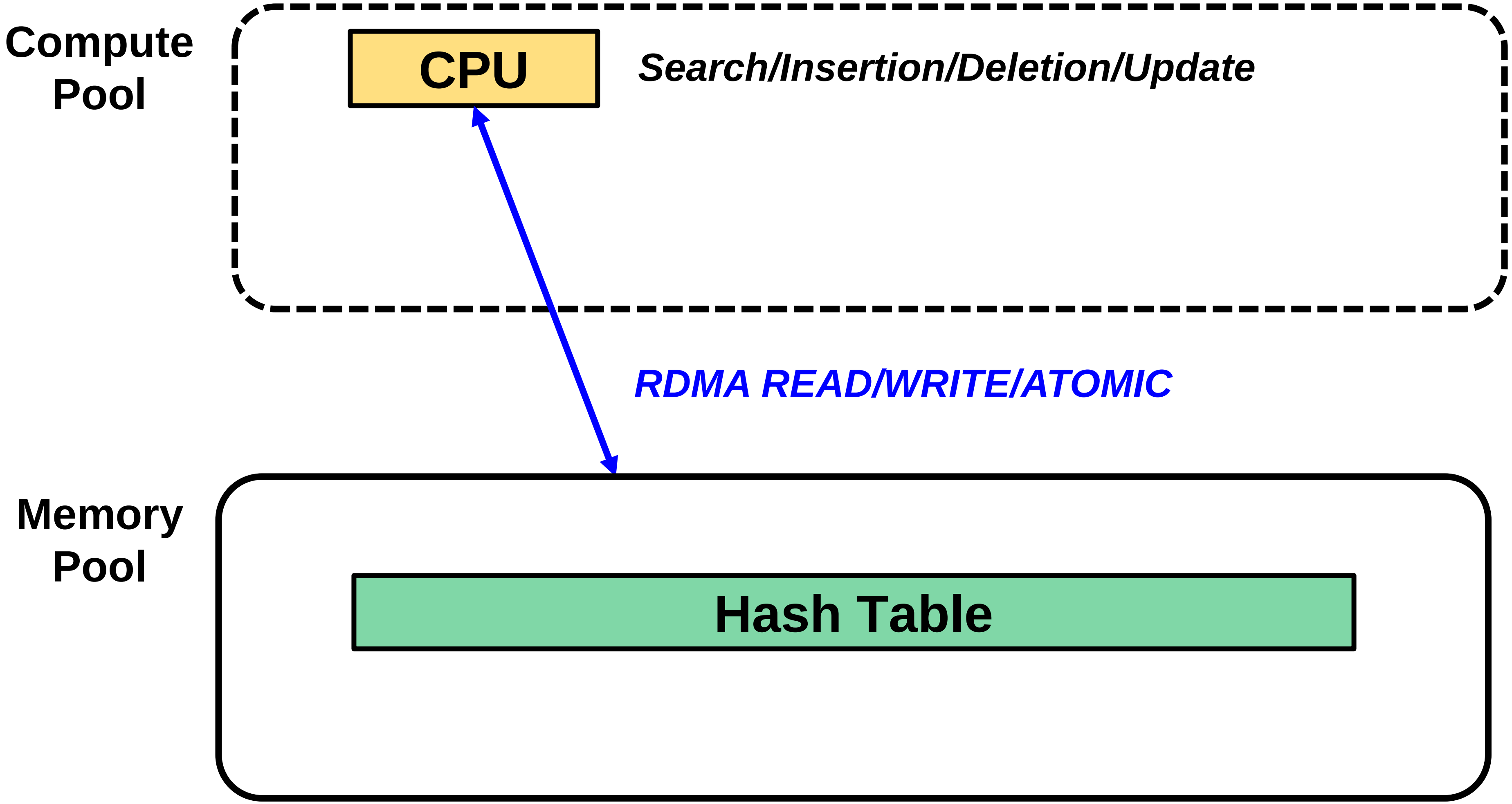
Hashing Indexes in Local Memory

Hashing Indexes in Disaggregated Memory

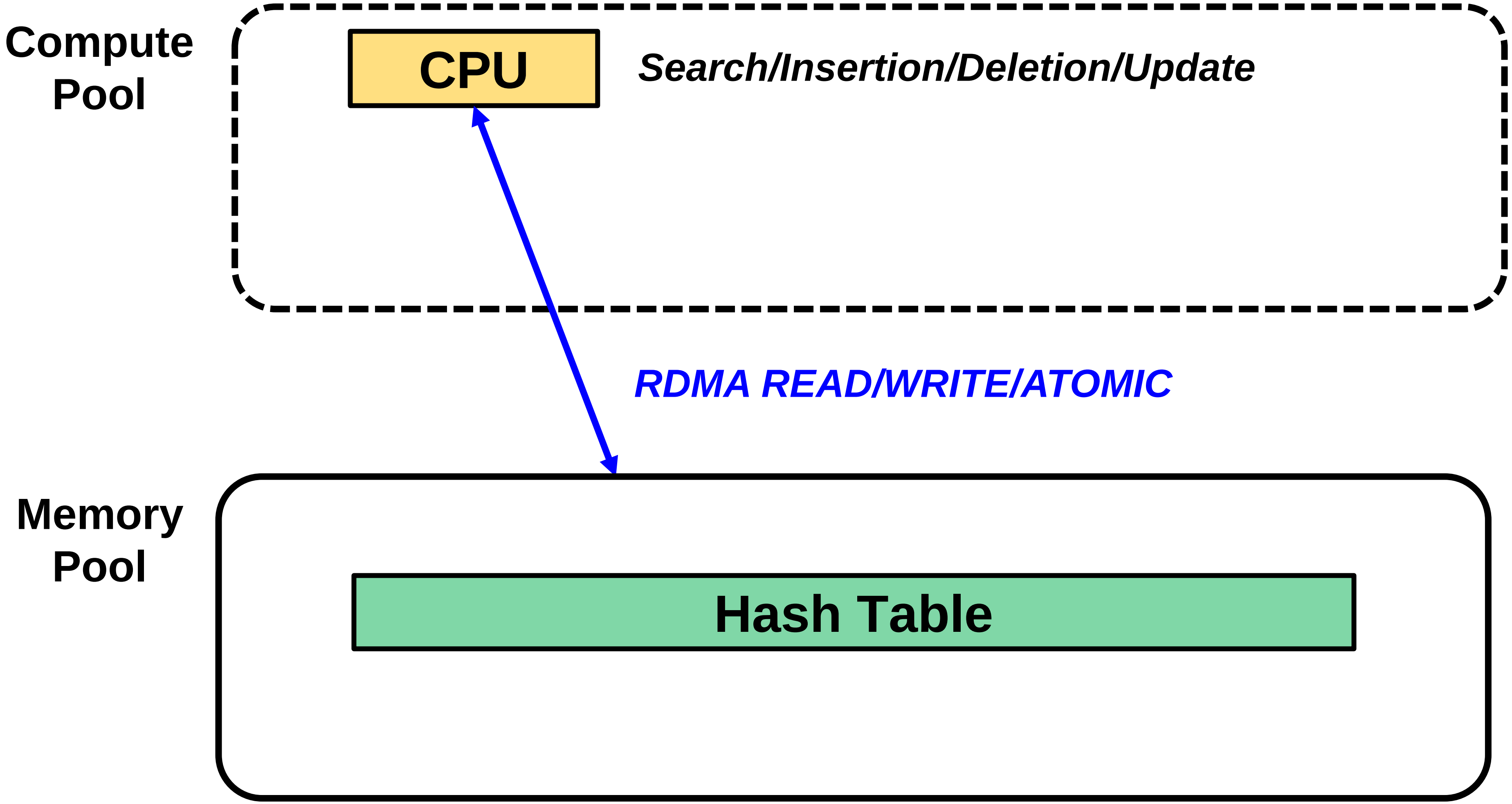


HUAWEI

Hashing Indexes in Disaggregated Memory

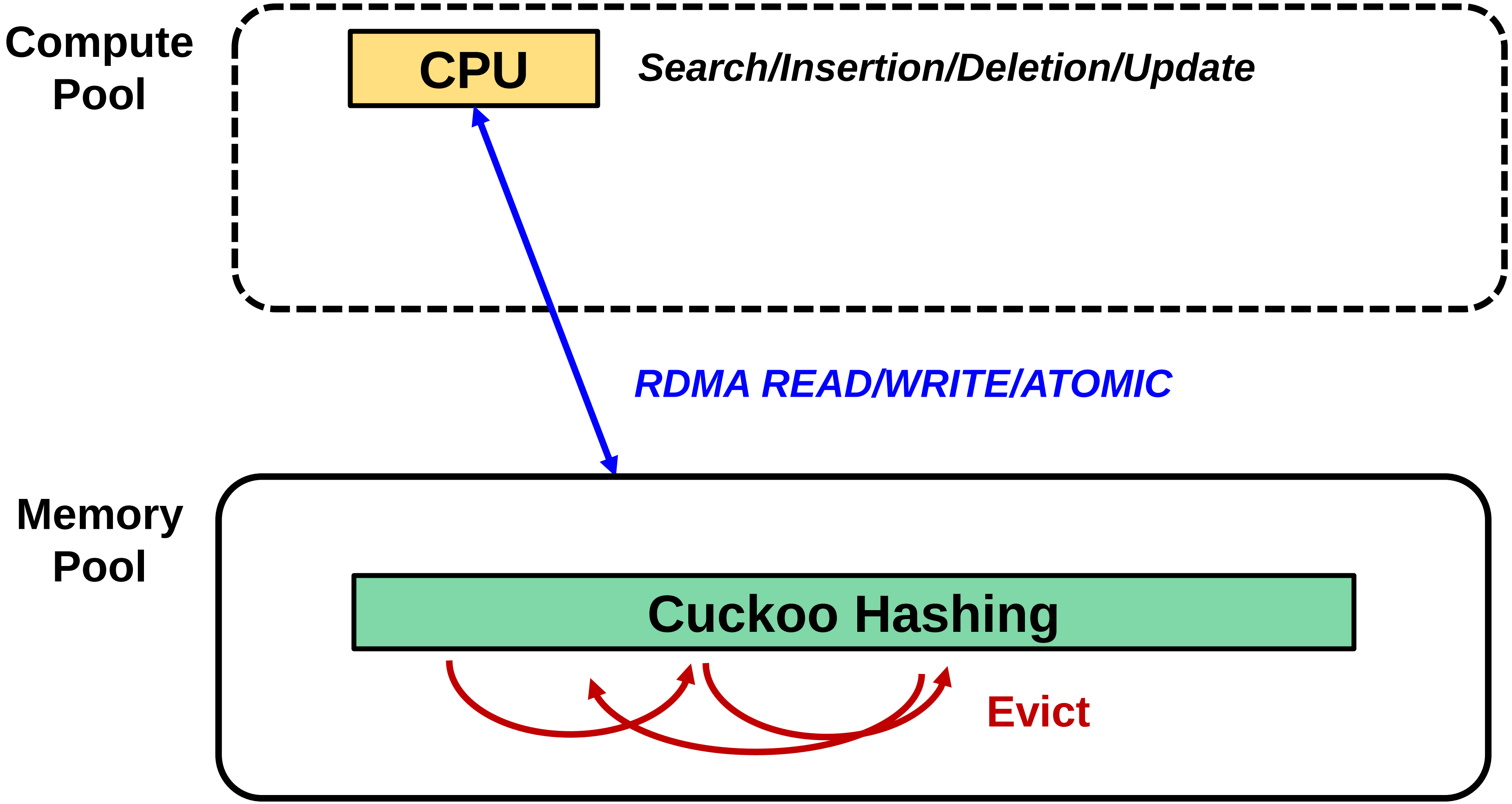


Challenge 1: Hash Collision



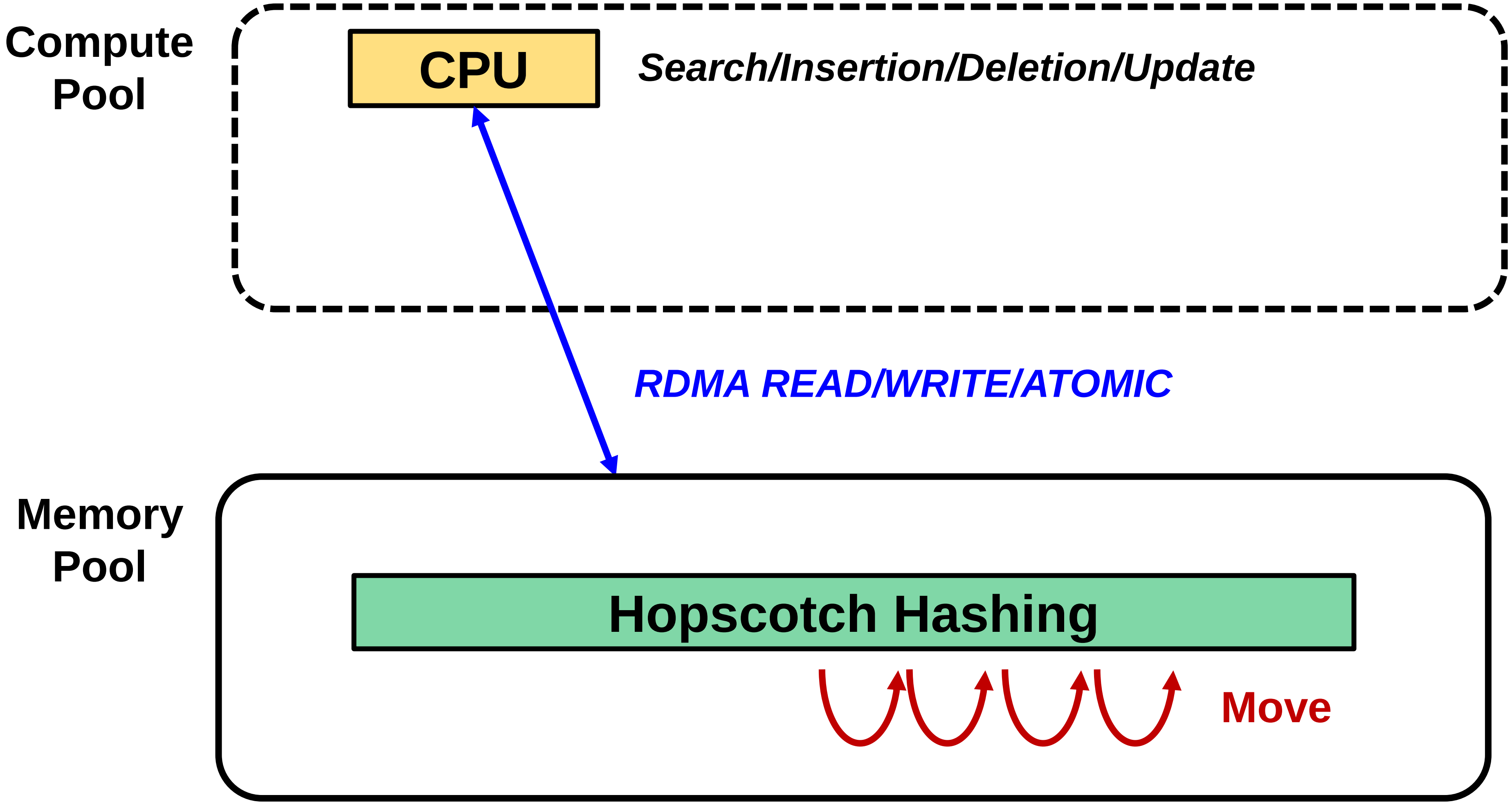
