Building Efficient Distributed In-memory Computing Systems by *Combing Hardware Features*

RONG CHEN

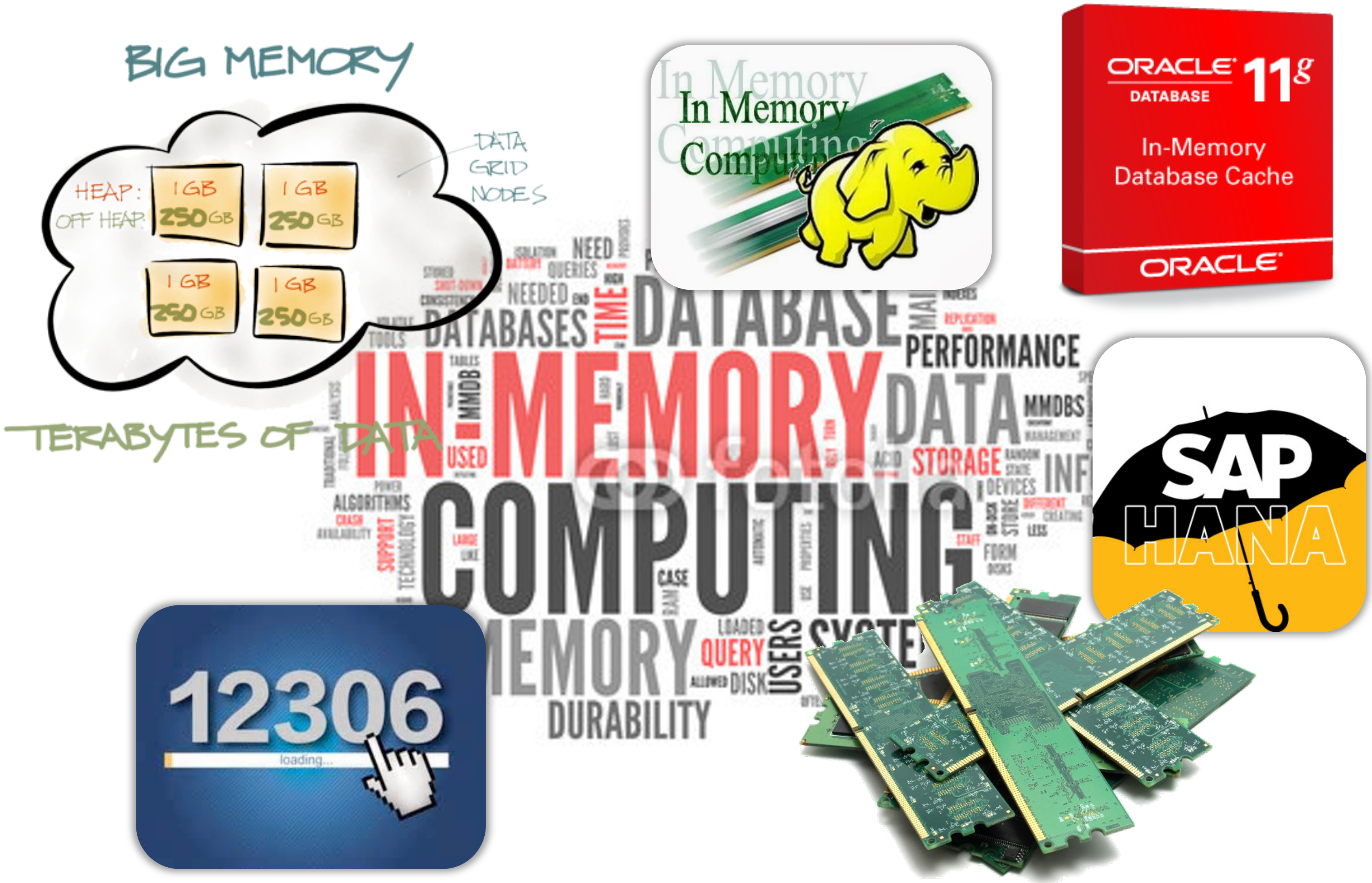
Institute of Parallel and Distributed Systems

Shanghai Jiao Tong University

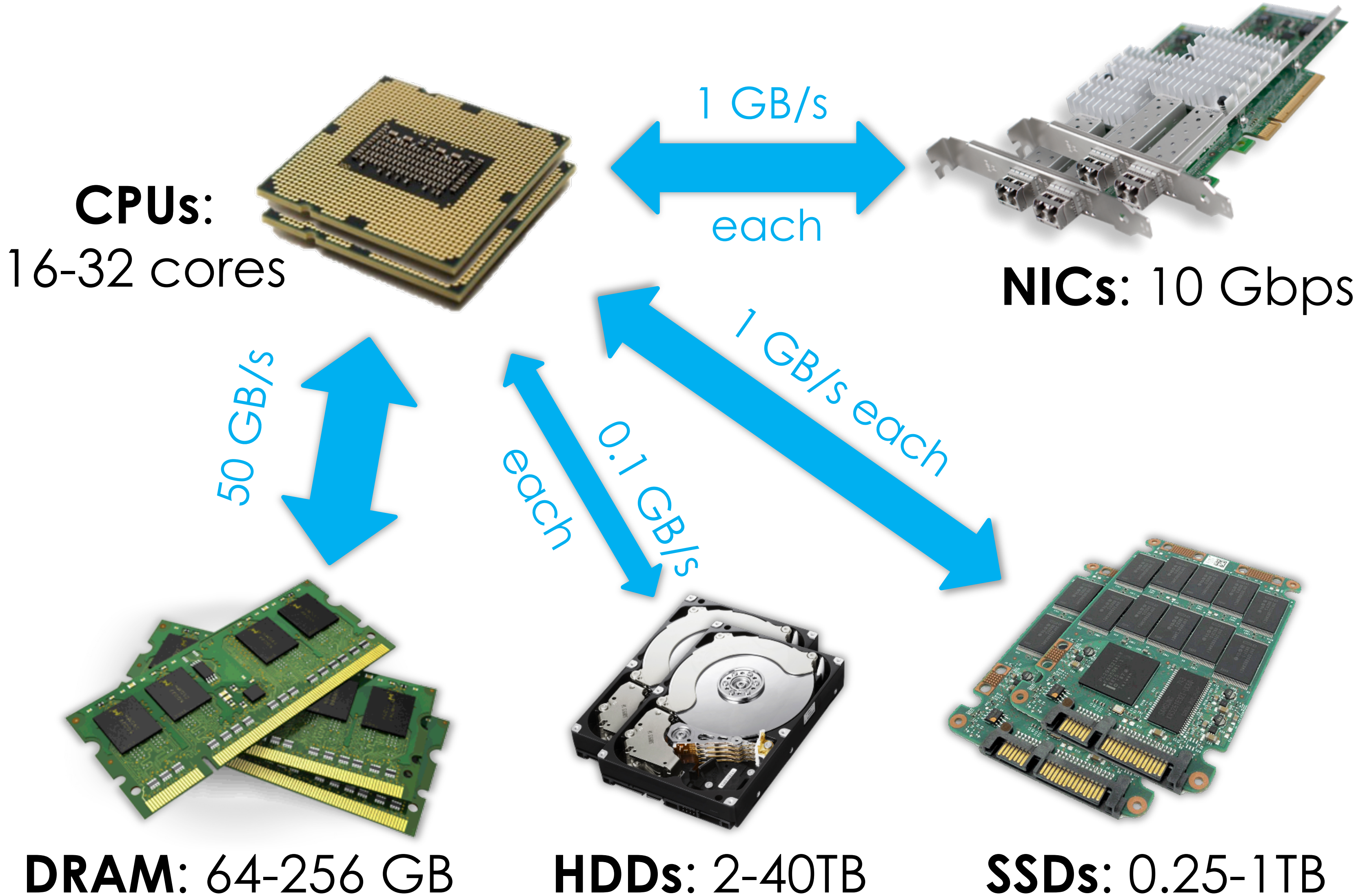
National Annual Conference on HPC, Hefei, 2017

Joint work with *Haibo, Xingda, Jiaxin, Yanzhe, Zhaoguo, Youyang, Yunhao, Siyuan and Chang* @IPDADS

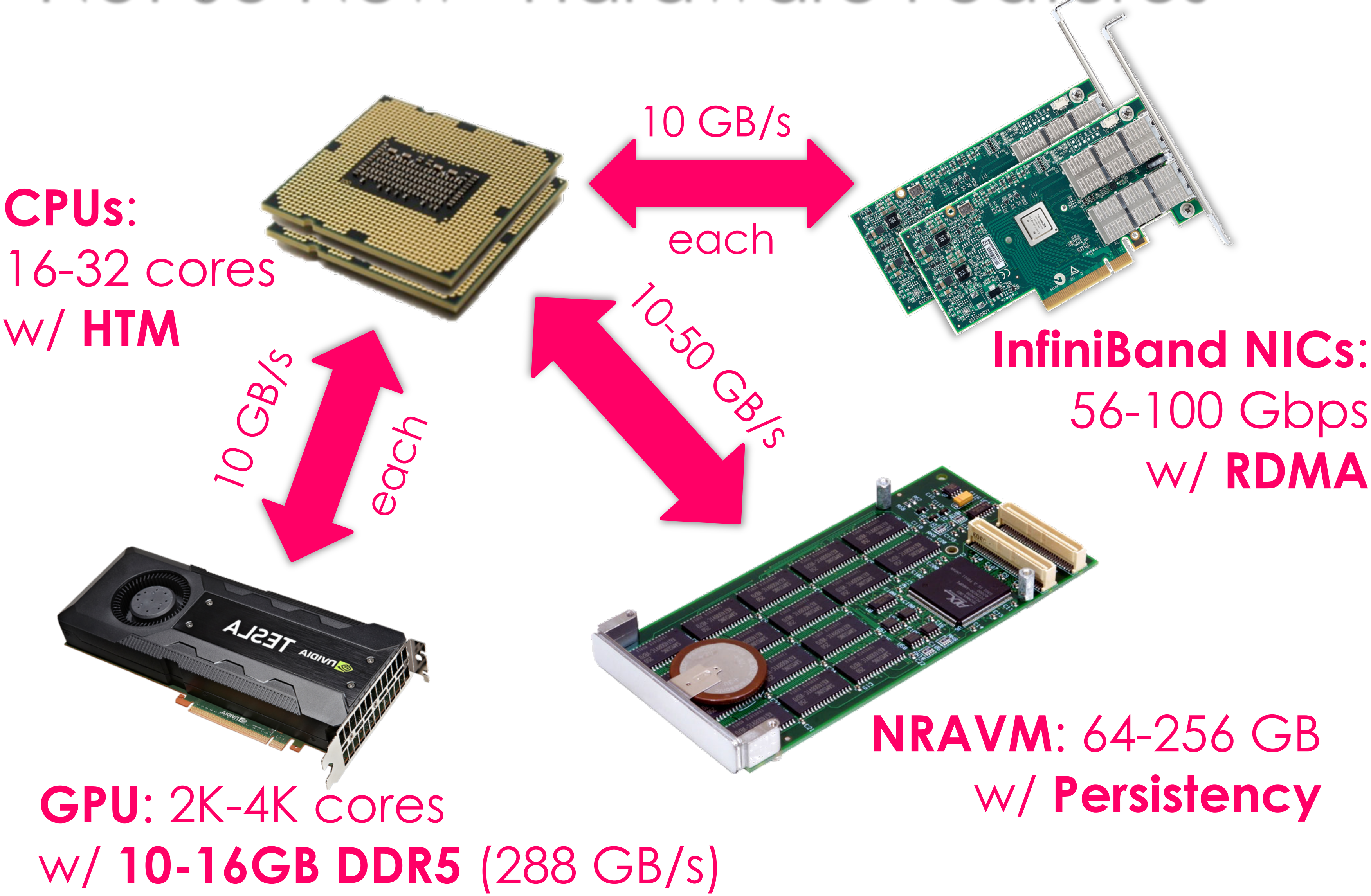
In-Memory Computing



Typical Server @Datacenter

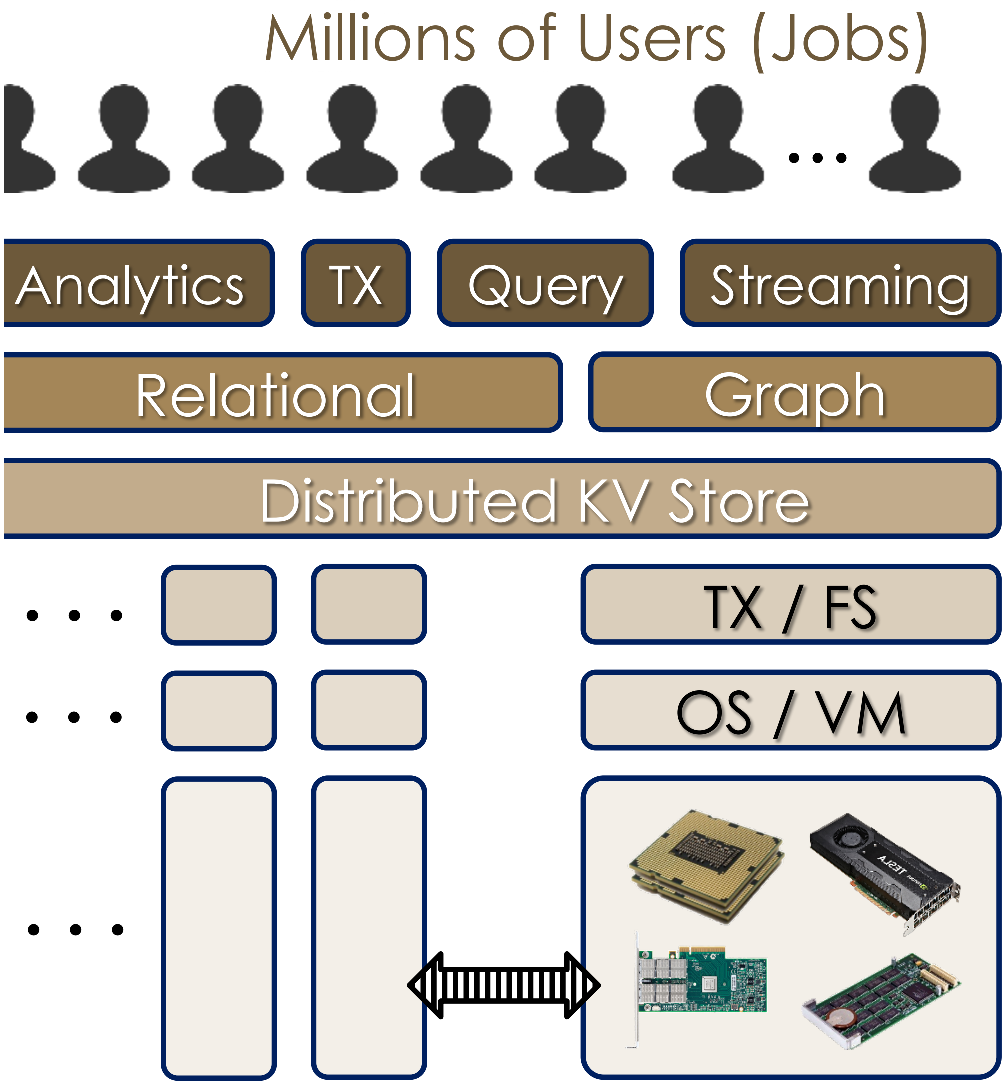


“Not So New” Hardware Features



Our Recent Work

In-memory Systems w/ new hardware features



Wukong
@SOSP’17
@OSDI’16

DrTM
@ACM TOCS’17
@USENIX ATC’17
@EUROSYS’16
@SOSP’15

DBX
@PPOPP’17
@SIGMOD’16
@EUROSYS’14

SoupFS
@USENIX ATC’17

PVM
@SoCC’16

This Talk

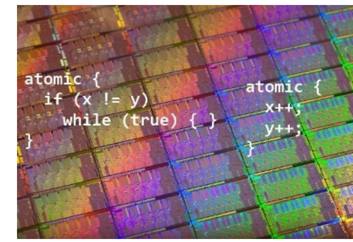
101 of Advanced Hardware Features

Fast Distributed Transactions: HTM + RDMA

High Availability Transactions: HTM + NVM + RDMA

Hybrid Graph Querying: GPU & RDMA

HTM: Hardware Transactional Memory



Herlihy & Moss — **1993**

Sun Rock — **2009**

AMD TX Extension — **2010**

. . . — **2012**

Intel Haswell — **2013**

Massively available

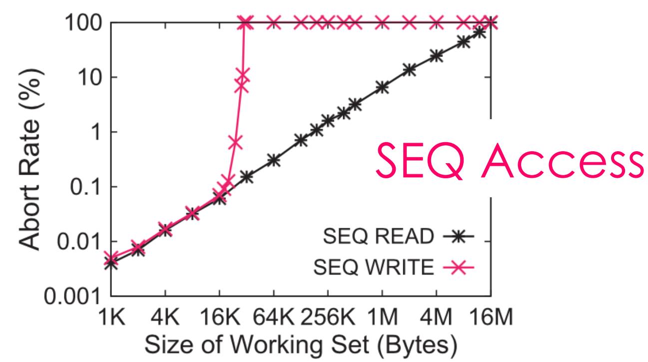
Intel Restricted Transactional Memory

Restricted Transactional Memory (RTM)