Challenge 1: Many remote reads&writes for handling hash collisions

Challenge 1: Hash Collision



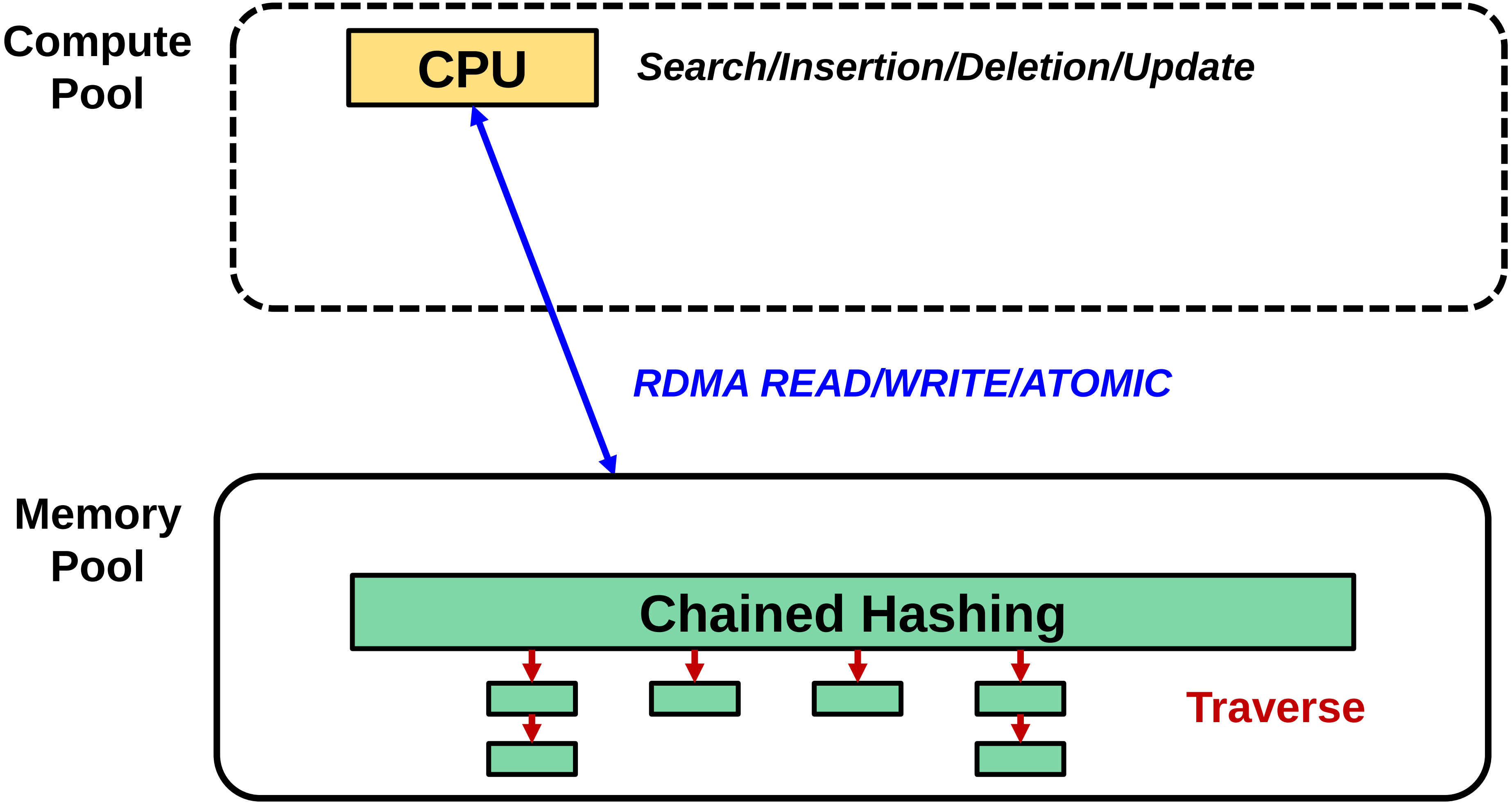
Challenge 1: Many remote reads&writes for handling hash collisions

Challenge 1: Hash Collision



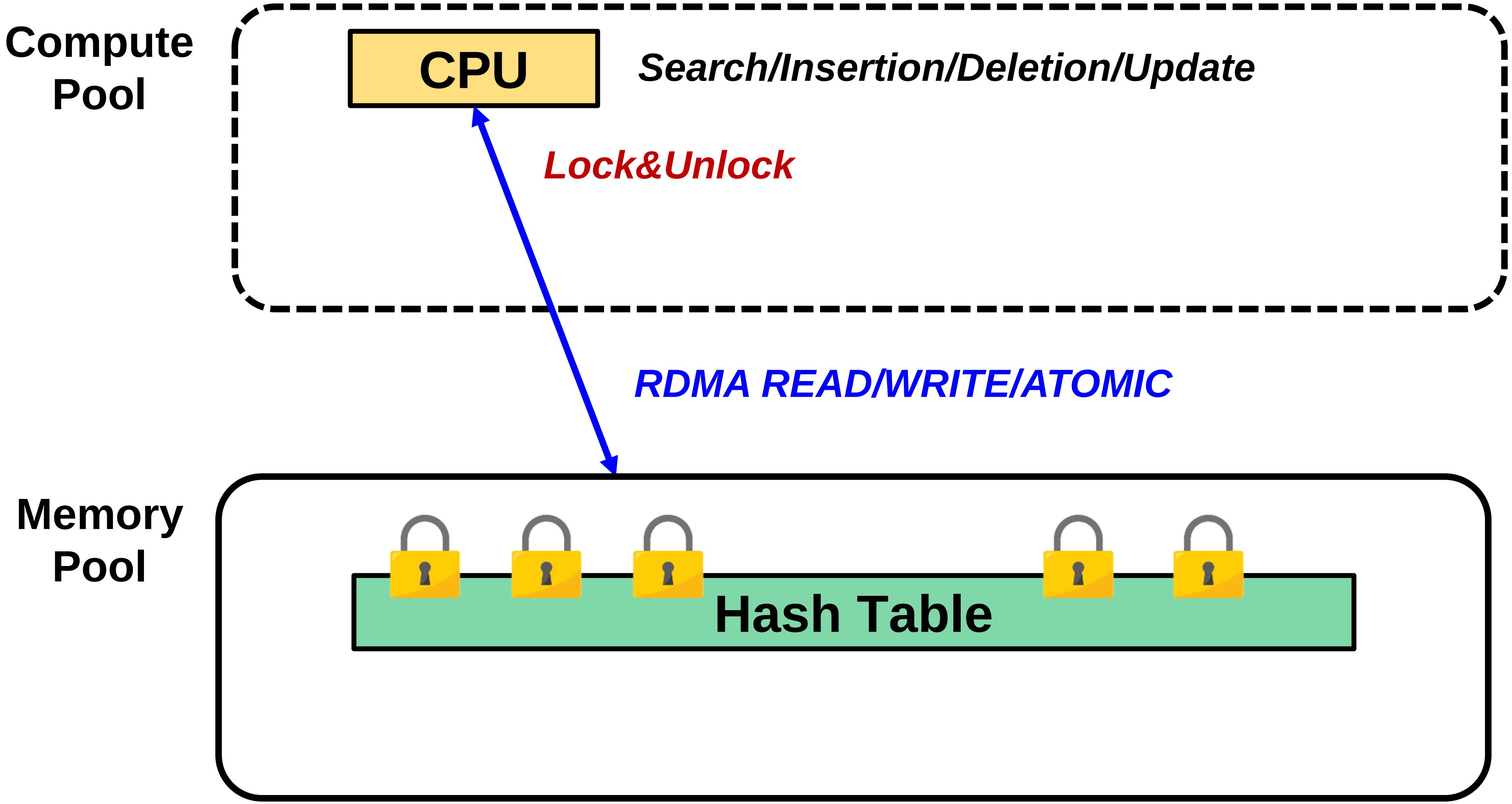
Challenge 1: Many remote reads&writes for handling hash collisions