- Commercial hardware TM with limitations

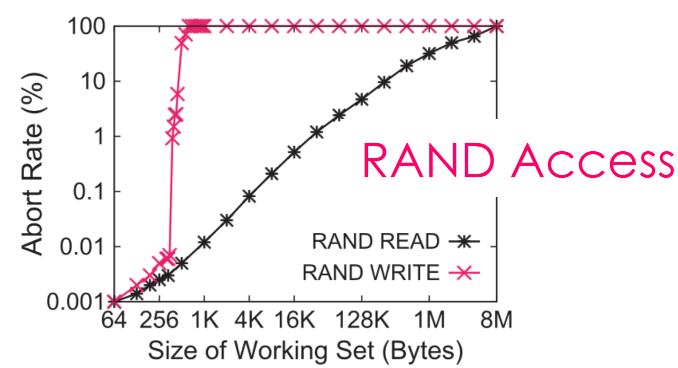
New Instructions

- Xbegin, Xend, Xabort



Major Limitations

1. Working set is limited
2. System events abort TX



RDMA: Remote Direct Memory Access

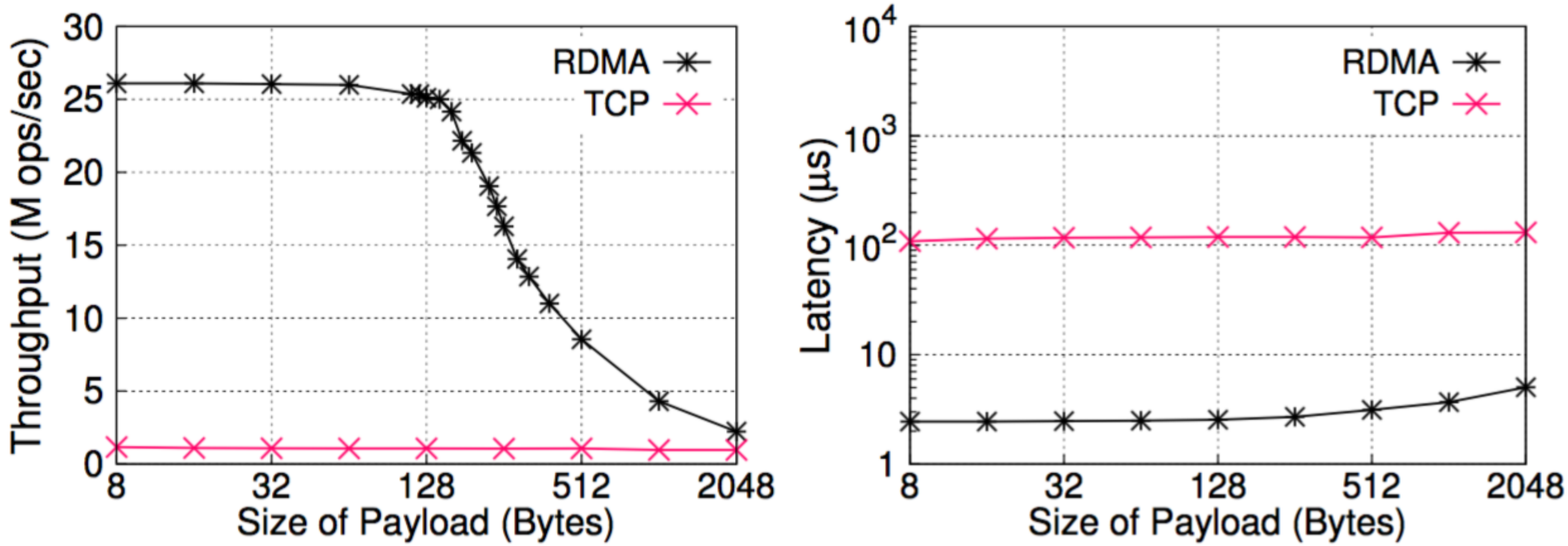
A network feature that allows direct access to the memory of a remote computer

High speed, low latency & low CPU overhead

- Interface: IPoIB, SEND/RECV Verbs, and one-sided RDMA (READ/WRITE/CAS), etc.
- Bypasses OS kernels: Zero copy
- Round-trip time: one-sided/~1-3μs, verb msg/~7μs, IPoIB/~100 μs

One-sided RDMA Performance

Performance of Random Read[1]

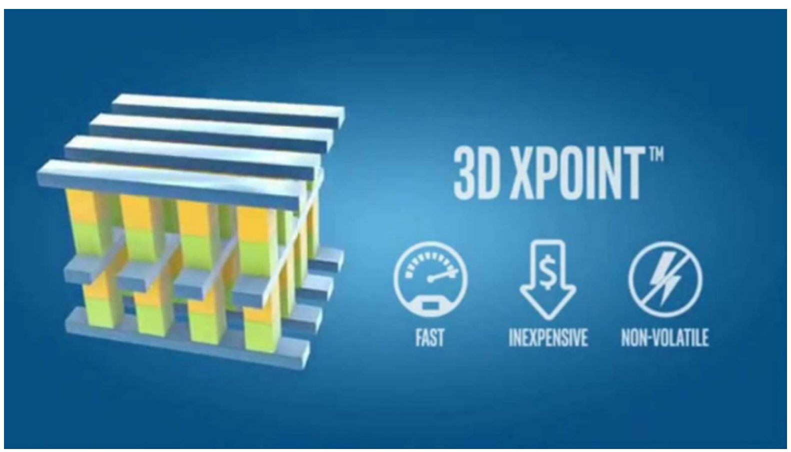


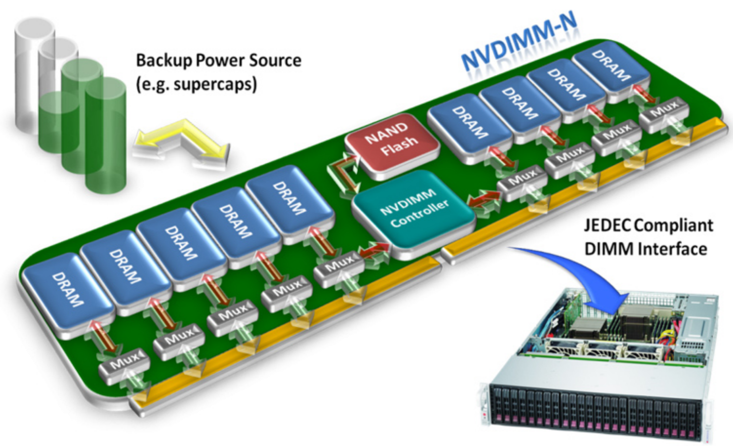
Insensitive to payload size: high/near constant throughput and low latency when payload is smaller than a threshold

[1] Mellanox ConnectX-3 MCX353A 56Gbps InfiniBand NIC

NVM: Non-Volatile Memory

- ✓ Non-volatile
- ✓ Byte-addressable
- ✓ High throughput and low latency





Fast In-memory Transaction Processing using RDMA and HTM

Transaction: Key Pillar for Many Systems

Alibaba.com®

$9.3 billion/day

Demand Speedy Distributed Transaction Over Large Data Volumes

12306

9.56 million tickets/day

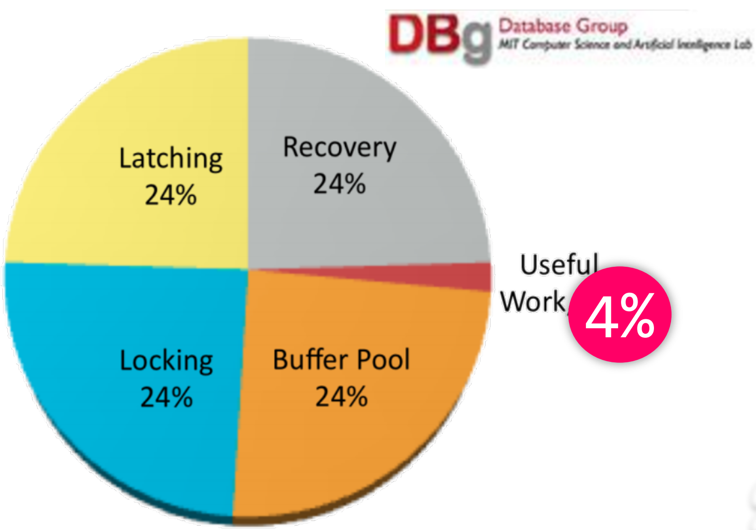
PayPal™

11. 6 million payments/day

Why (Distributed) TXs are Slow?

Only **4% of wall-clock time** spent on useful data processing, while the rest is occupied with **buffer pools**, **locking**, **latching**, **recovery**.[1]

-- Michael *Stonebraker*



Michael Stonebraker
2014 Turing Award

[1] "The Traditional RDBMS Wisdom is All Wrong"

Opportunities: (not so) New HW Features

HTM: **H**ardware **T**ransaction **M**emory

- Allow a group of load & store instructions to execute in an atomic, consistent and isolated (ACI) way

RDMA: **R**emote **D**irect **M**emory **A**ccess

- Provide cross-machine accesses with high speed, low latency and low CPU overhead

Rethink the design of fast in-memory TX systems by **fusing new hardware features**

Opportunities with HTM & RDMA

HTM: **H**ardware **T**ransaction **M**emory

a non-transactional code will unconditionally abort a transaction when their accesses conflict

Strong Atomicity

RDMA: **R**emote **D**irect **M**emory **A**ccess

Opportunities with HTM & RDMA

HTM: **H**ardware **T**ransaction **M**emory

a non-transactional code will unconditionally abort a transaction when their accesses conflict

Strong Atomicity

RDMA: **R**emote **D**irect **M**emory **A**ccess

one-sided RDMA operations are cache-coherent with local accesses

Strong Consistency

Opportunities with HTM & RDMA

HTM: **H**ardware **T**ransaction **M**emory

a non-transactional code will unconditionally abort a transaction when their accesses conflict

RDMA: **R**emote **D**irect **M**emory **A**ccess

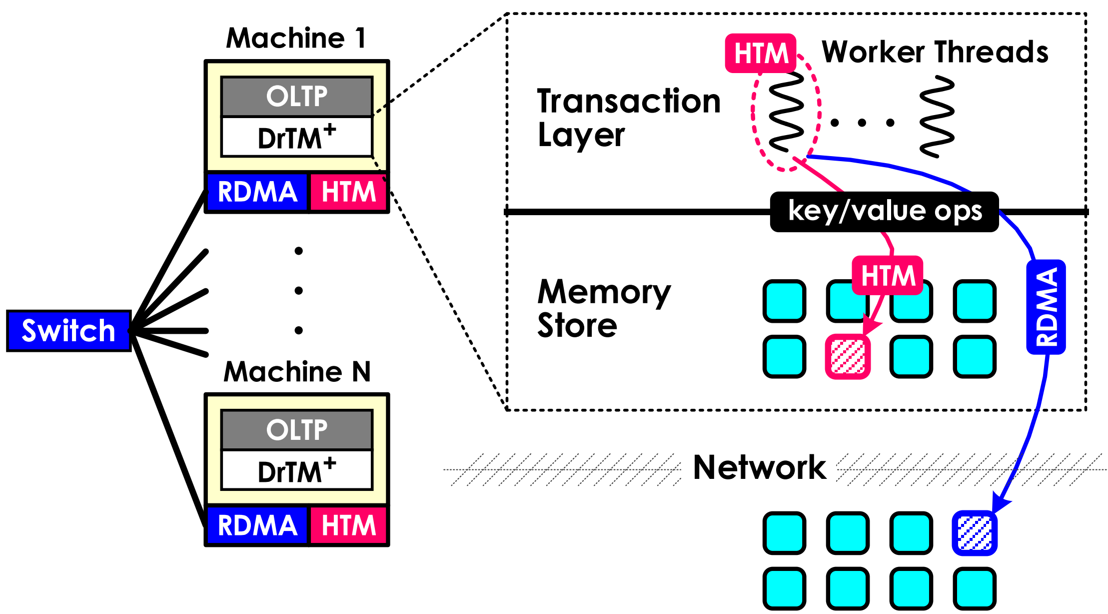
one-sided RDMA operations are cache-coherent with local accesses

HTM Strong Atomicity + **RDMA Strong Consistency** ➡ **Distributed HTM**

System Overview

DrTM : an in-memory transaction processing systems with HTM and RDMA

- Target: OLTP workloads over large volume of data
- Two components: Transaction Layer & Store Layer
- Hybrid OCC Protocol: **HTM + RDMA → Distributed HTM**



How to use HTM for DB Transactions?

Strawman: Using RTM protect {DB TX}

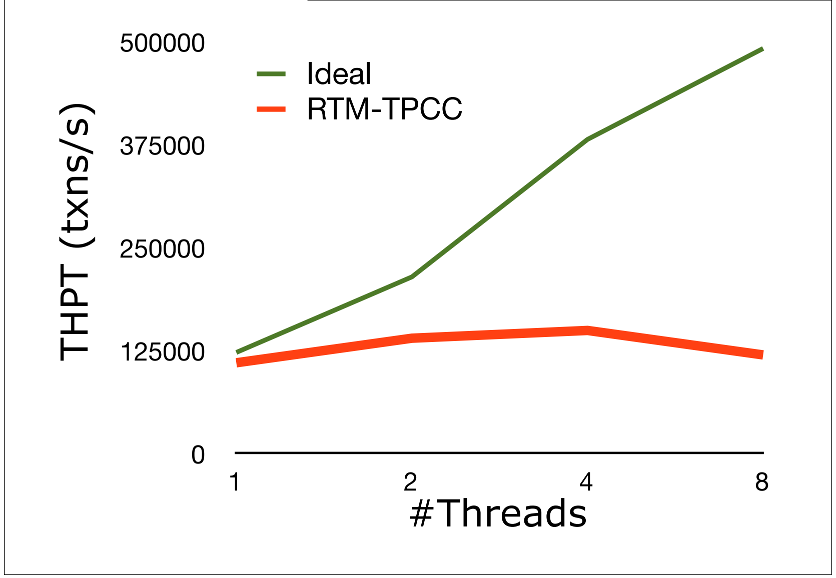
- Observation: RTM (Intel HTM) provides ACI
- Very poor performance!