Challenge 1: Hash Collision



Challenge 1: Many remote reads&writes for handling hash collisions

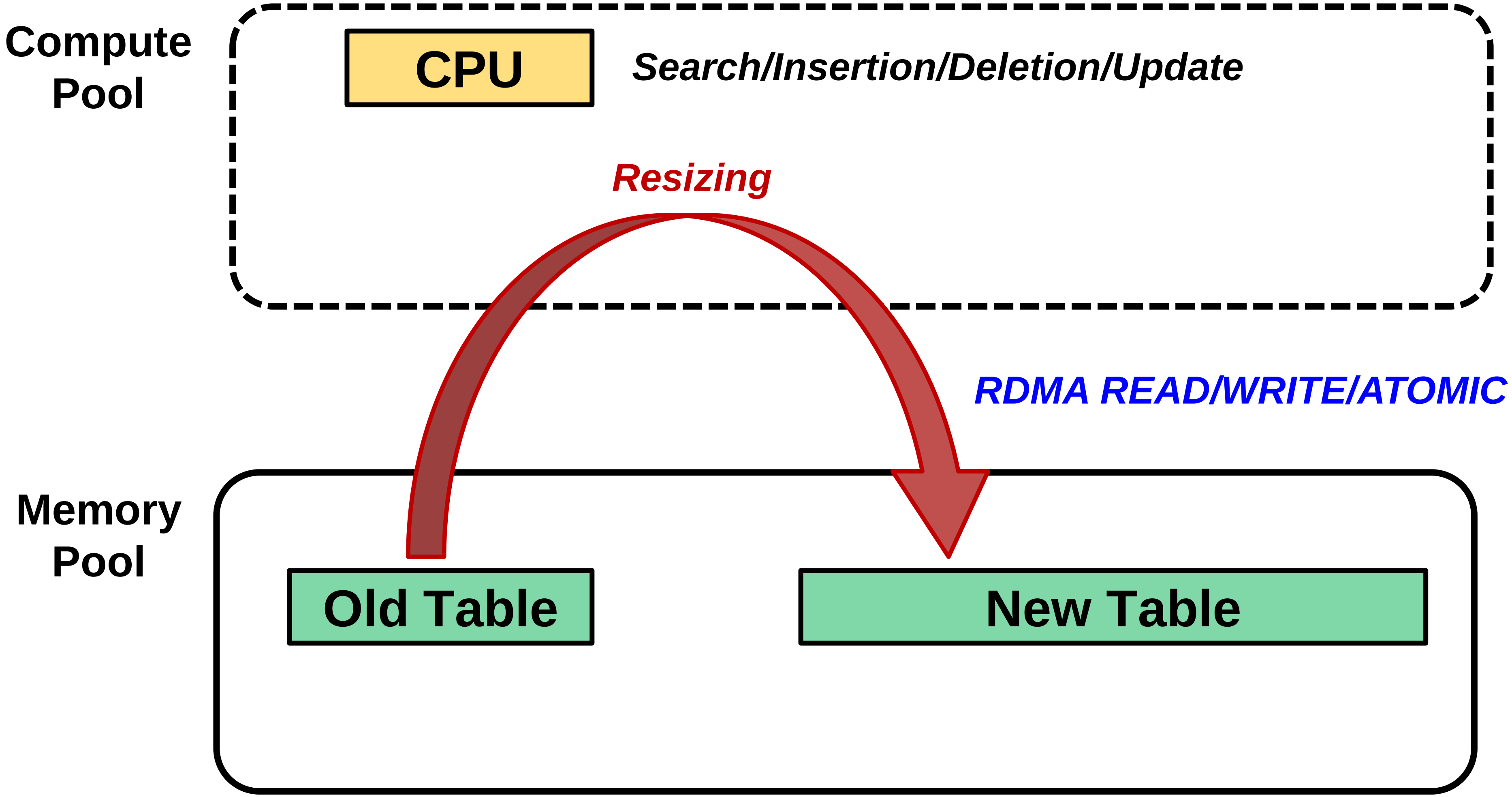
Challenge 2: Concurrency Control



Challenge 2: Locking has high overhead

HUAWEI

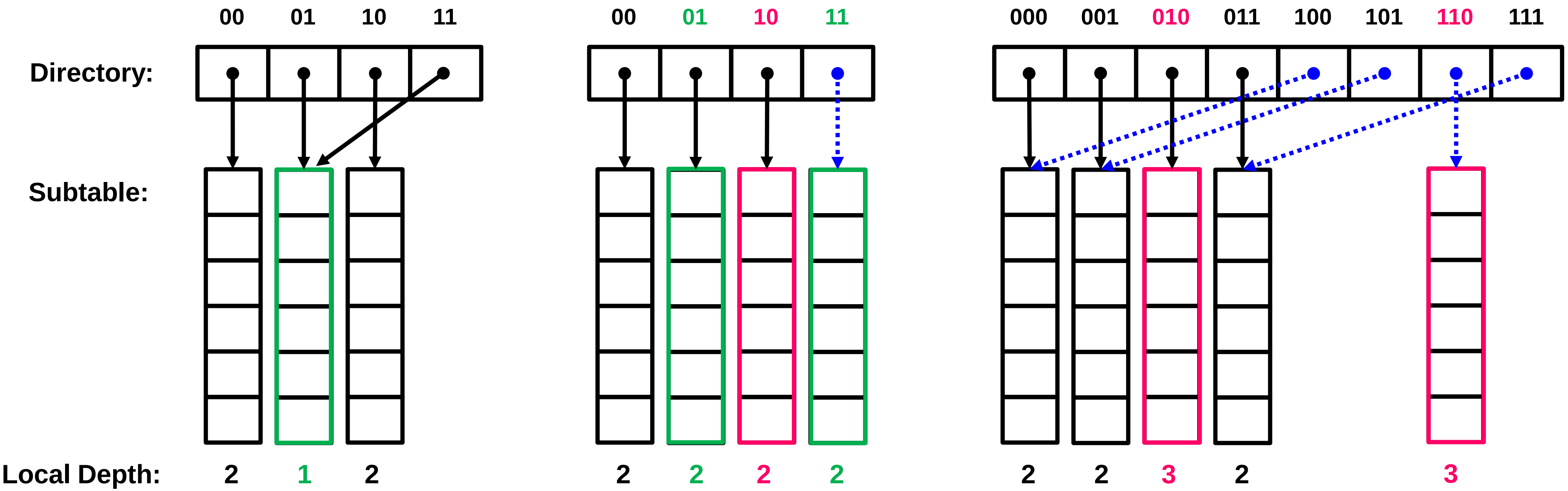
Challenge 3: Remote Resizing



Challenge 3: Moving items from the old table to the new table

HUAWEI

Challenge 3: Extendible Resizing



(a) A hash table (global depth = 2)

(b) Subtable resizing (global depth = 2)

(c) Directory resizing (global depth = 3)

- Challenge of using extendible resizing in disaggregated memory
  - One extra RDMA for accessing the directory

Challenge Summary

1. **Many remote reads&writes for handling hash collisions**
   - Cuckoo hashing, hopscotch hashing, chained hashing

2. **Concurrency control for remote access**
   - One RDMA RTT for locking or unlocking

3. **Tricky remote resizing of hash tables**
   - One extra RDMA RTT for accessing the directory

RDMA-Conscious Extendible (RACE) Hashing

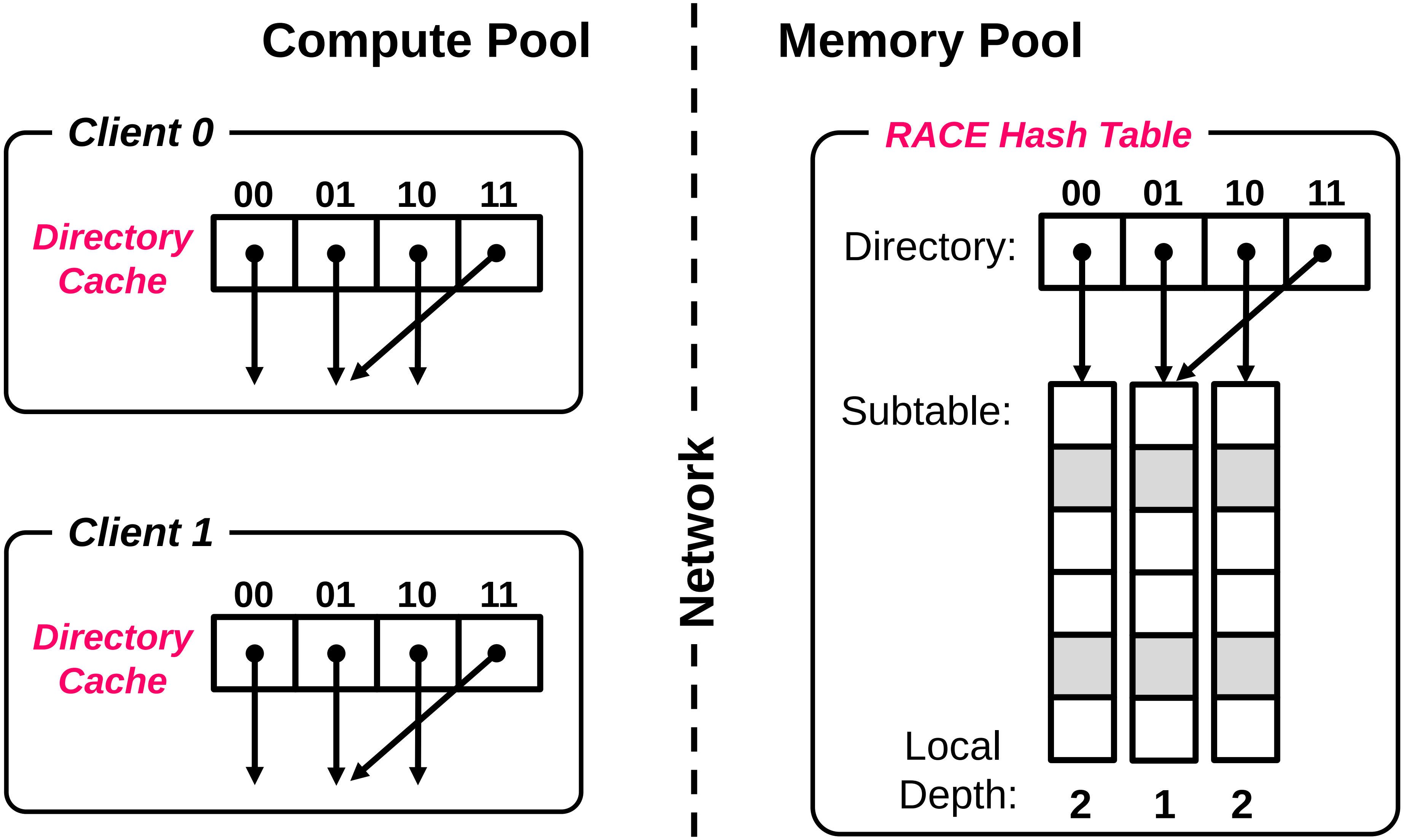
1. **Many remote reads&writes for handling hash collisions**
   - Cuckoo hashing, hopscotch hashing, chained hashing
   - ***Solution**: One-sided RDMA-conscious table structure*