DB TX

```
RTM Begin
...
DB TX's R/W ops
...
RTM End
```

High RTM abort rate
1. Large working set
2. Random access pattern

TPC-C Benchmark

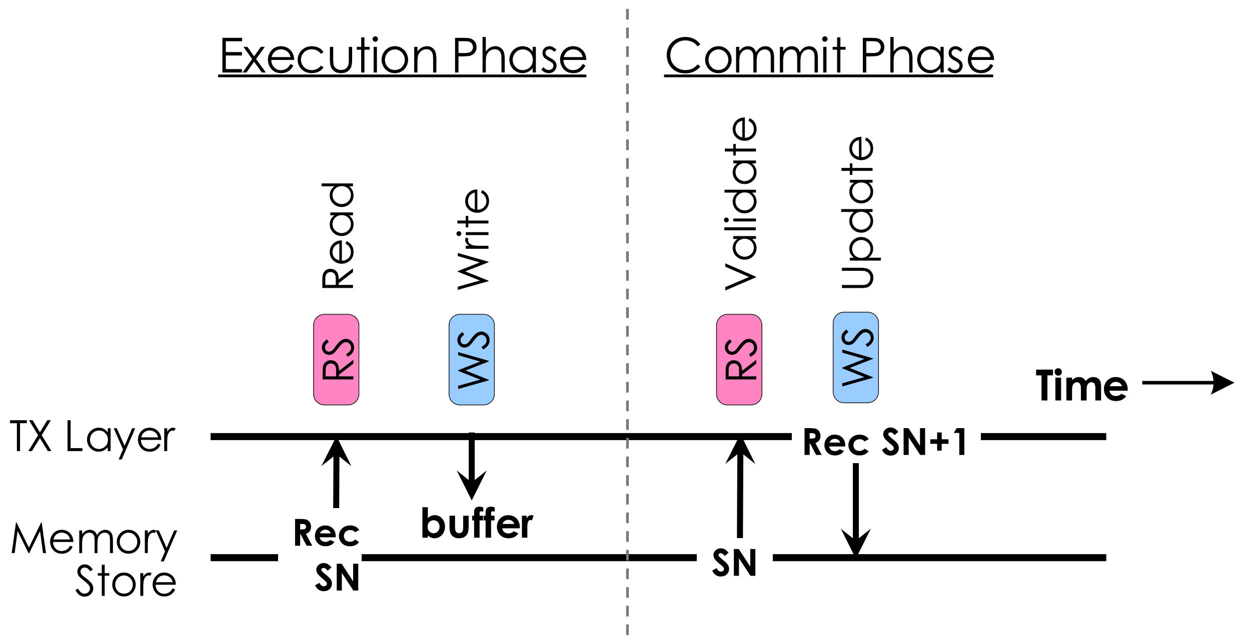


Implement CC protocol with RTM

Optimistic Concurrency Control (OCC)

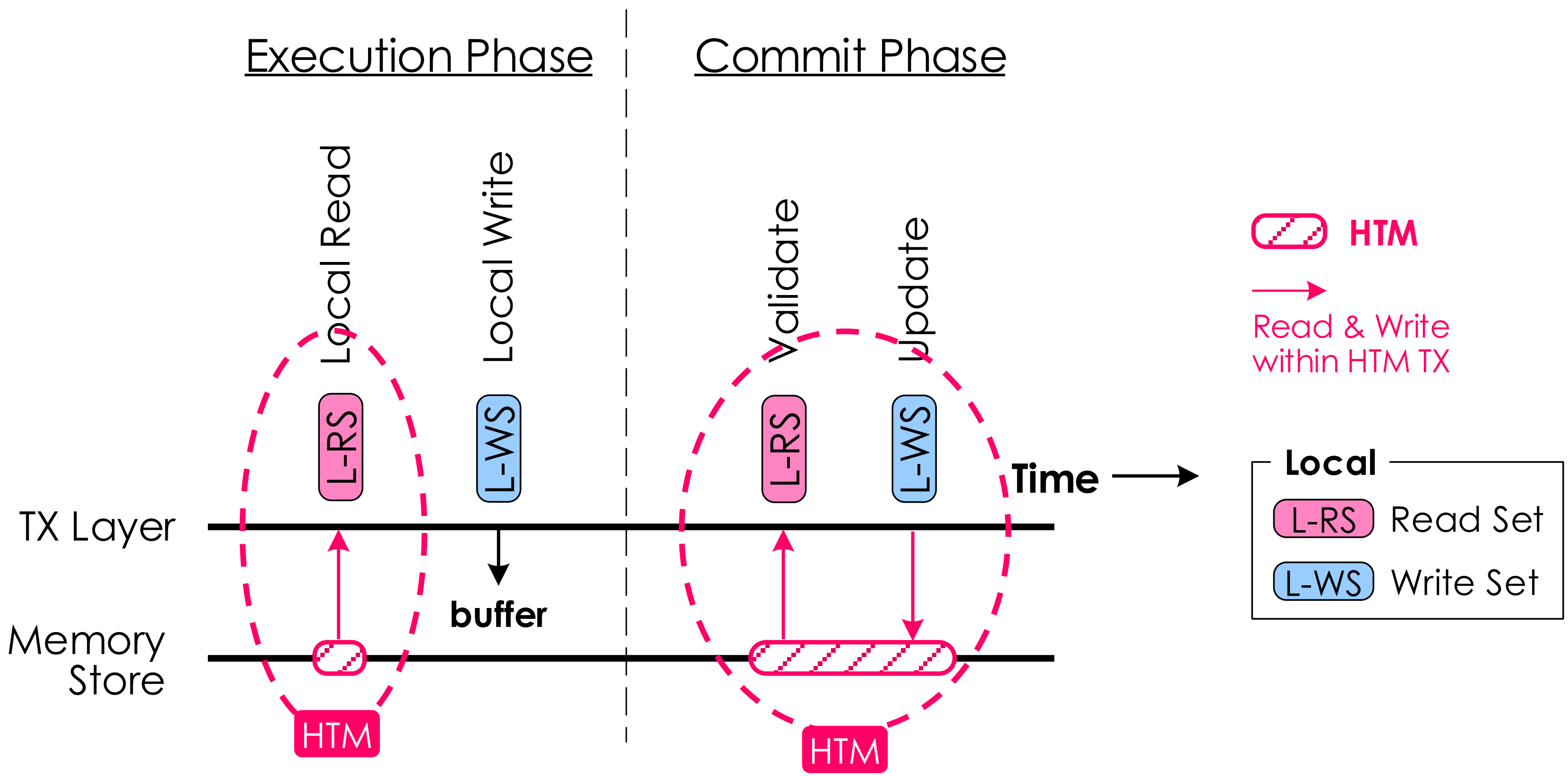
* *Each record with a version (SN=sequence number)*

1. *Optimistic* Execution: **Read**/rec+SN, **Write**/local buf
2. Commit: **Validate** R's SN, then **Update** W's buf



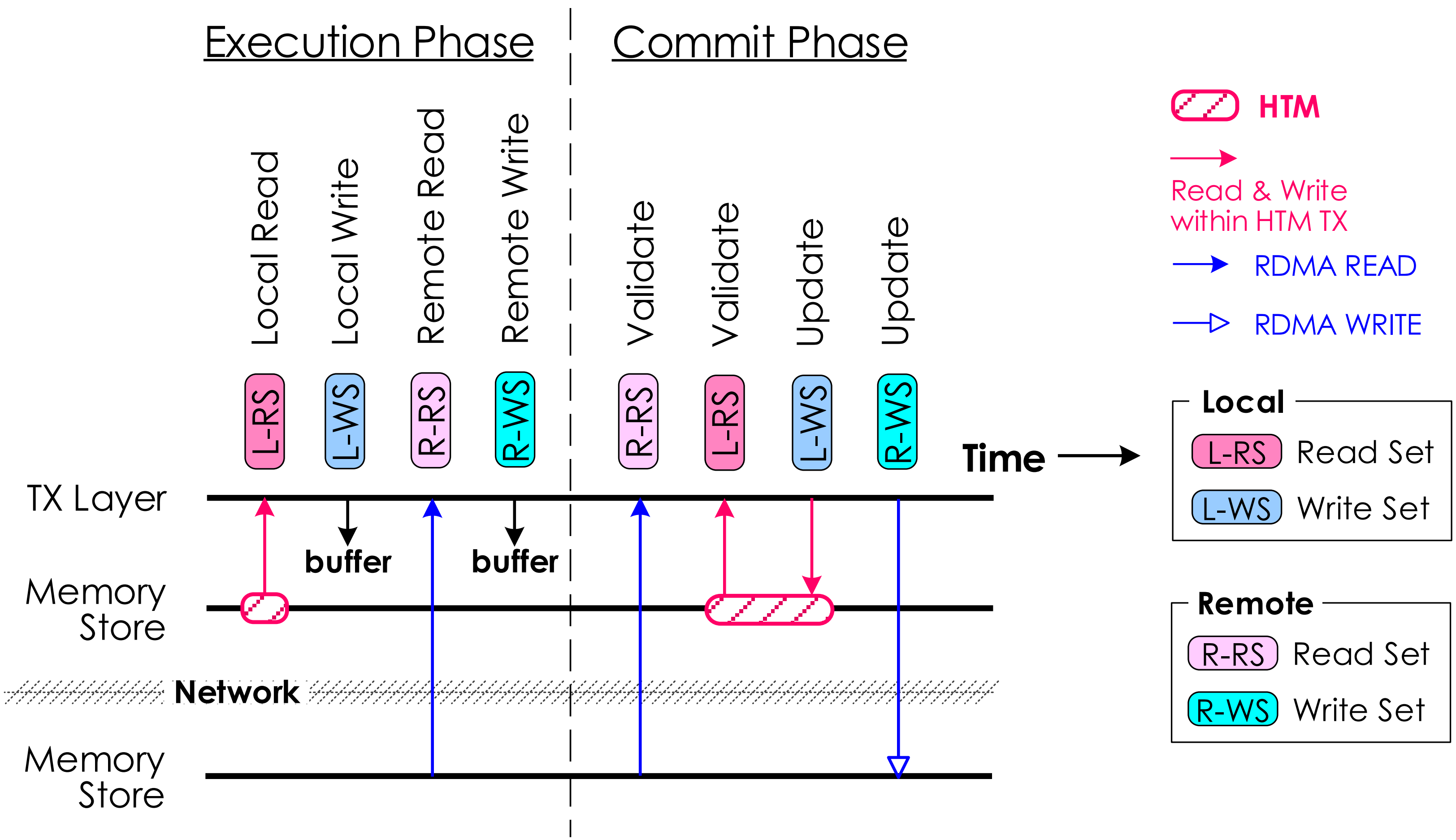
Single Machine: DBX[EuroSys'14]

HTM-friendly Optimistic Concurrency Control



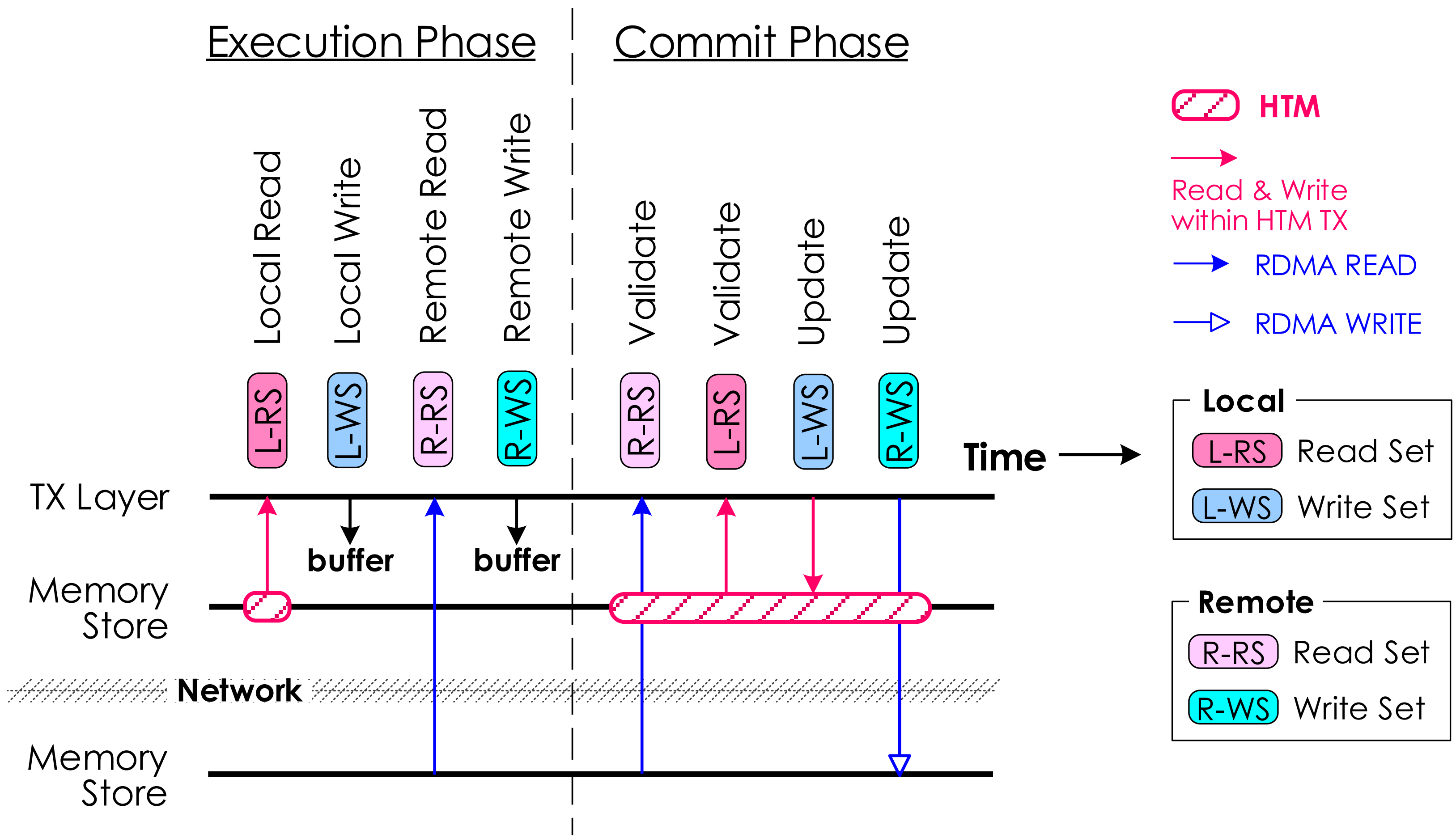
First Attempt

Use **RDMA** to fetch and commit remote records



First Attempt

Use **RDMA** to fetch and commit remote records



Challenge#1: Restriction of HTM

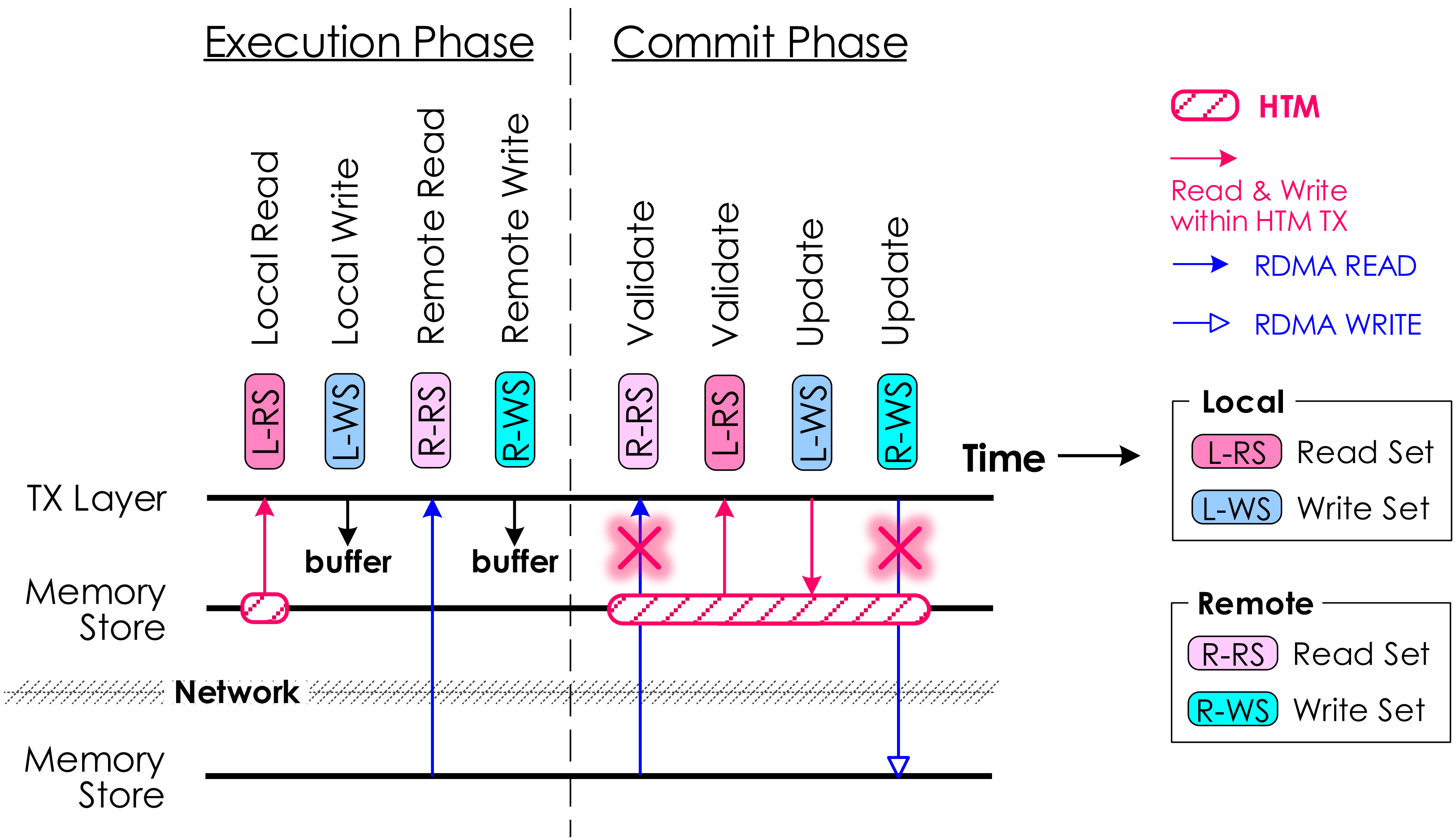
HTM is only a compelling hardware feature for single machine platform

Some instructions & system events (e.g., I/O) will unconditionally abort **HTM** transactions

- Like any RDMA ops: READ/WRITE, CAS

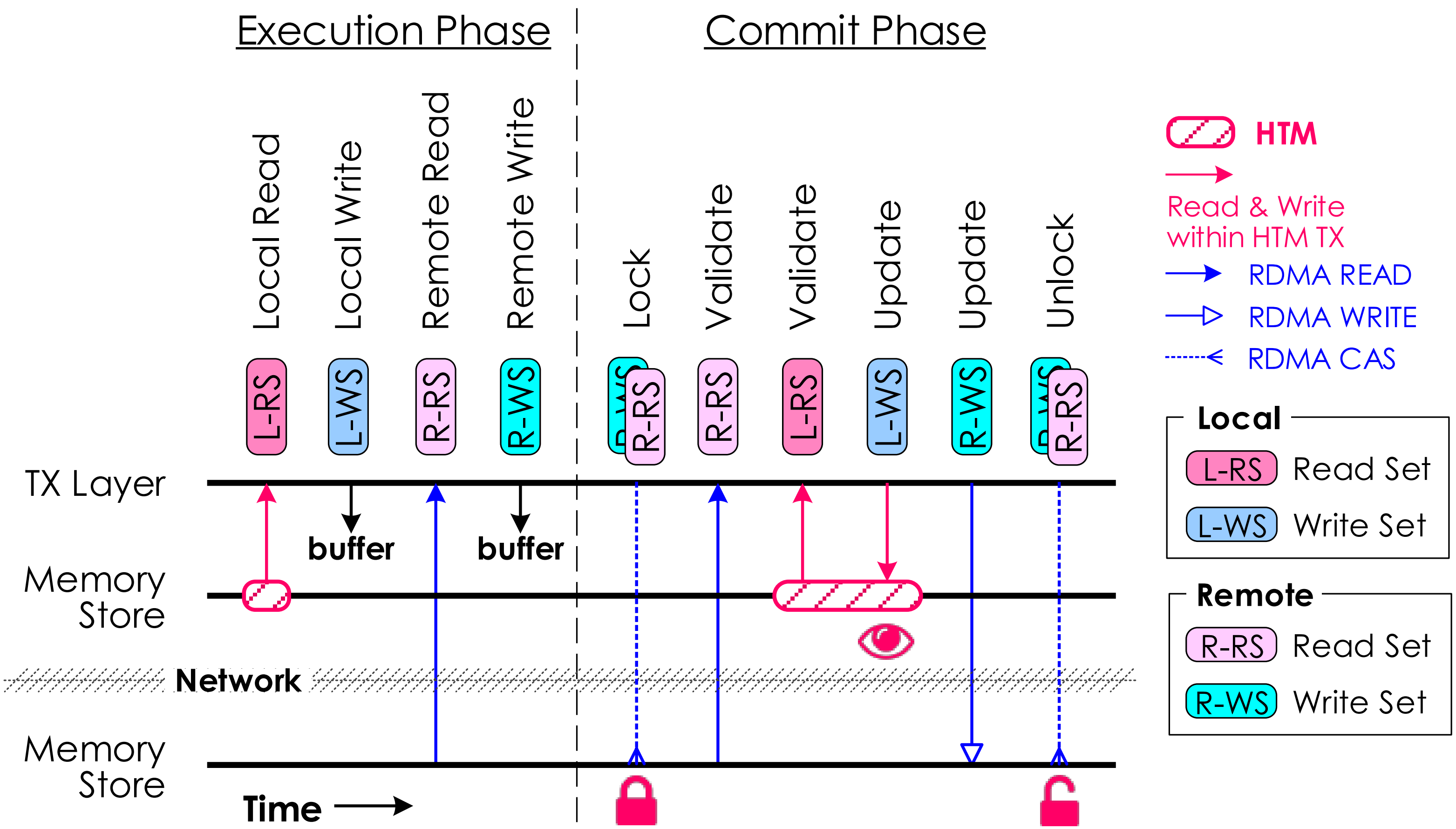
Challenge#1: Restriction of HTM

RDMA ops will abort HTM transactions



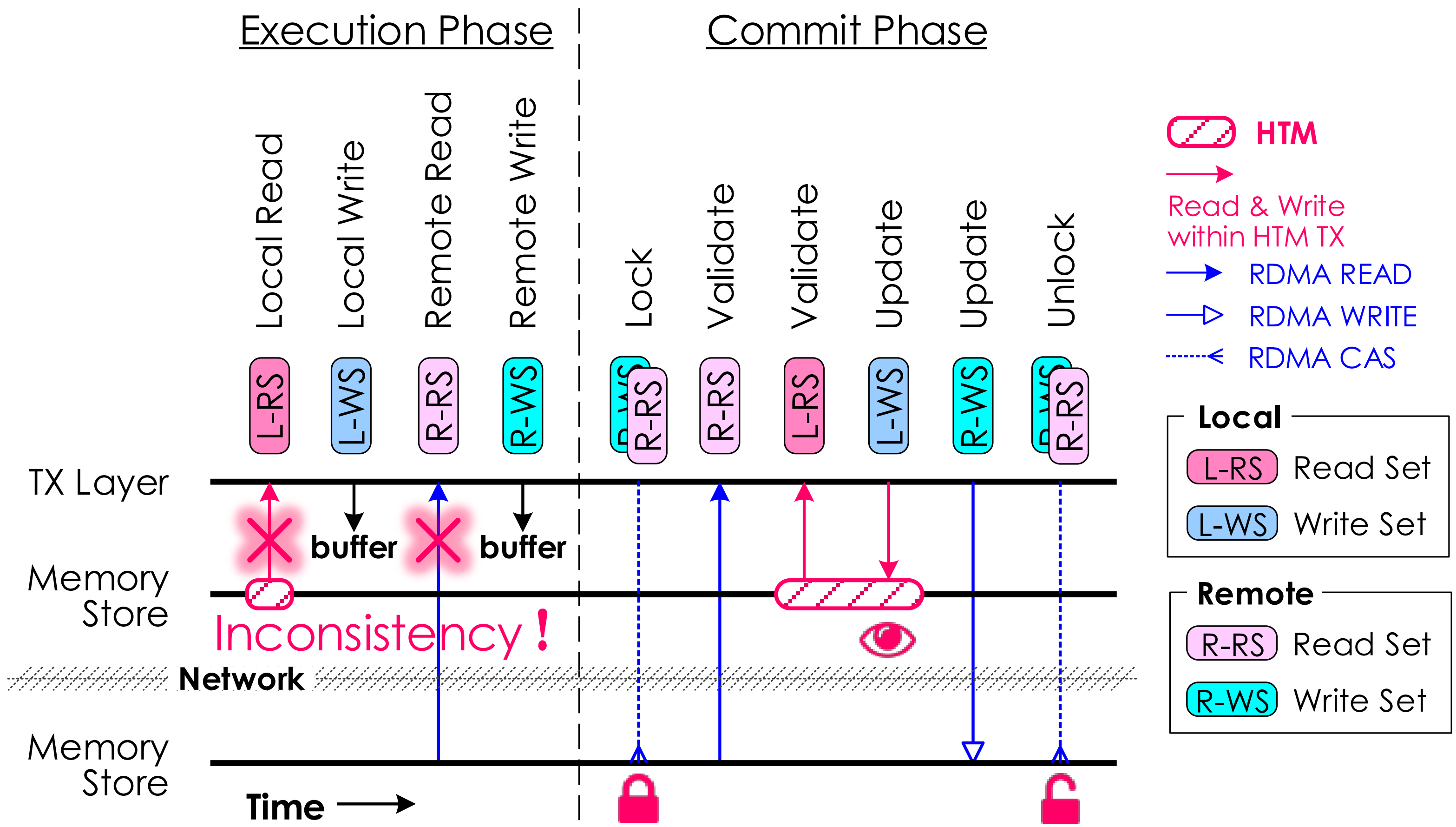
Combining OCC with locking

+RDMA-based locking prior to commit phase



Challenge#2: Consistency of RDMA

Local HTM ops meet Remote RDMA ops



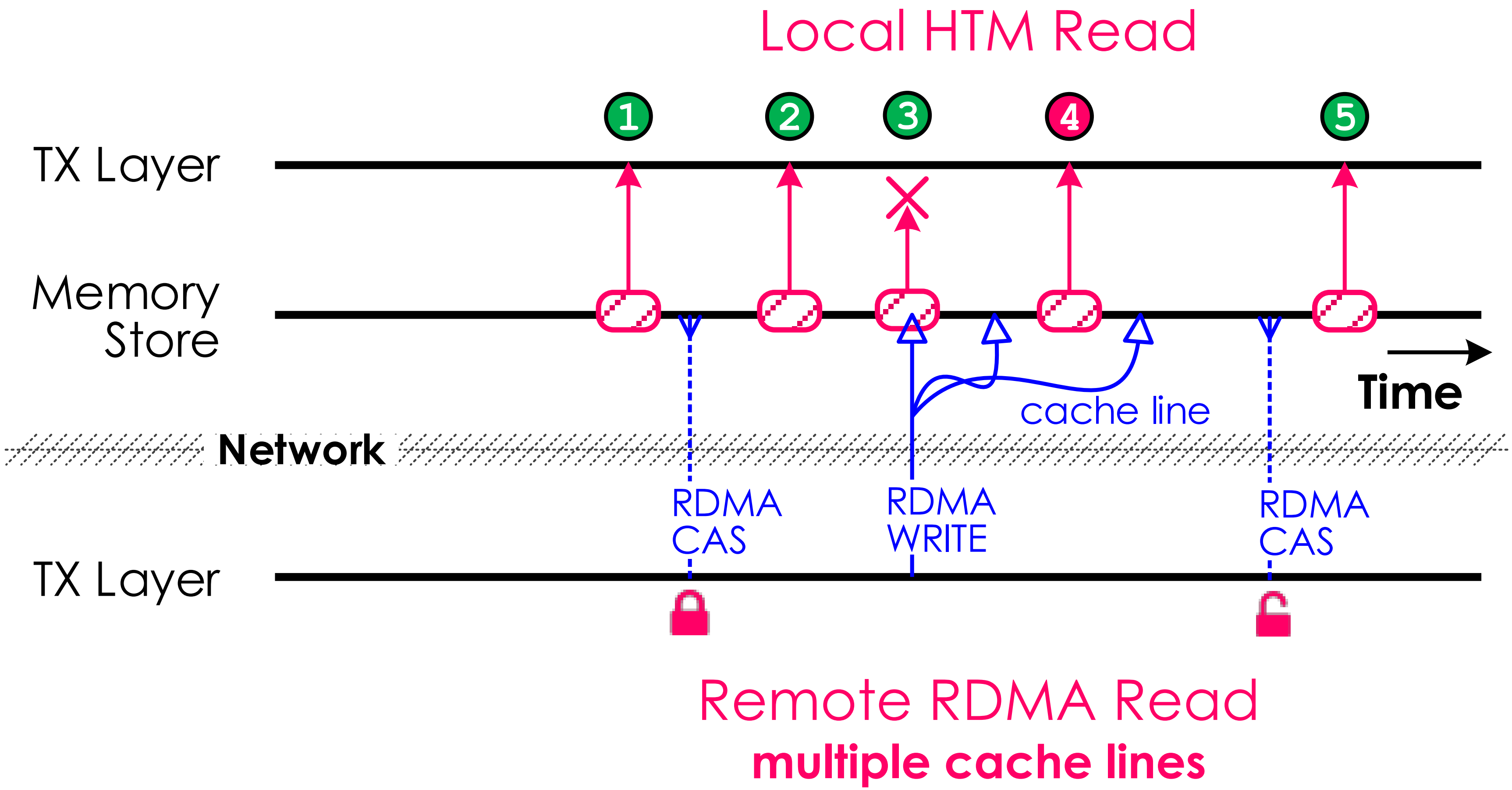
Challenge#2: Consistency of RDMA

One-sided **RDMA** ops are only cache-coherent within a single cache line

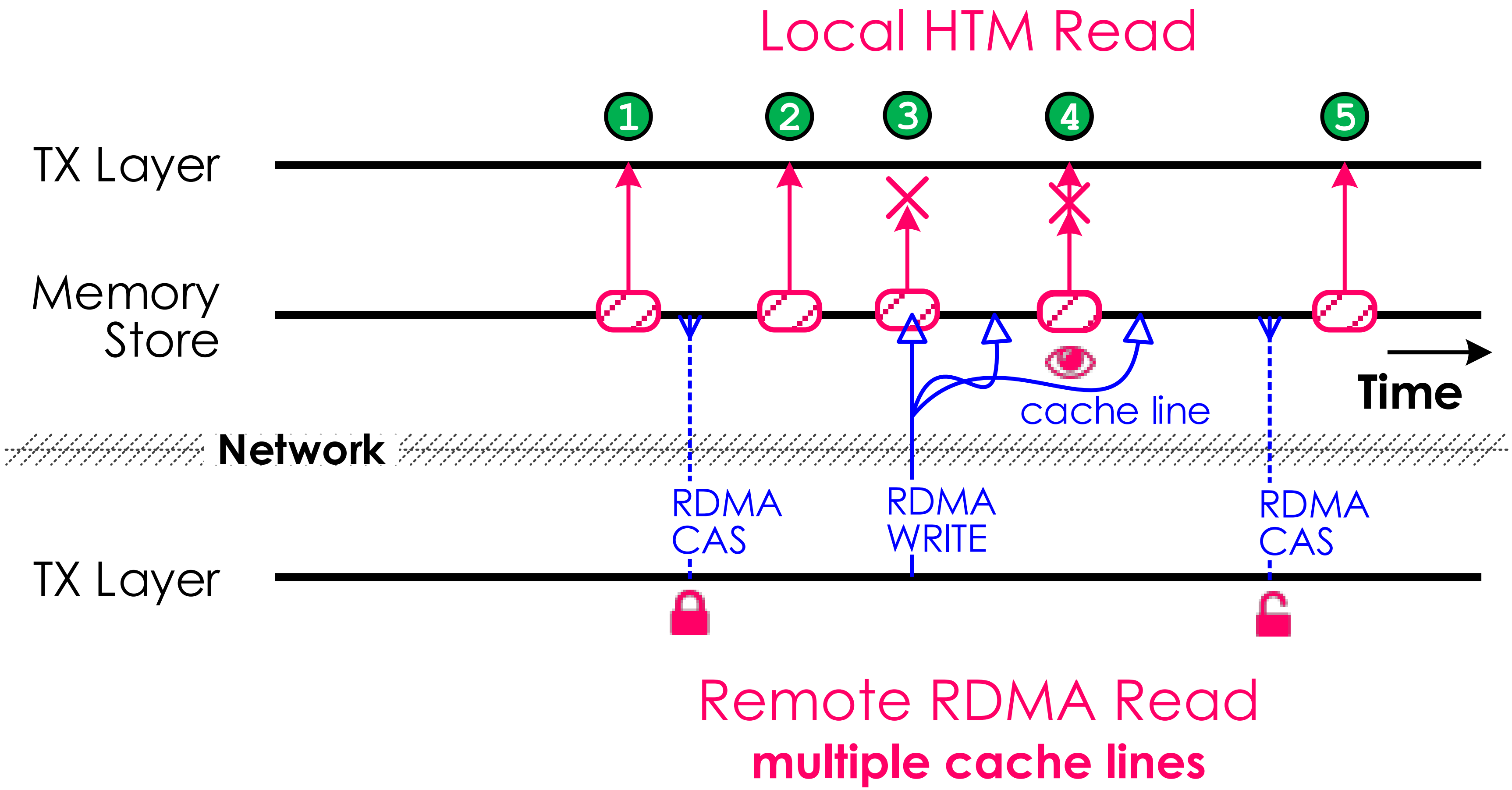
One RDMA op is multiple sub-ops to the target machine