2. **Concurrency control for remote access**
   - One RDMA RTT for locking or unlocking
   - ***Solution**: Lock-free remote concurrency control*

3. **Tricky remote resizing of hash tables**
   - One extra RDMA RTT for accessing the directory
   - ***Solution**: Extendible remote resizing with stale-read caching*

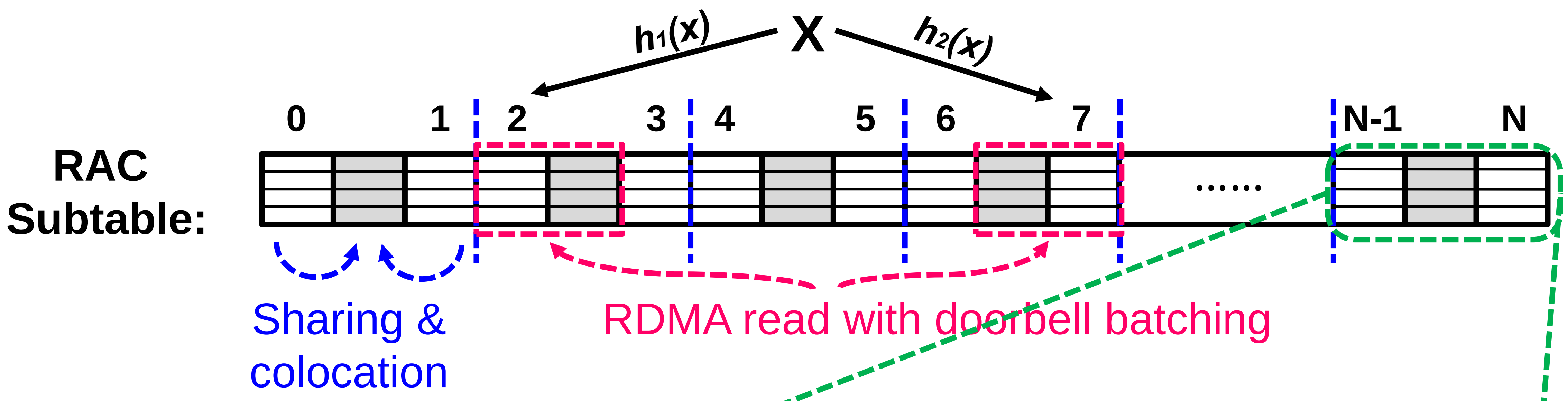
HUAWEI

Architectural Overview

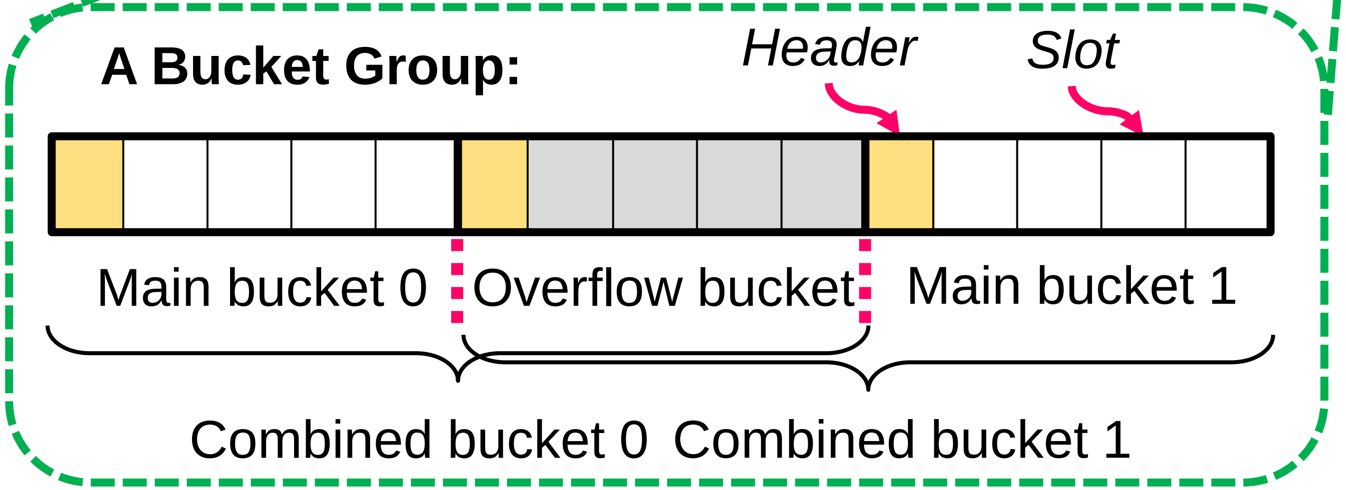


HUAWEI

One-sided RDMA-Conscious (RAC) Subtable

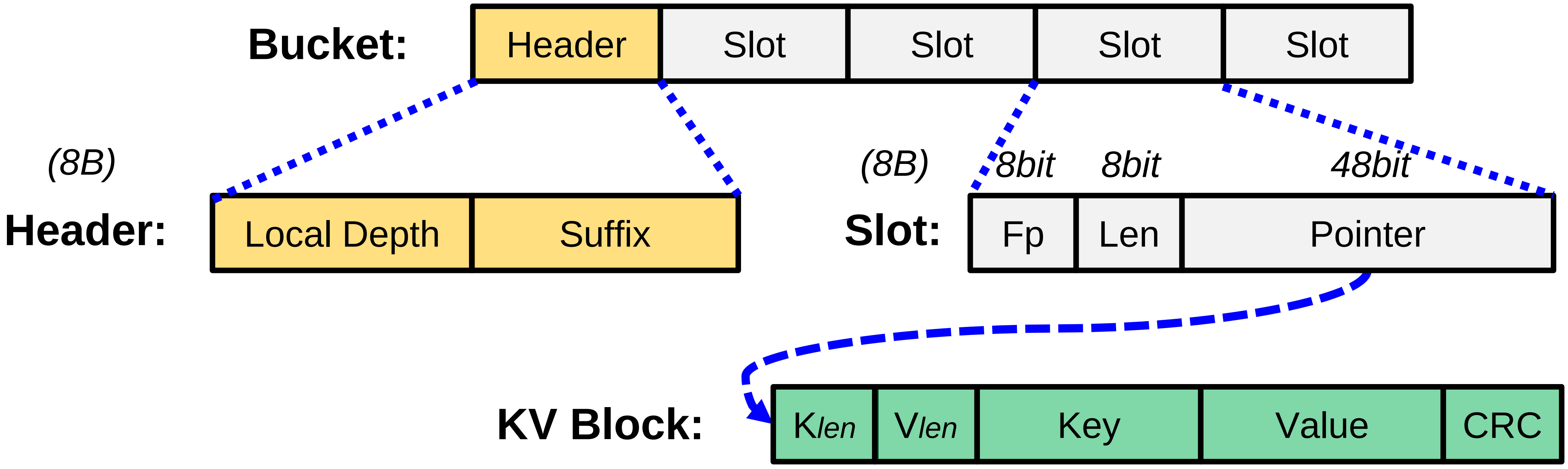


- **Design decisions**
  - **Associativity**
  - **Two choices**
  - **Overflow colocation**
- **Strengths**
  - **RDMA-search friendly**
  - **RDMA-IDU friendly**
  - **High memory efficiency**