- Consistency of multiple sub-ops is **NOT** guaranteed

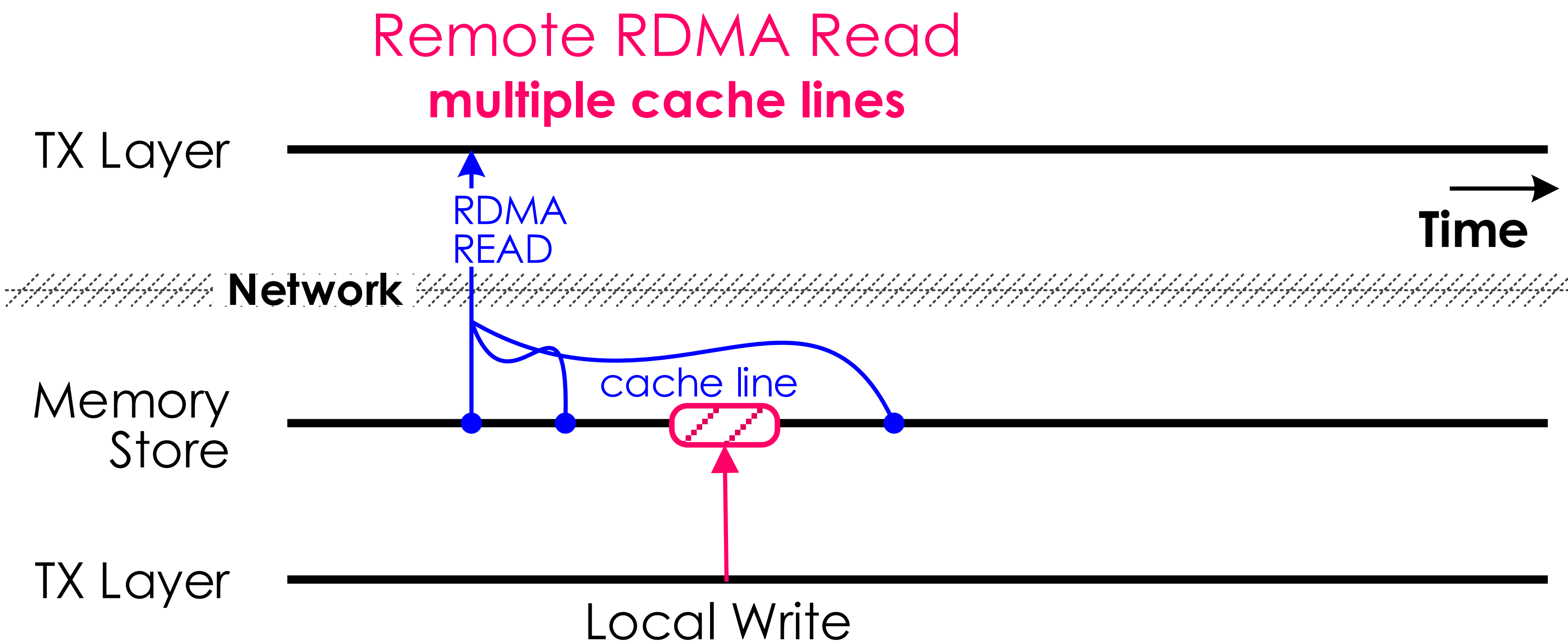
Local Read Consistency



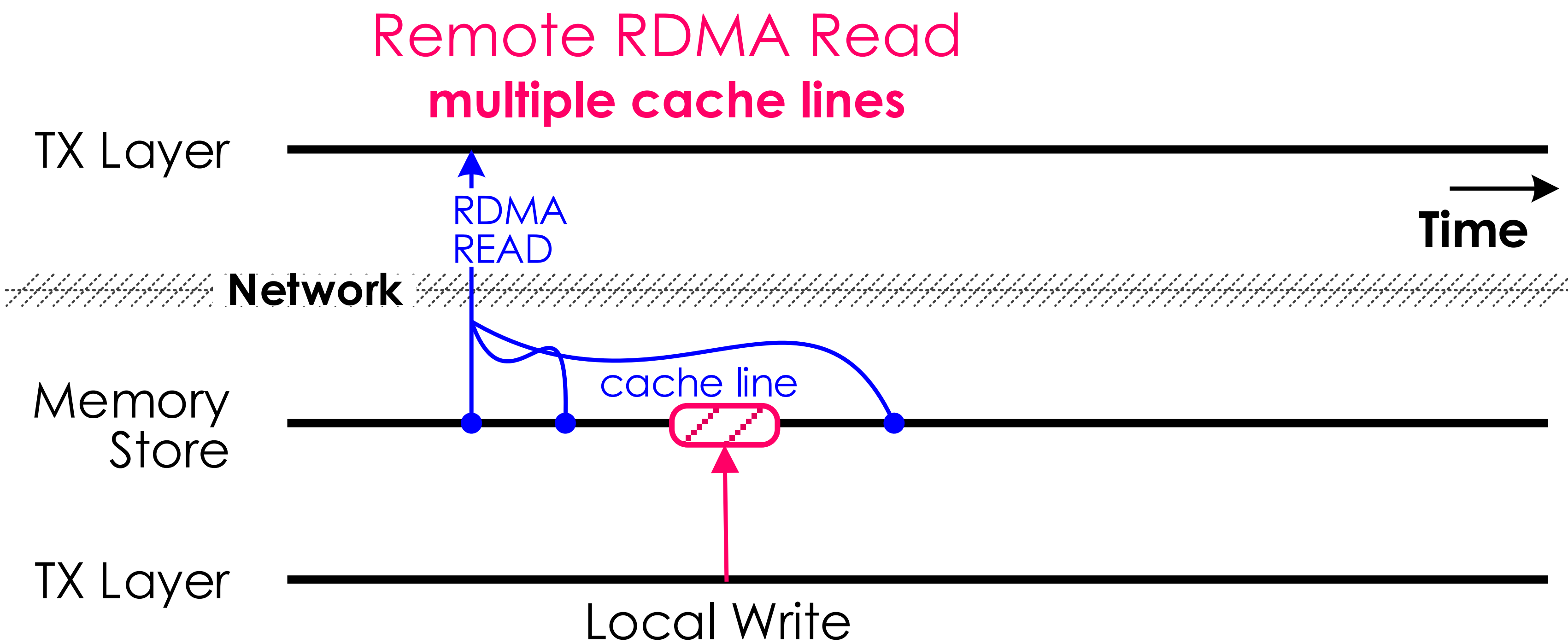
Check the lock in Local Read

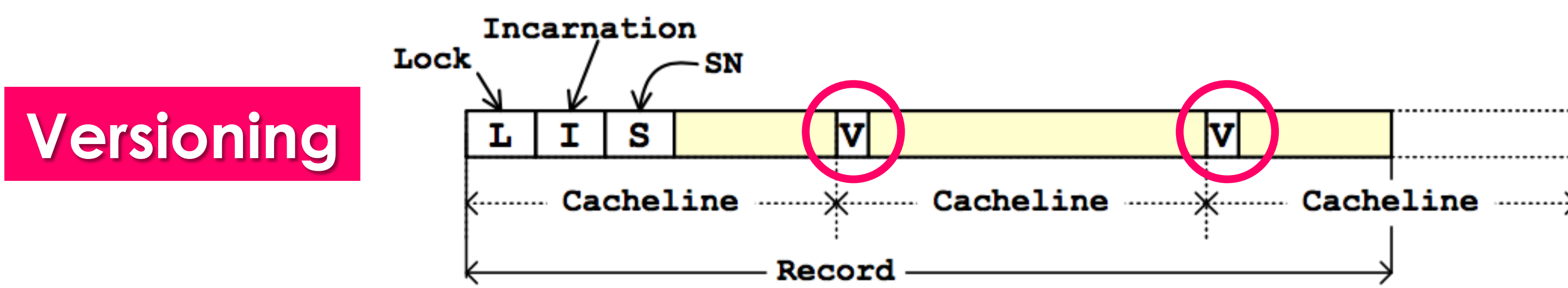


Remote Read Consistency

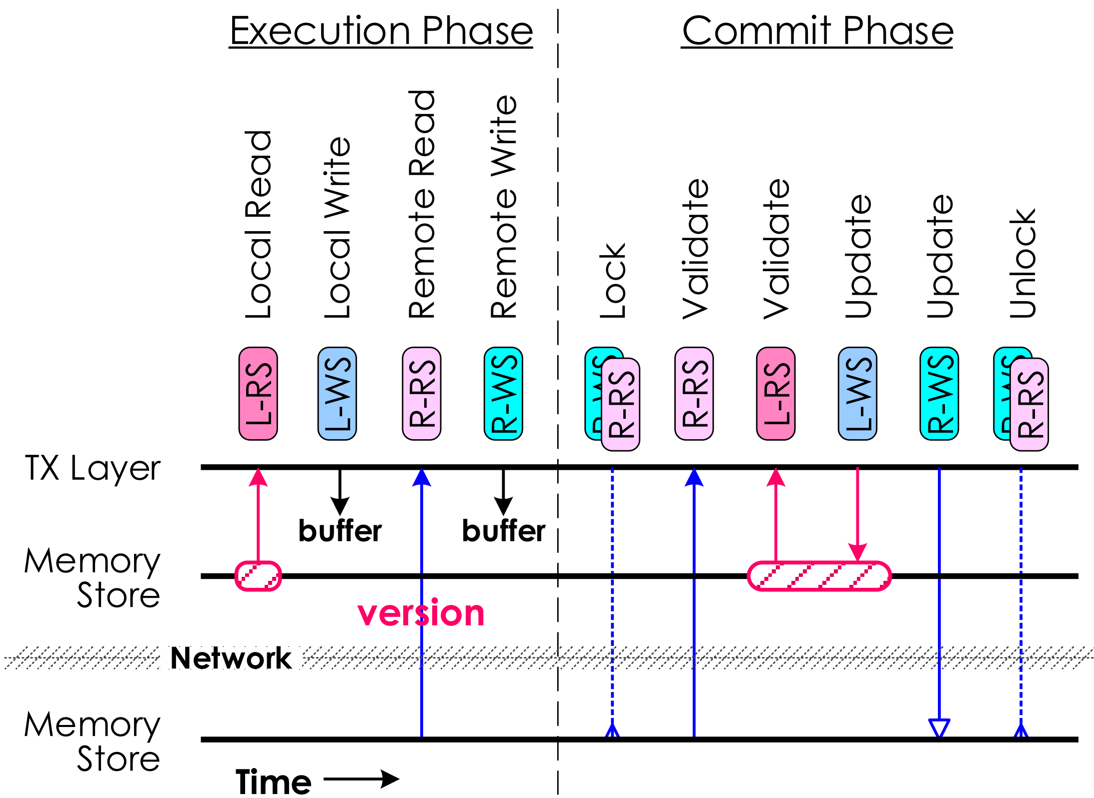


Check the version in Remote Read





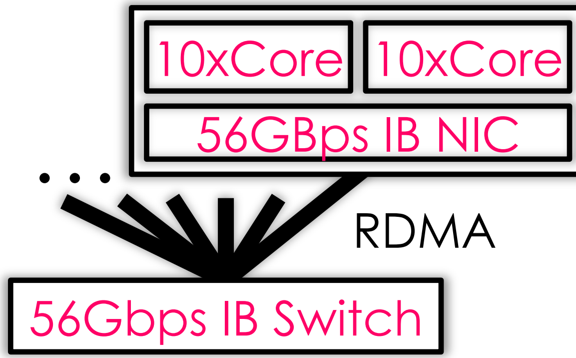
Hybrid OCC Protocol Summary



ISOLATION	COMMIT/L	COMMIT/R
COMMIT/L	**HTM**	
COMMIT/R		

CONSISTENCY	COMMIT/L	COMMIT/R
READ/L	**HTM**	
READ/R		

Evaluation



Baseline: *Calvin*

Platforms: A small-scale 6-machine cluster

- Each: two 10-cores, RTM-enabled Intel Xeon E5-2650 (disabled HT), 64GB DRAM, Mellanox ConnectX-3 MCX353A 56Gbps InfiniBand NIC w/ RDMA[1]

Benchmarks[2]

- TPC-C
- SmallBank

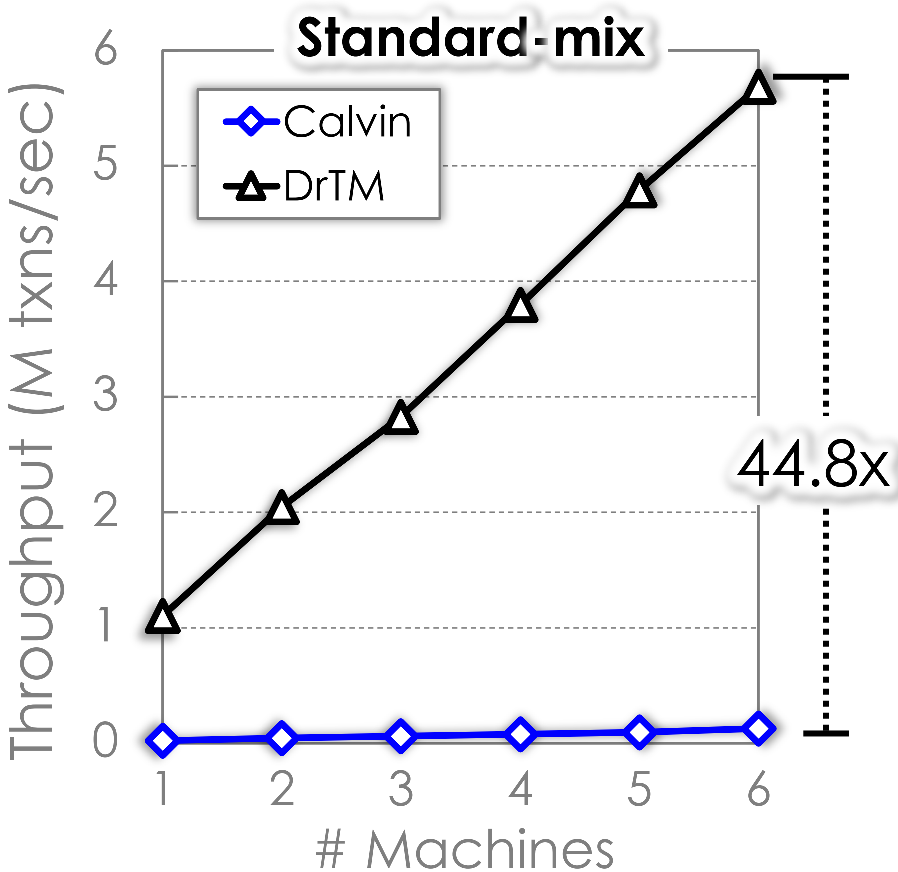
TPC-C	NEW	PAY	DLY	OS	SL
Ratio	45%	43%	4%	4%	4%
Type	d+rw	d+rw	l+rw	l+ro	l+ro

SmallBank	SP	AMG	BAL	DC	WC	TS
Ratio	25%	15%	15%	15%	15%	15%
Type	d+rw	d+rw	l+ro	l+rw	l+rw	l+rw

[1] All machines run Ubuntu 14.04 with Mellanox OFED v3.0-2.0.1 stack.
[2] **d** and **l** stand for distributed and local. **rw** and **ro** stand for read-write and read-only.