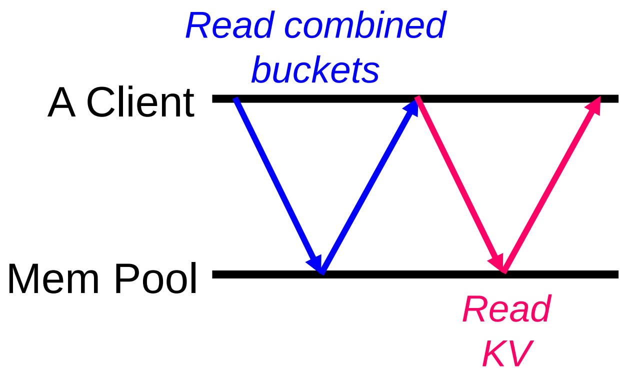
Lock-free Remote Concurrency Control

- Bucket Structure: supporting the RDMA CAS

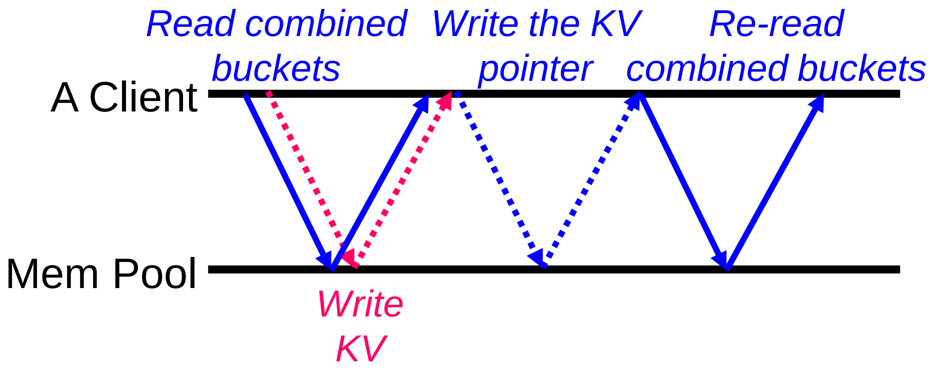


HUAWEI

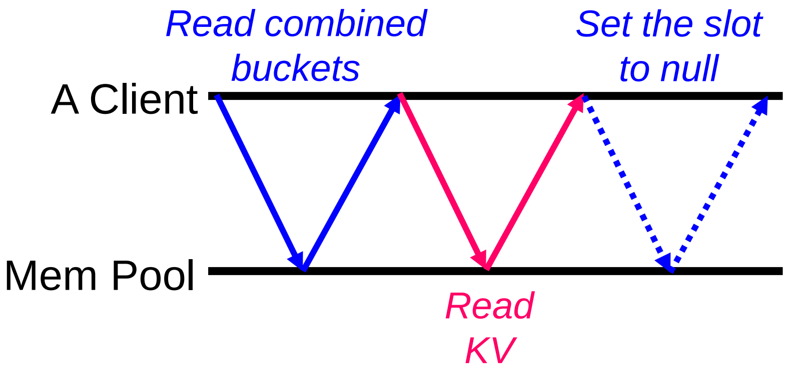
Lock-free Remote Concurrency Control



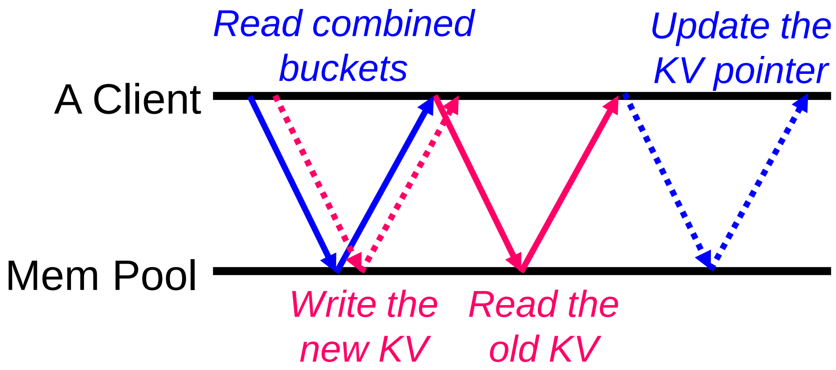
(a) Search



(b) Insertion

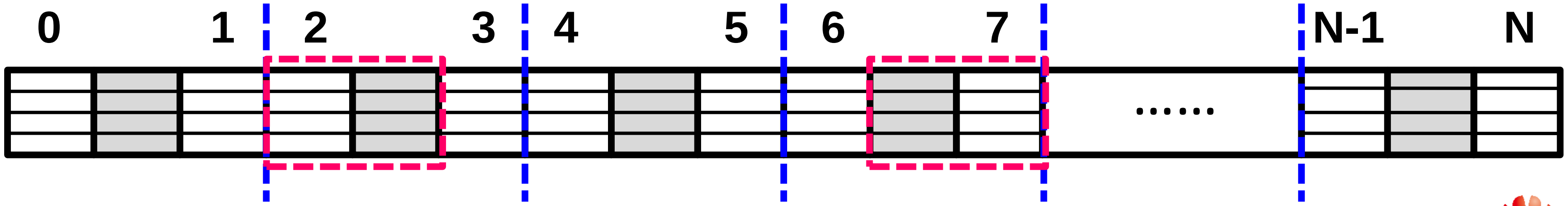


(c) Deletion



(d) Update

RAC Subtable:

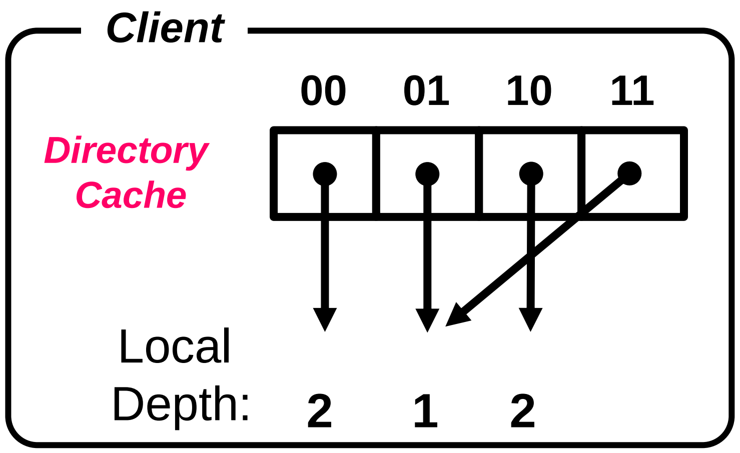


HUAWEI

Extendible Remote Resizing

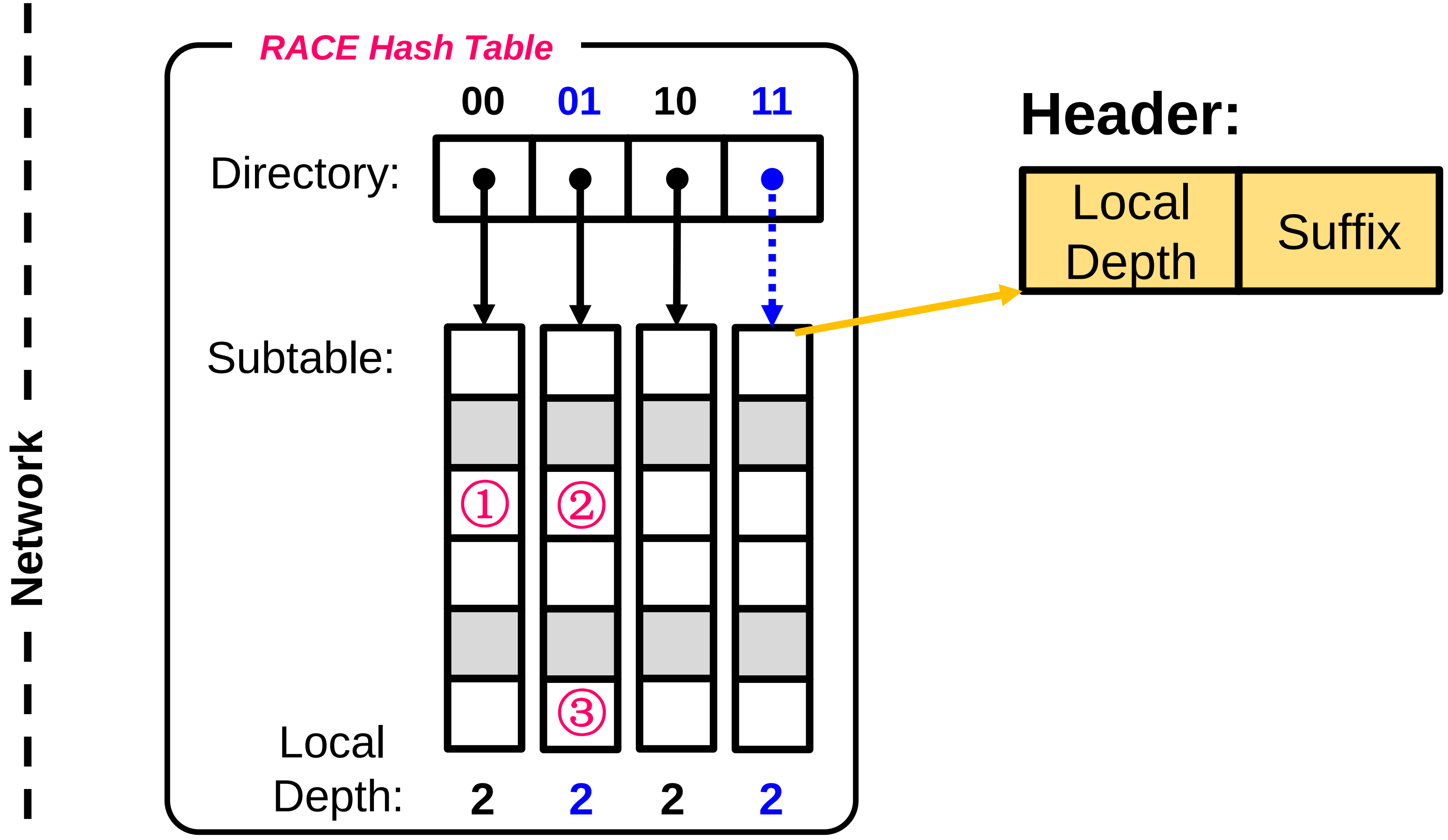
- Client Directory Cache with Stale Reads

Compute Pool | **Memory Pool**



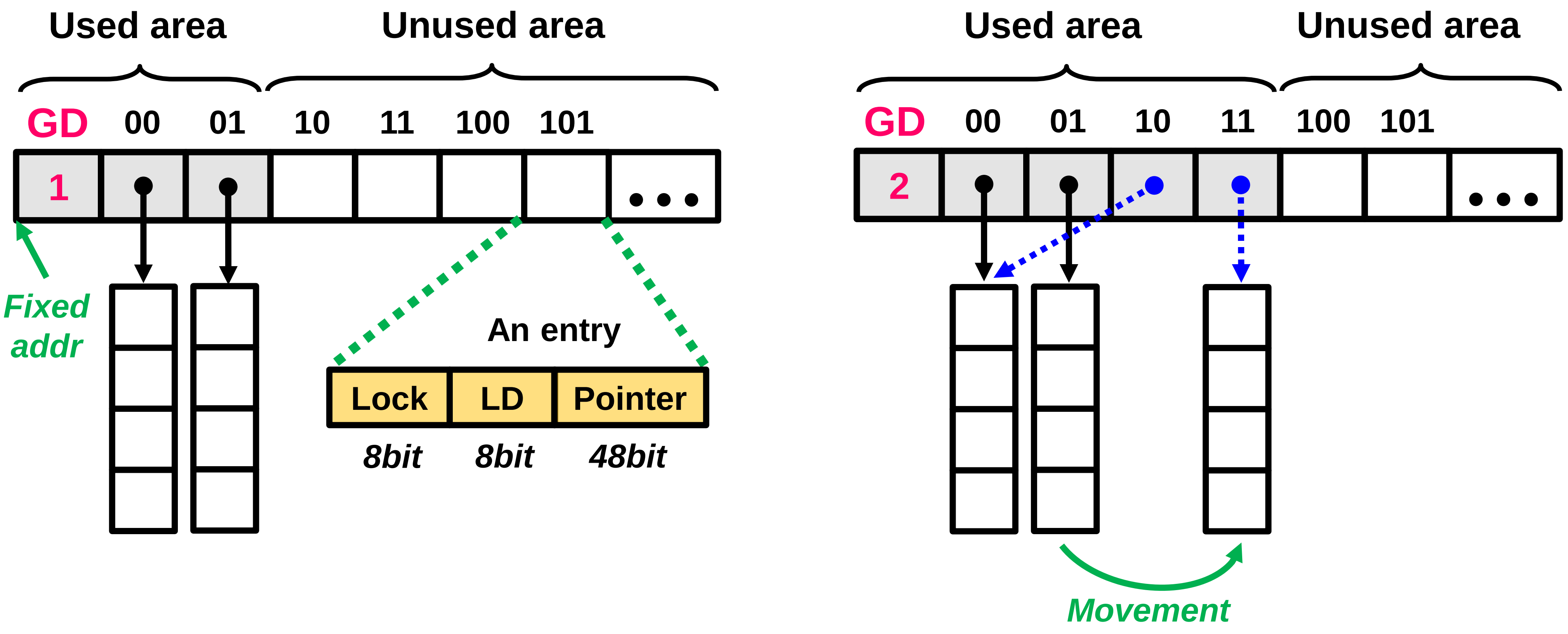
Different cases of stale reads:

	Key	Bucket		Correct?
①	XX00	2	00	Yes
②	XX01	2	01	Yes
③	XX11	2	01	No



Extendible Remote Resizing

- Resize a directory



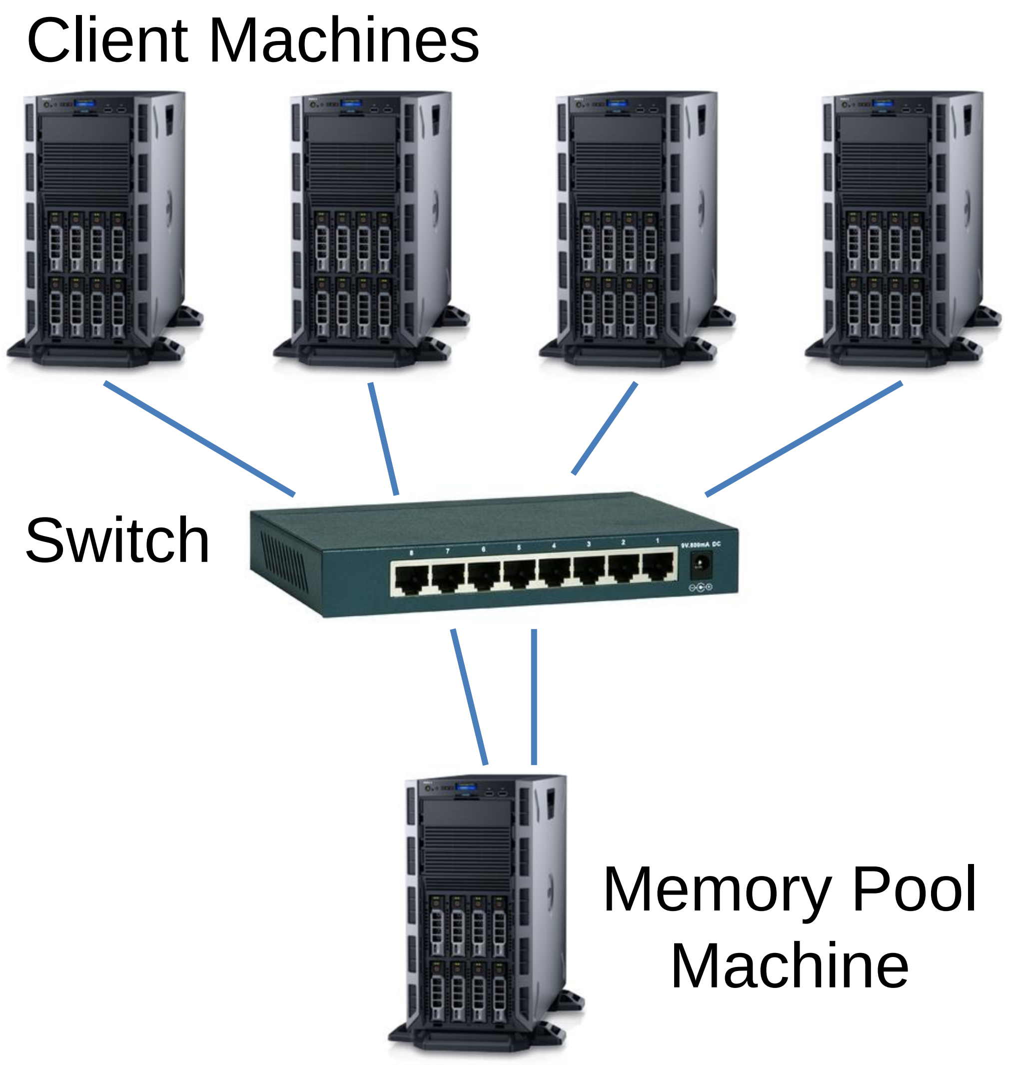
(a) Before the directory resizing

(b) After the directory resizing

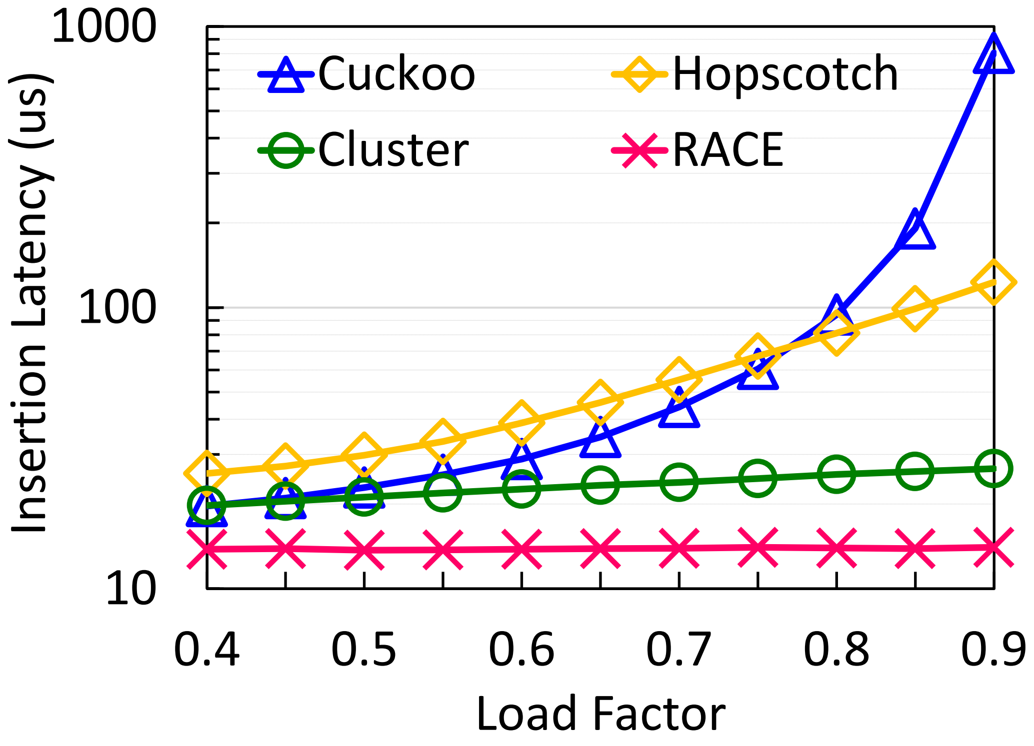
HUAWEI

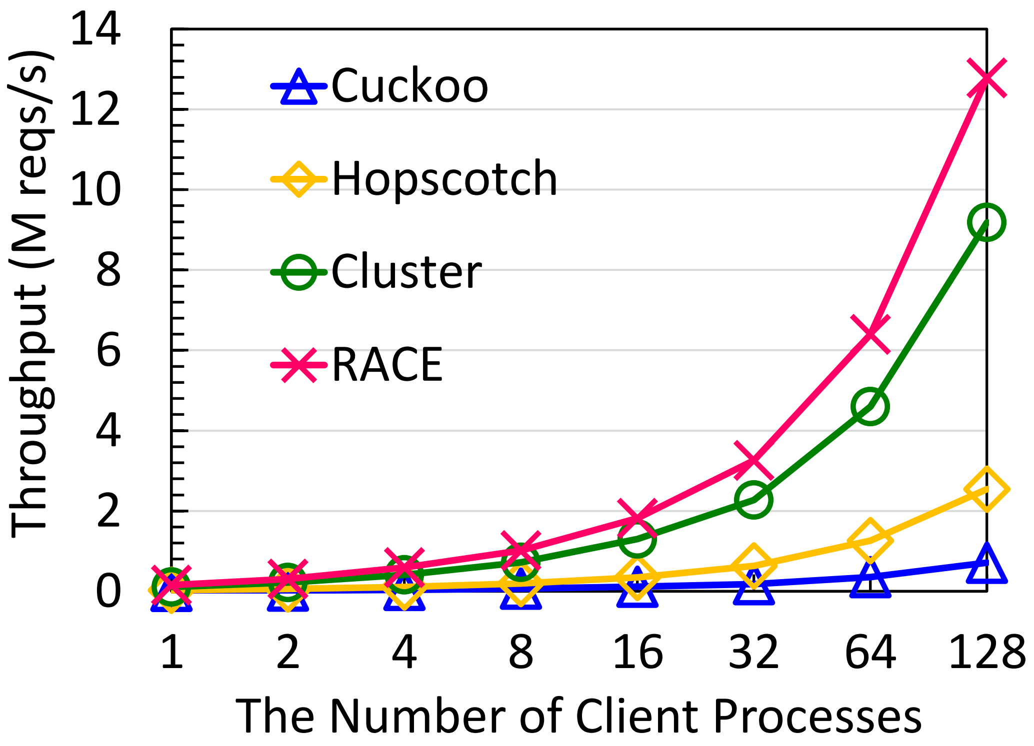
Experimental Setup

- **Testbed**
  - 4 client machines + 1 memory machine
- **Workloads**
  - YCSB, 16B key + 32B value
- **Comparisons**
  - Pilaf cuckoo hashing [ATC'15]
  - FaRM hopscotch hashing [NSDI'14]
  - DrTM cluster hashing [SOSP'15]



Insertion

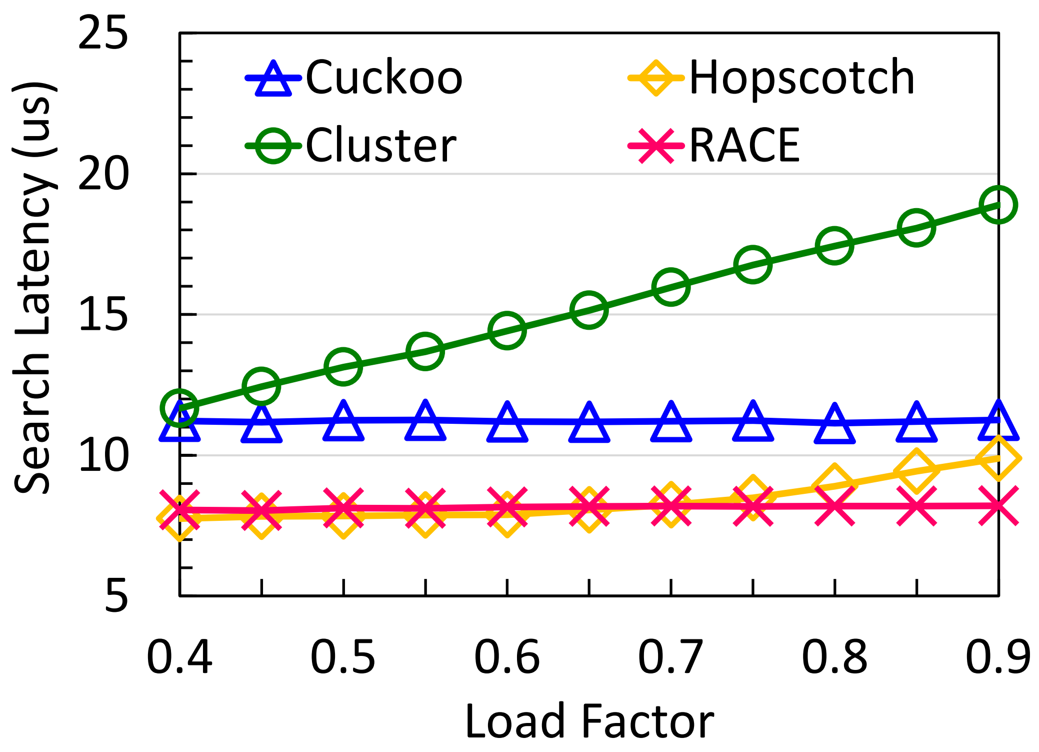




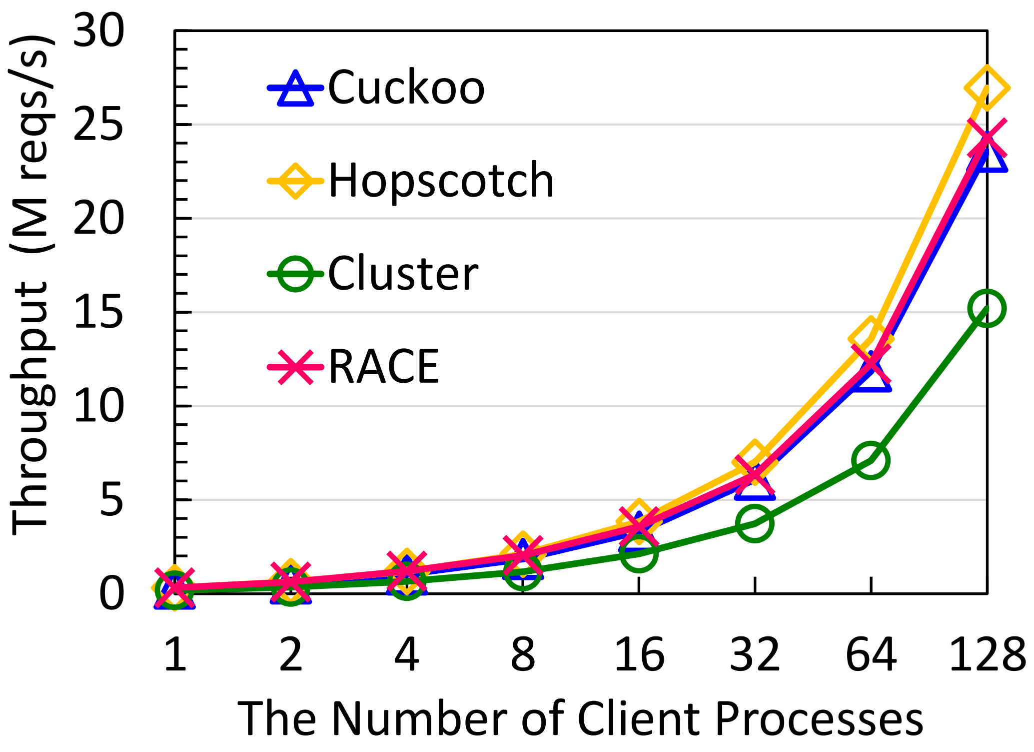
- RACE hashing improves the insertion throughput by 1.4~16.9×

HUAWEI

Search

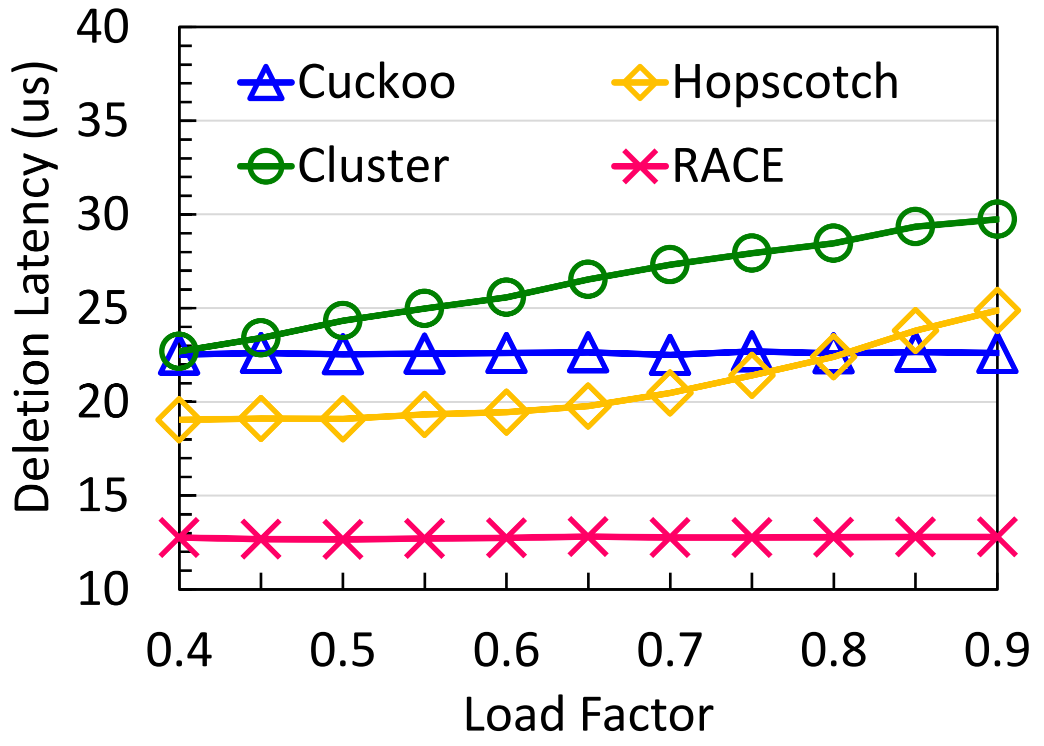


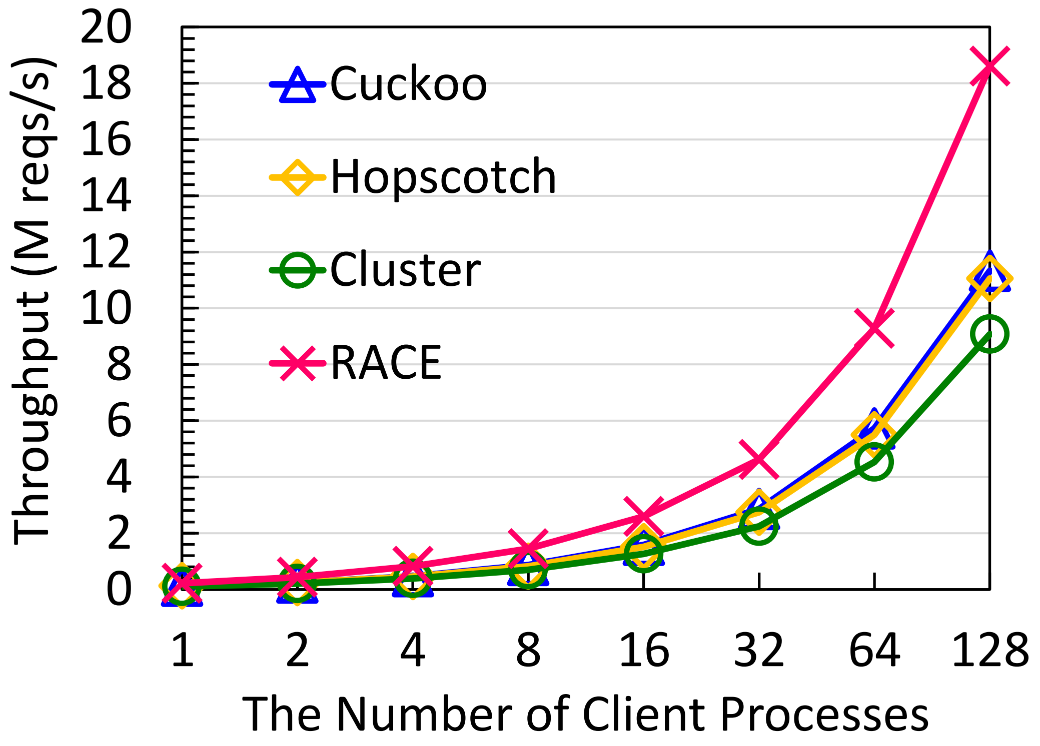
25
20
15
10
5
Search Latency (us)
Cuckoo
Hopscotch
Cluster
RACE
0.4
0.5
0.6
0.7
0.8
0.9
Load Factor



30
25
20
15
10
5
0
Throughput (M reqs/s)
Cuckoo
Hopscotch
Cluster
RACE
1
2
4
8
16
32
64
128
The Number of Client Processes