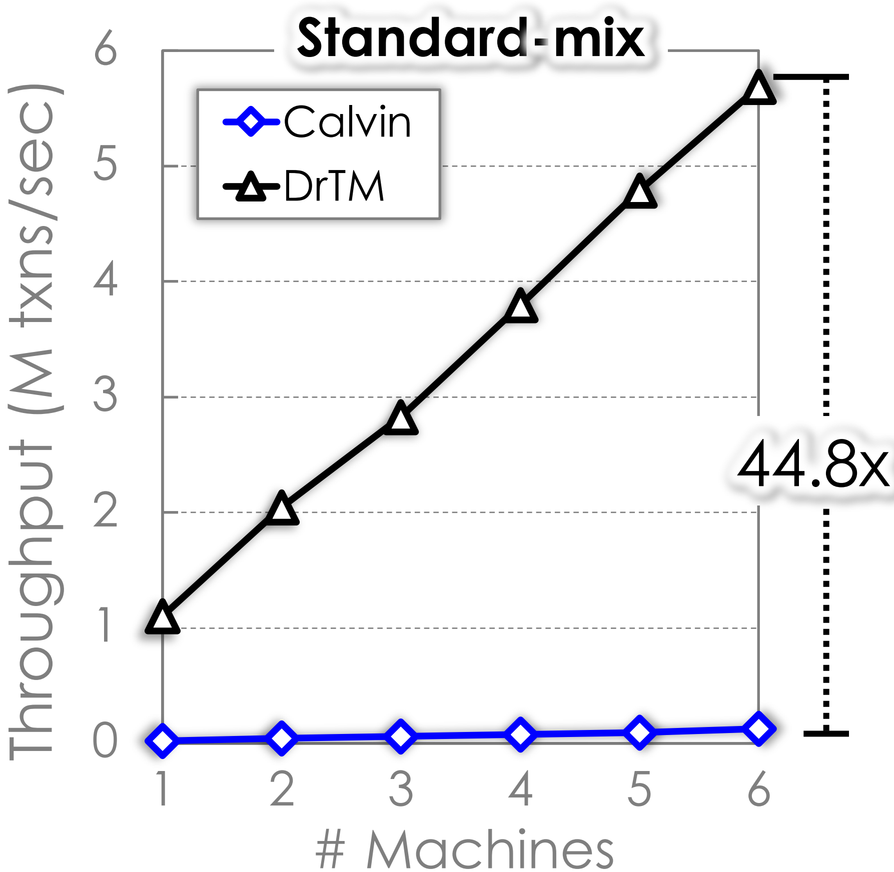
Performance on TPC-C

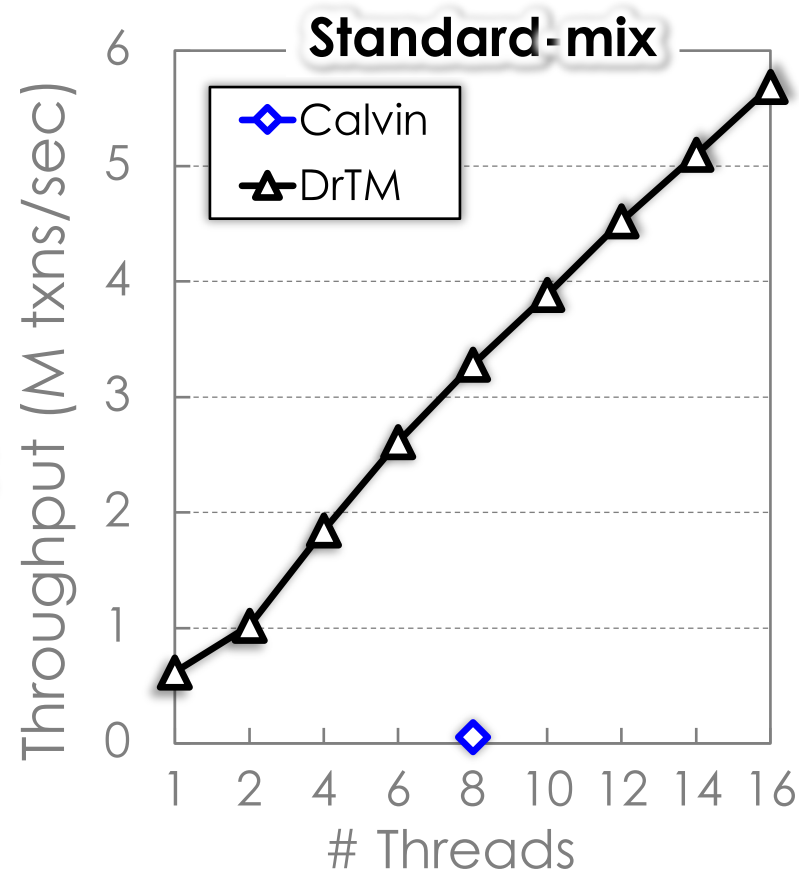
New-order TX
≈ Standard-mix x45%



Performance on TPC-C

New-order TX
≈ Standard-mix x45%

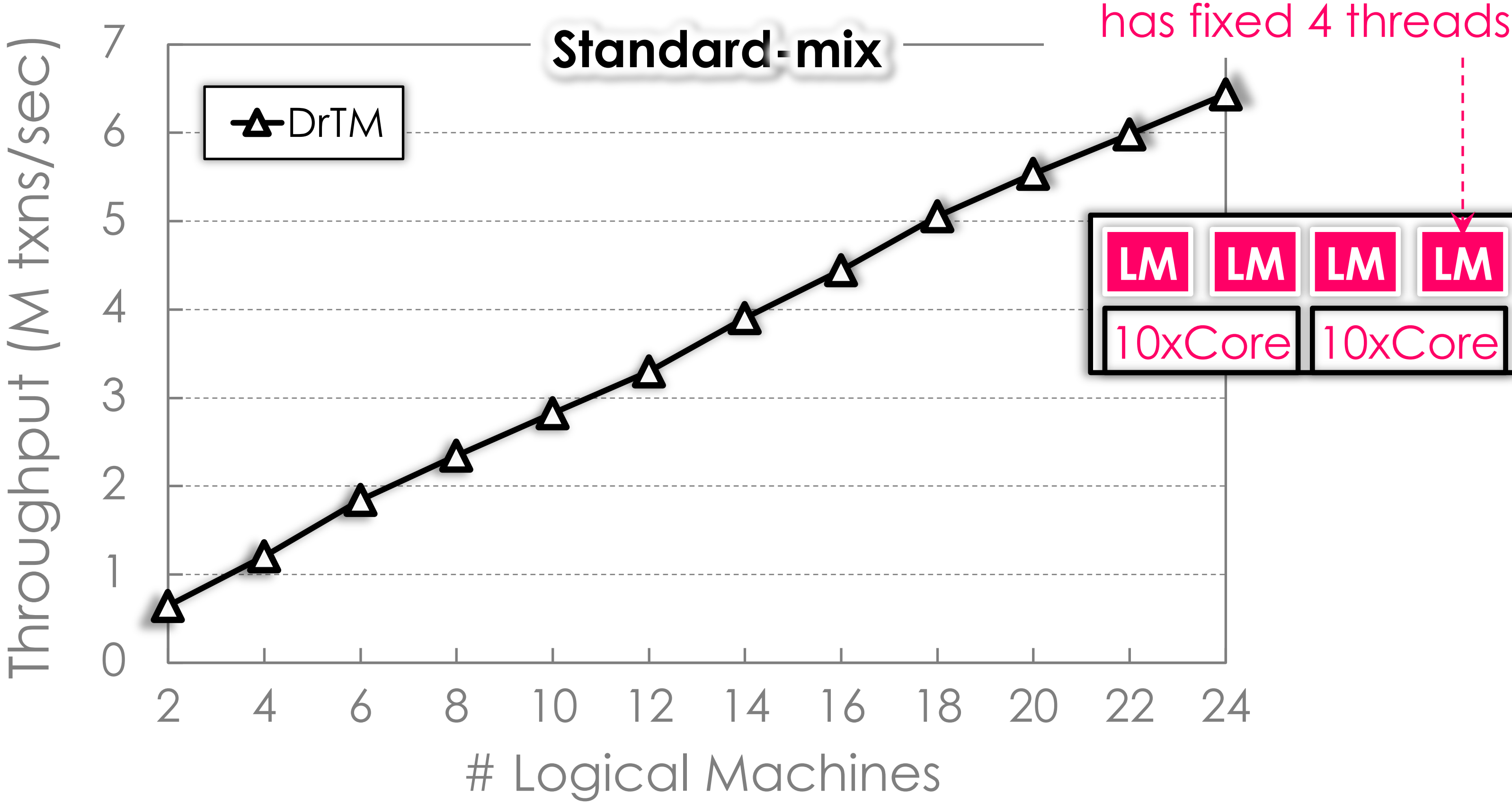




Scalability on TPC-C

New-order TX
≈ Standard-mix x45%

Each logical machine
has fixed 4 threads



NOTE: the interaction btw. two logical nodes sharing the same machine still uses our RDMA-friendly OCC protocol

High Availability of In-memory Transaction Processing with NVM

I Can't Believe It's Not Availability!

GitLab suffers major backup failure after data deletion incident

Posted Feb 1, 2017 by Natasha Lomas (@riptari)

Microsoft loses Sidekick users' personal data

Microsoft subsidiary Danger s
disruption, will almost certair

Google apologizes for 11 hours of Gmail issues affecting 29% of emails, blames dual network failure

Massive Amazon cloud service outage disrupts sites

Amazon's massive AWS outa

One incorrect command and the whole internet suffers.

BY JASON DEL REY | @DELREY | MAR 2, 2017, 2:20PM EST

Aliyun cloud suffers prolonged disruption in Hong Kong

Opportunities with HTM & RDMA

HTM: **H**ardware **T**ransaction **M**emory

> *a non-transactional code will unconditionally abort a transaction when their accesses conflict*

Strong Atomicity

RDMA: **R**emote **D**irect **M**emory **A**ccess

> *one-sided RDMA operations are cache-coherent with local accesses*

Strong Consistency

Opportunities with NVM

HTM: **H**ardware **T**ransaction **M**emory

a non-transactional code will unconditionally abort a transaction when their accesses conflict

Strong Atomicity

RDMA: **R**emote **D**irect **M**emory **A**ccess

one-sided RDMA operations are cache-coherent with local accesses

Strong Consistency

NVM: **N**on-**V**olatile **M**emory

- Provide random-access memory that retains its information when power is turned off (non-volatile)

Opportunities with NVM

HTM: **H**ardware **T**ransaction **M**emory

a non-transactional code will unconditionally abort a transaction when their accesses conflict

Strong Atomicity

RDMA: **R**emote **D**irect **M**emory **A**ccess

one-sided RDMA operations are cache-coherent with local accesses

Strong Consistency

NVM: **N**on-**V**olatile **M**emory

logging TX to NVM before commit can ensure data durability after machine failure

Strong Durability

Opportunities with NVM

HTM: **H**ardware **T**ransaction **M**emory

> *a non-transactional code will unconditionally abort a transaction when their accesses conflict*

Strong Atomicity

RDMA: **R**emote **D**irect **M**emory **A**ccess

> *one-sided RDMA operations are cache-coherent with local accesses*

Strong Consistency

NVM: **N**on-**V**olatile **M**emory

> *logging TX to NVM before commit can ensure data durability after machine failure*

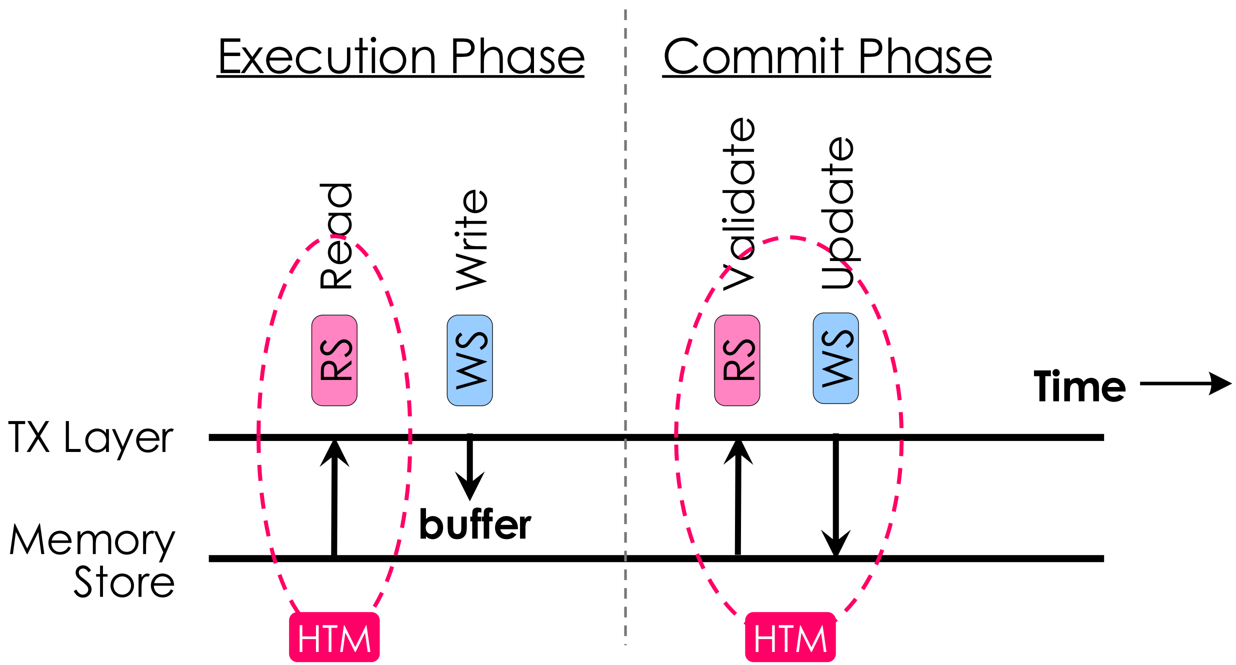
Strong Durability

NVM Strong Durability + **Replication** ➡ **High Availability**

System Overview

DrTM+R : Embrace performance & availability

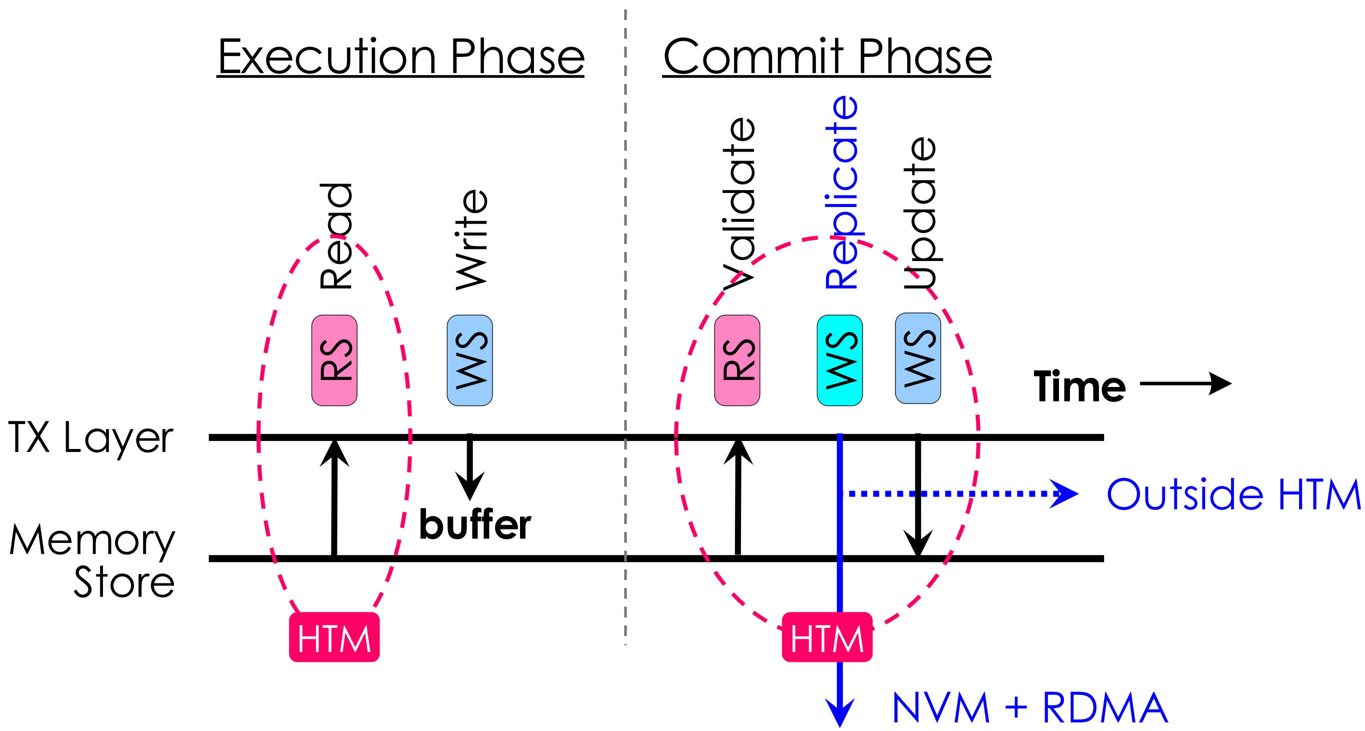
- Hybrid OCC Protocol with HTM and RDMA



System Overview

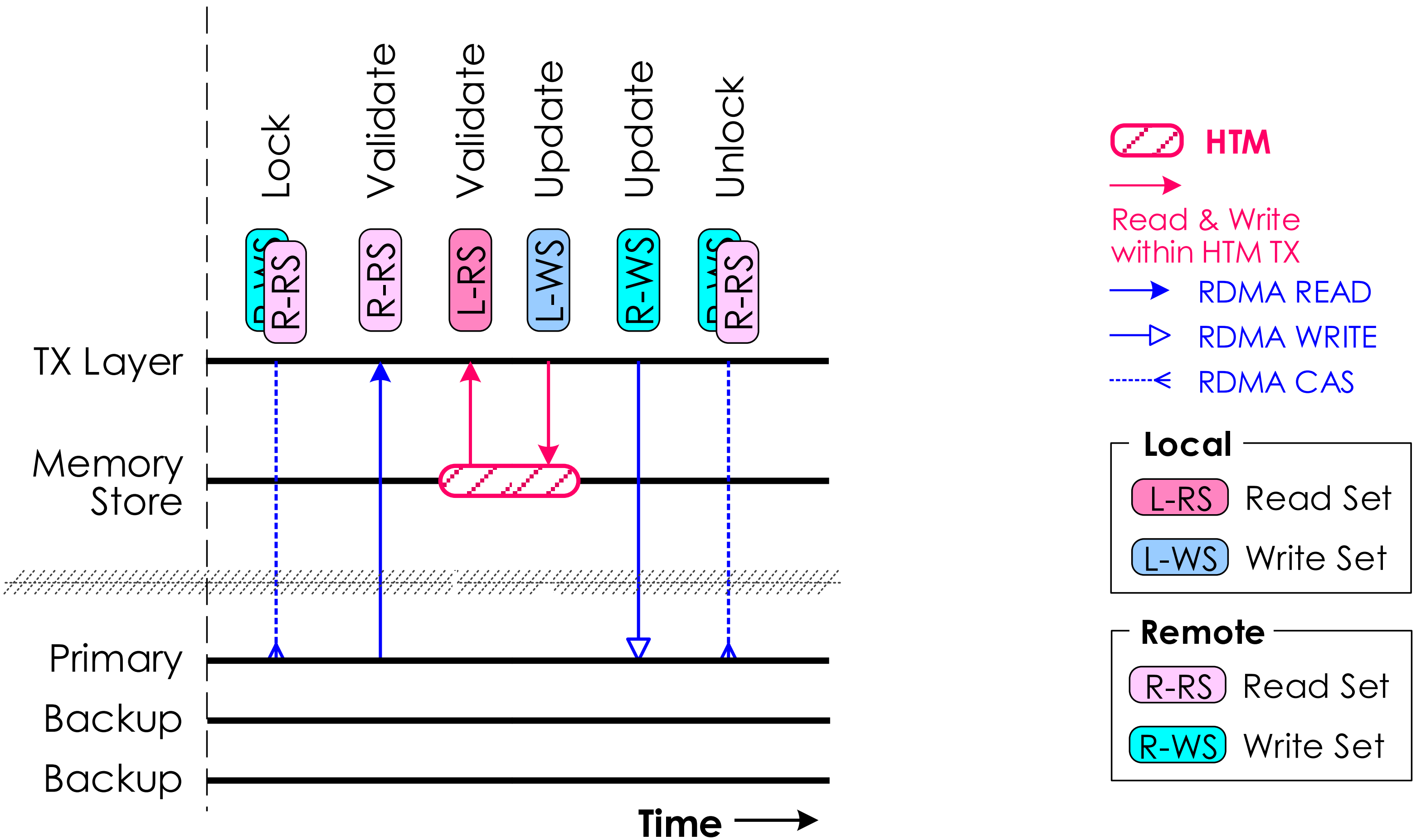
DrTM+R : Embrace performance & availability

- Hybrid OCC Protocol with HTM and RDMA
- Use Optimistic Replication with NVM to gain high availability