Deletion

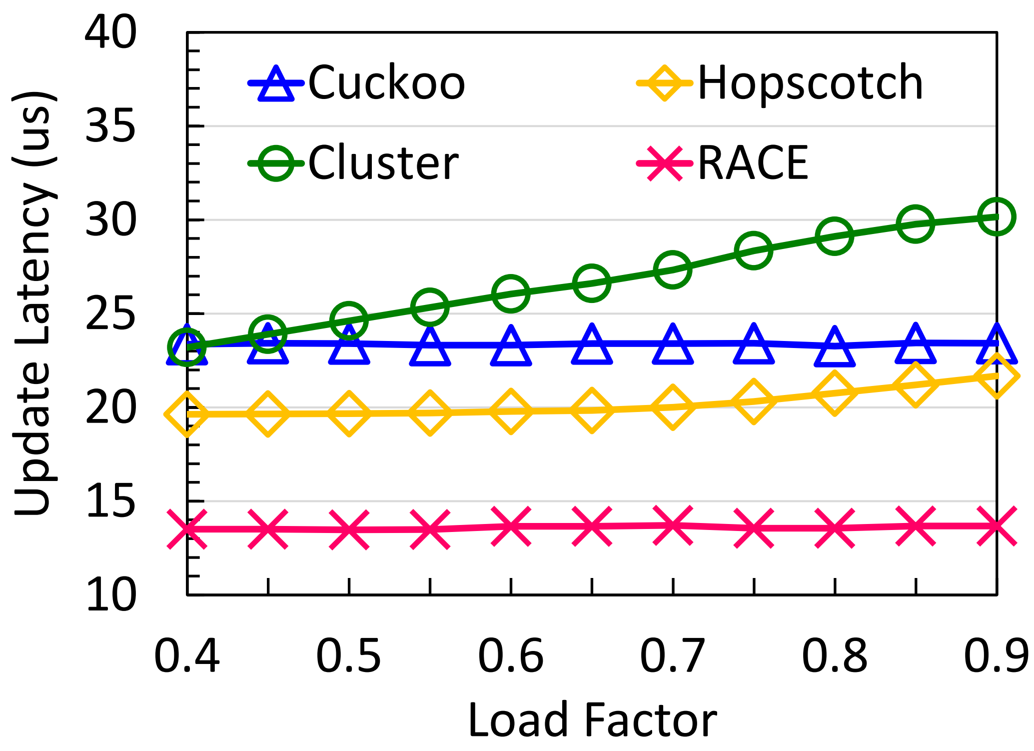


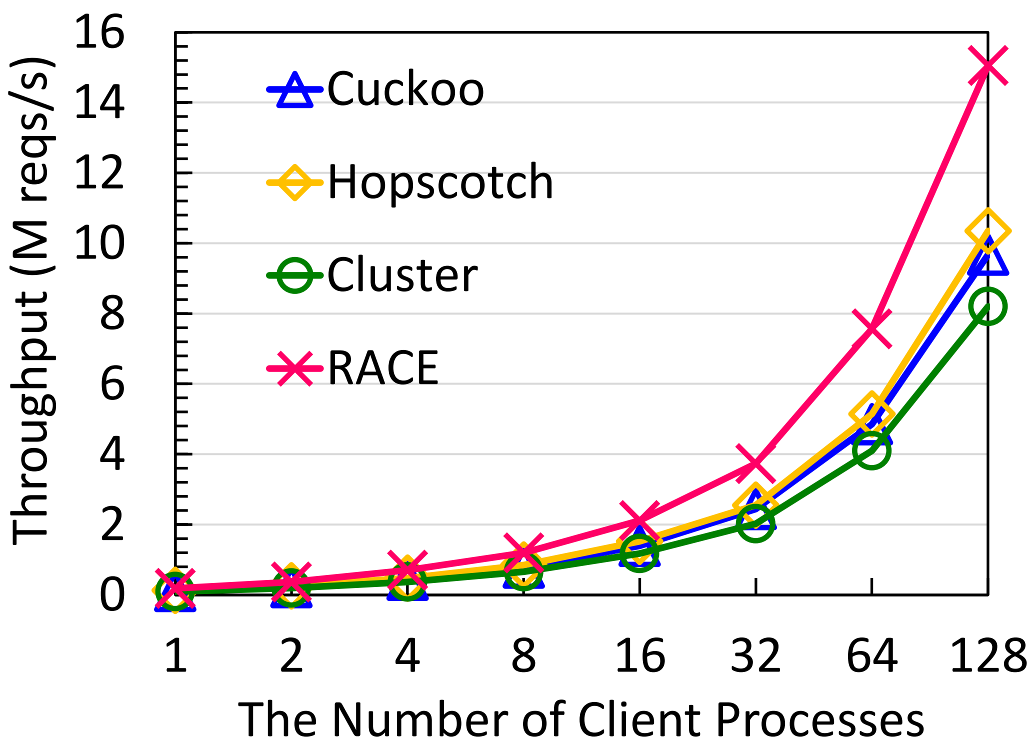


- RACE hashing improves the deletion throughput by 1.7~2.1×

HUAWEI

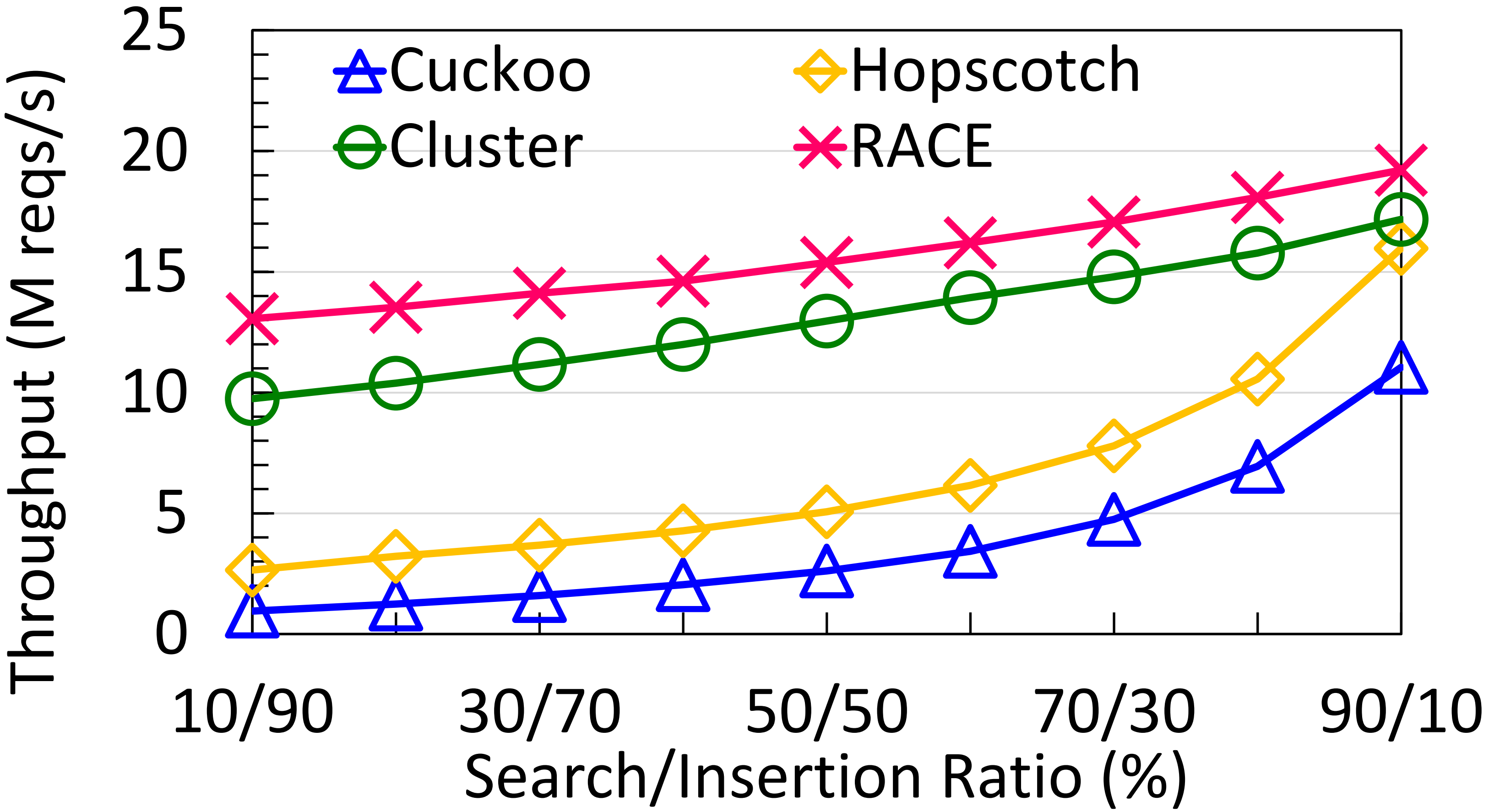
Update





- RACE hashing improves the deletion throughput by 1.5~1.9×

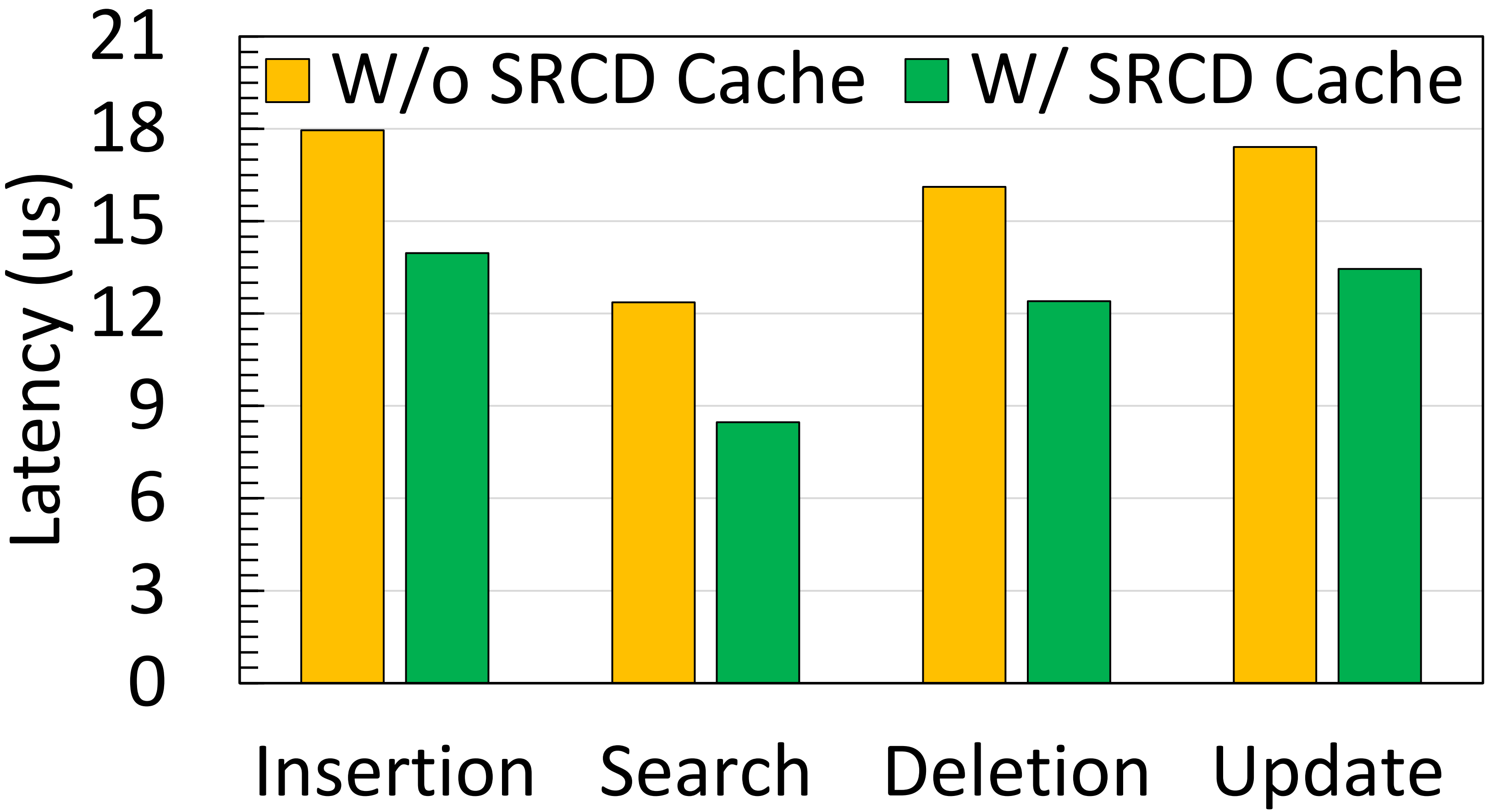
YCSB Hybrid Workloads



- RACE hashing speeds up the YCSB hybrid workloads by 1.4~1.3.7×

HUAWEI

The Stale-read Client Directory (SRCD) Cache



- The SRCD cache reduces the request latency by 23%~32%

HUAWEI

Conclusion

- Traditional distributed in-memory hashing indexes become inefficient in disaggregated memory
  - Many remote access, concurrency access, resizing
- We propose RACE hashing, the first hashing index designed for disaggregated memory
  - One-sided RDMA-conscious table structure
  - Lock-free remote concurrency control
  - Extendible remote resizing
- RACE Hashing outperforms state-of-the-art distributed in-memory hashing indexes by 1.4-13.7× in YCSB hybrid workloads

Thank you! Q&A