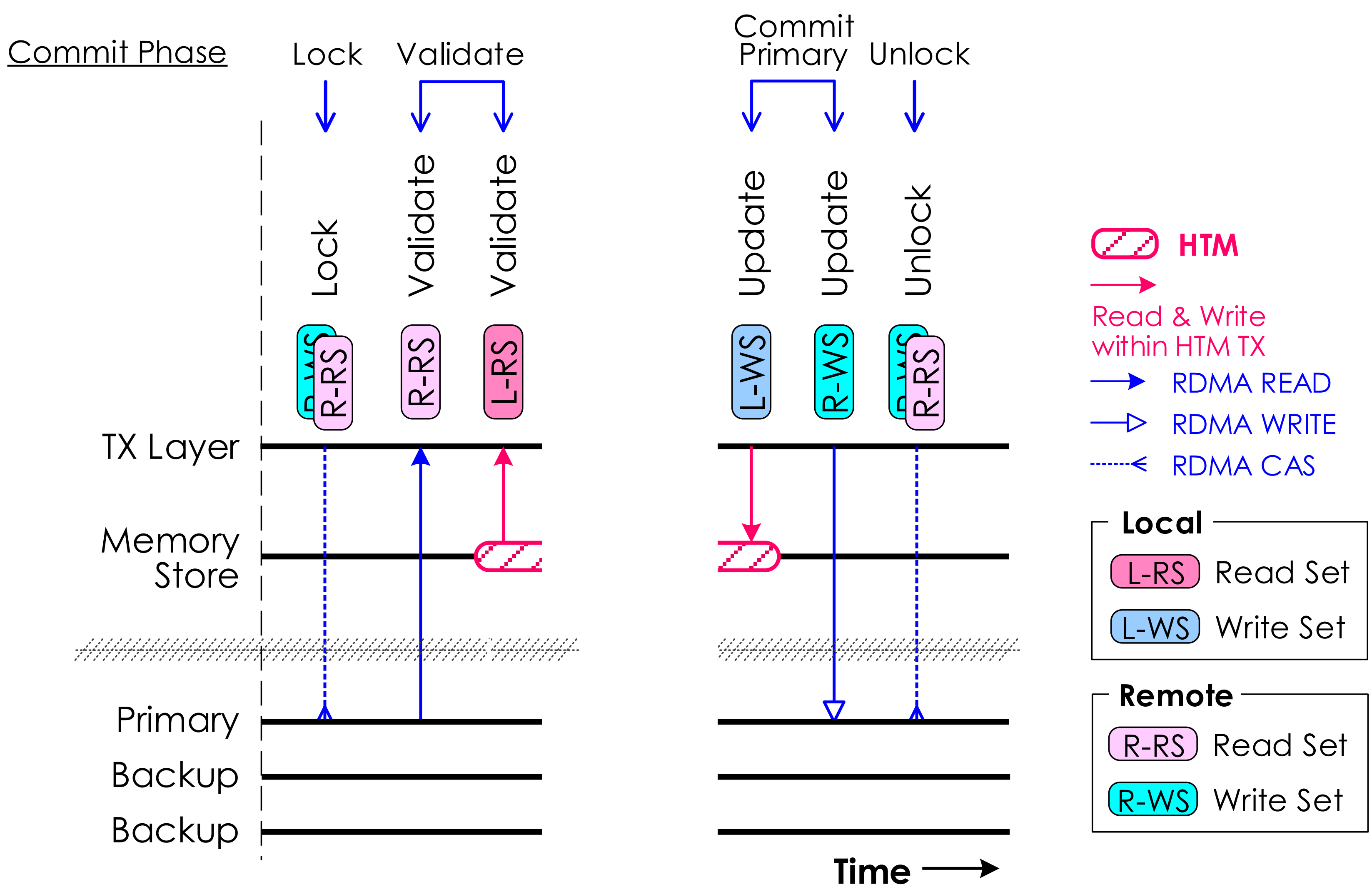


Primary-backup Replication

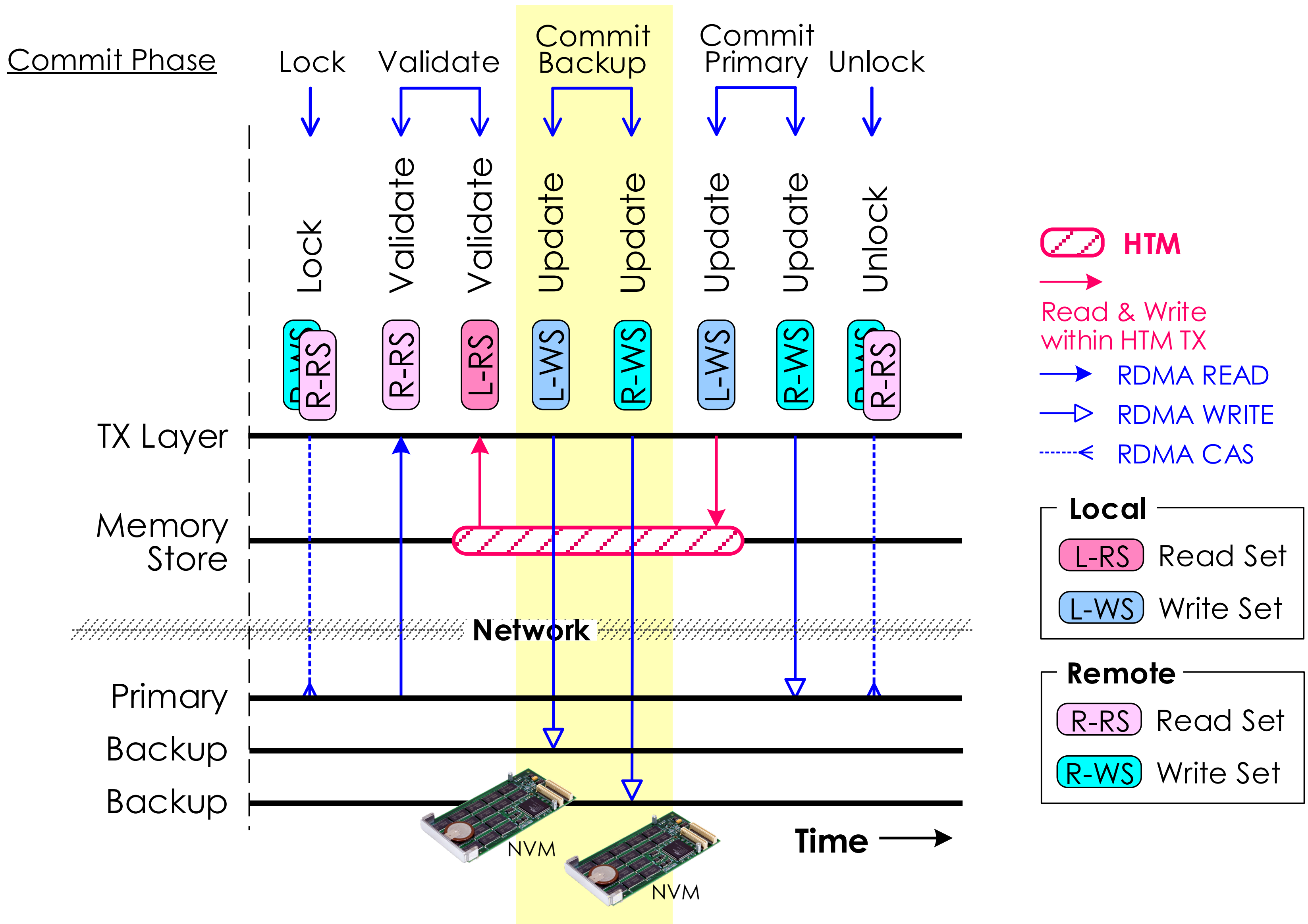
Commit Phase



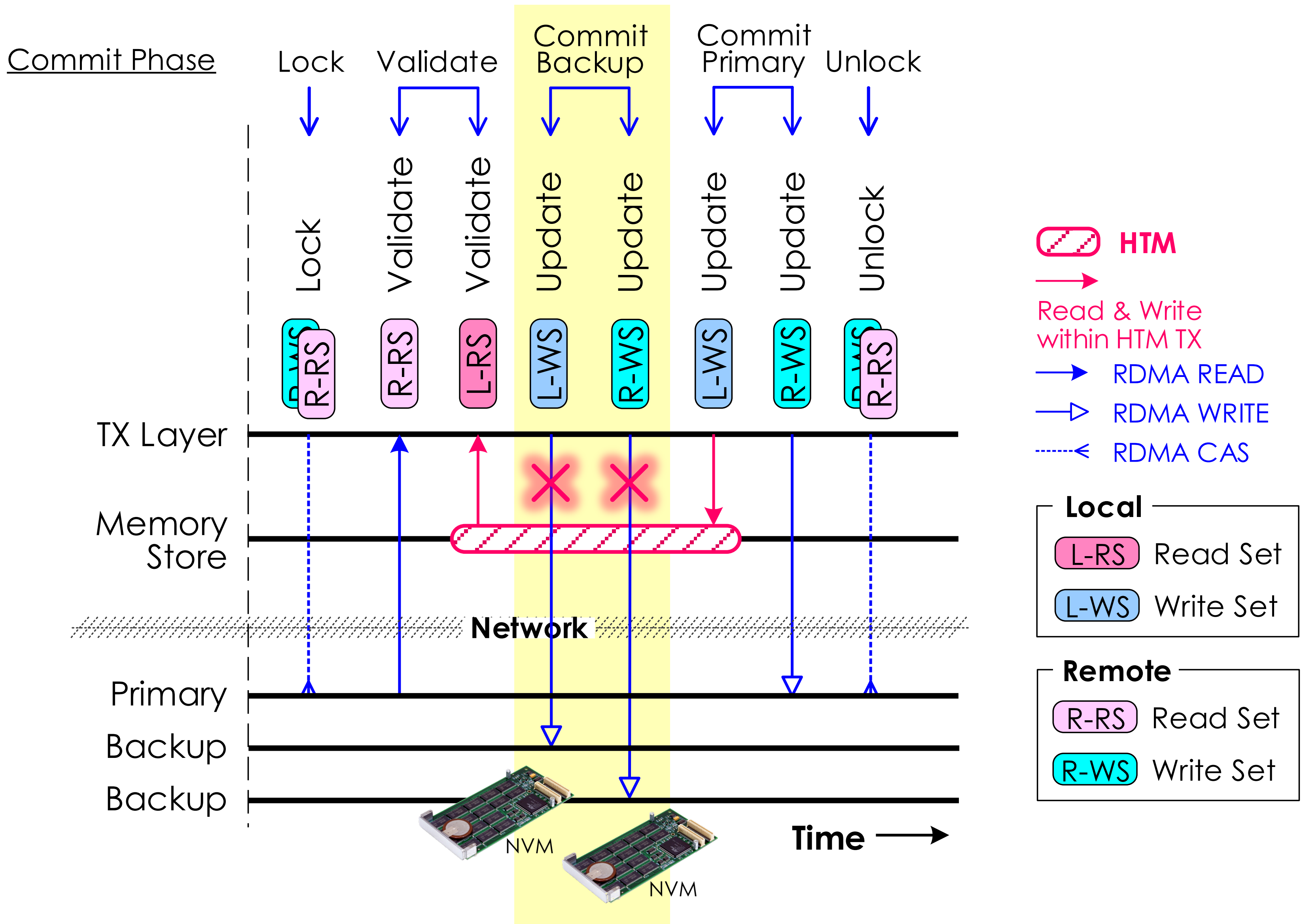
Primary-backup Replication



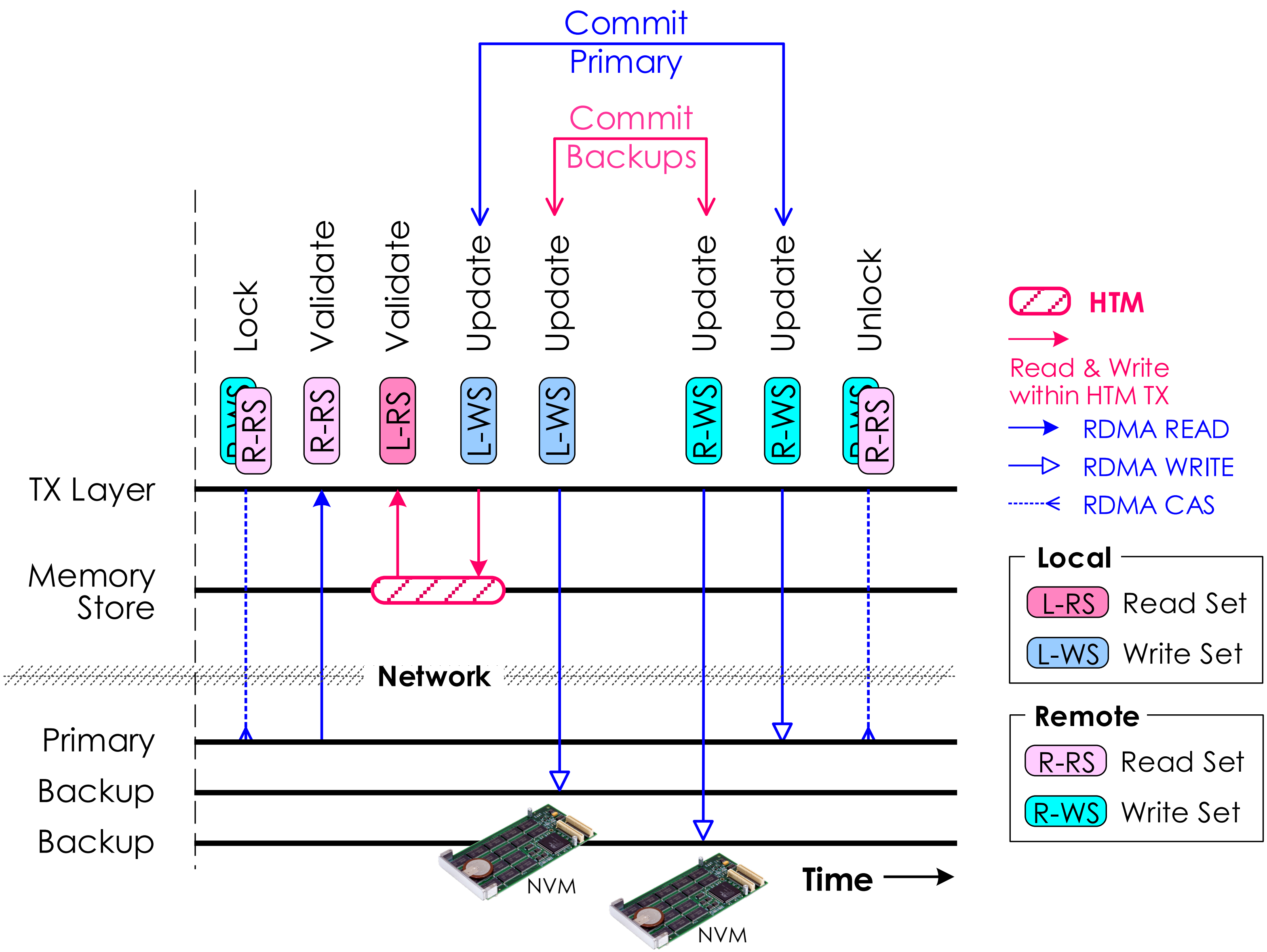
Primary-backup Replication



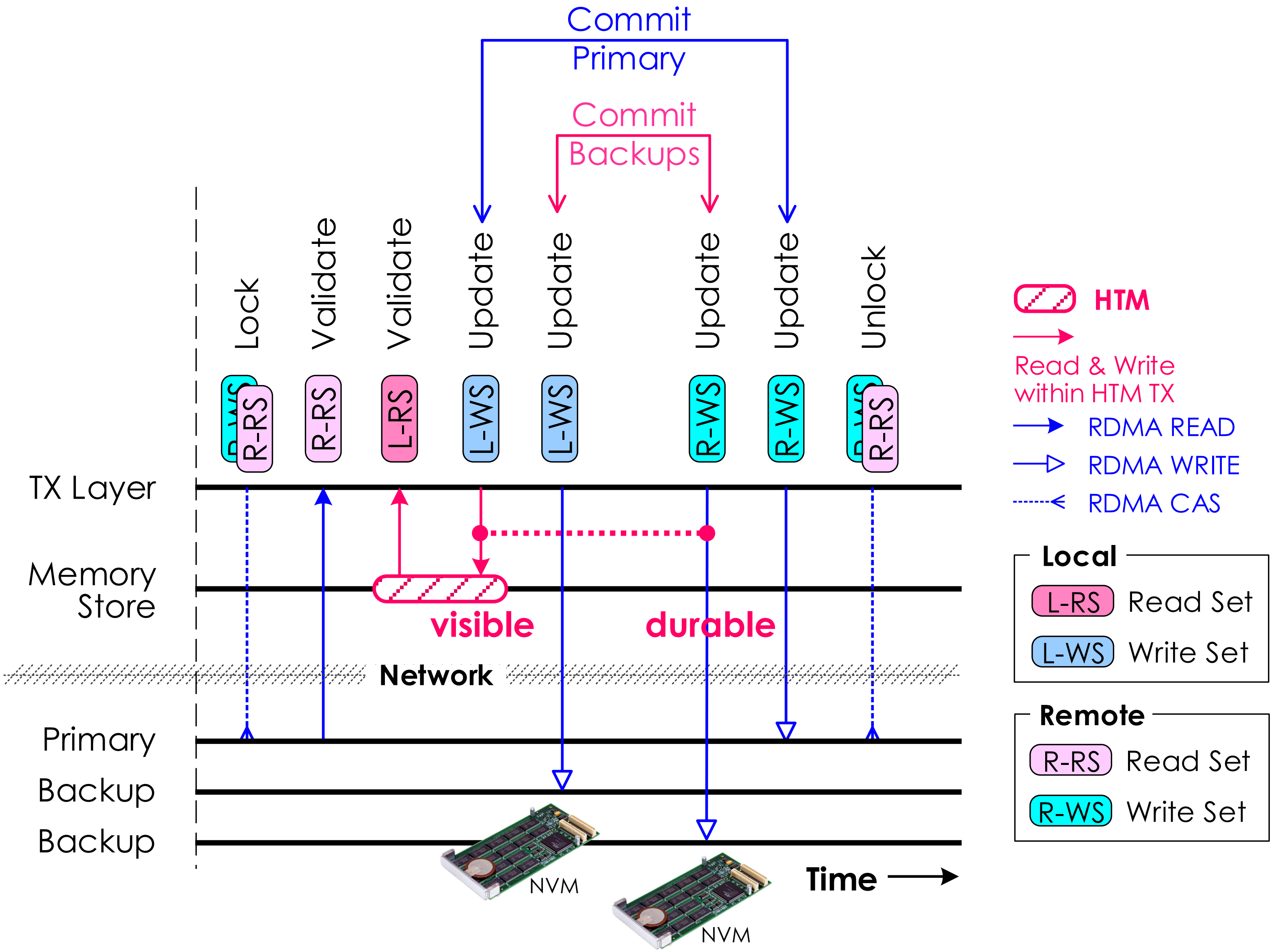
Challenge#1: Restriction of HTM



Move Replication outside the HTM



Challenge#2: Visibility of Commit



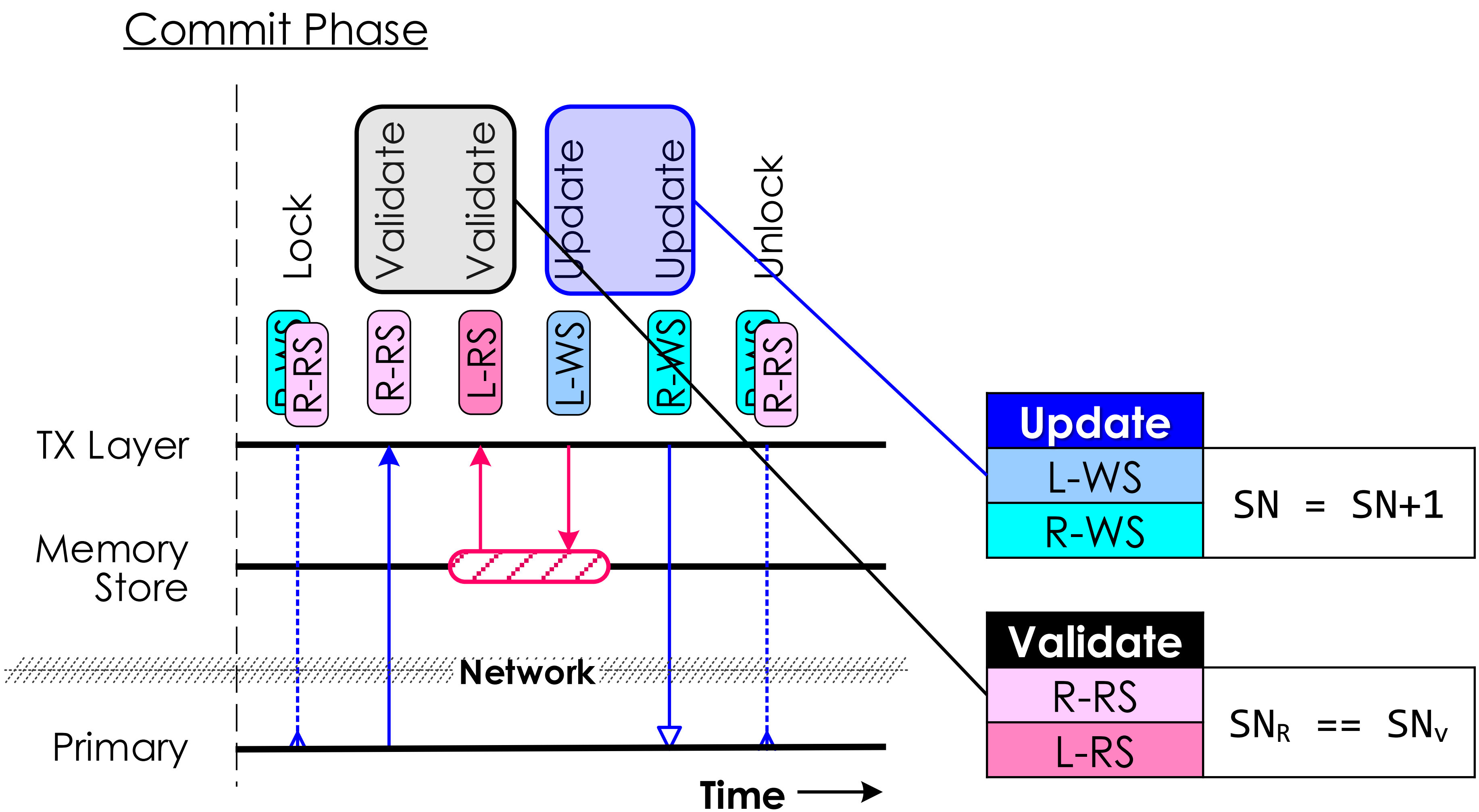
Optimistic Replication Scheme

Leave all locally written records in an uncommittable status after HTM commits

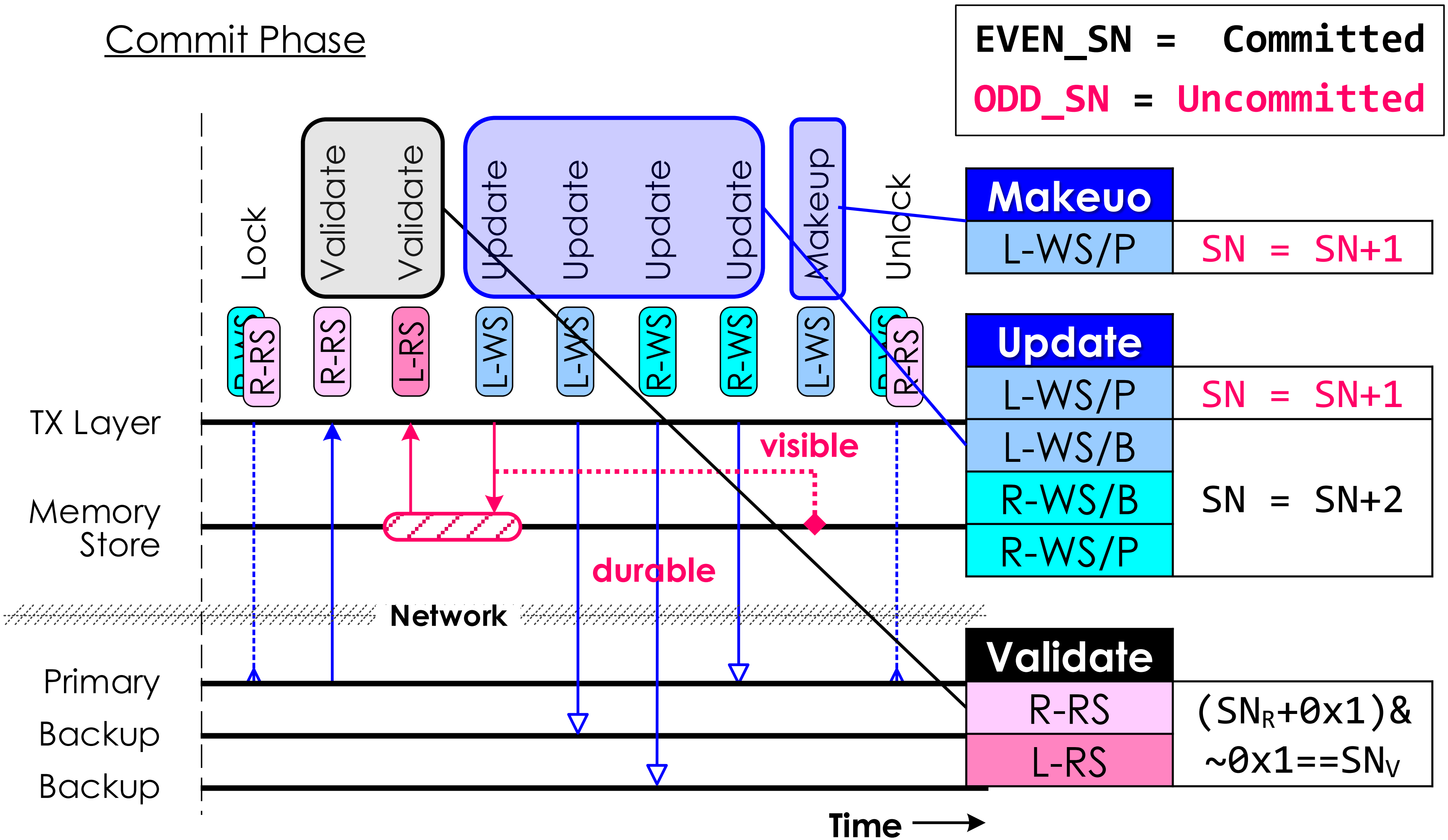
Transform to committable status until both primaries and backups of the written records have been applied

How to distinguish the **committable** status from **uncommittable** status ?

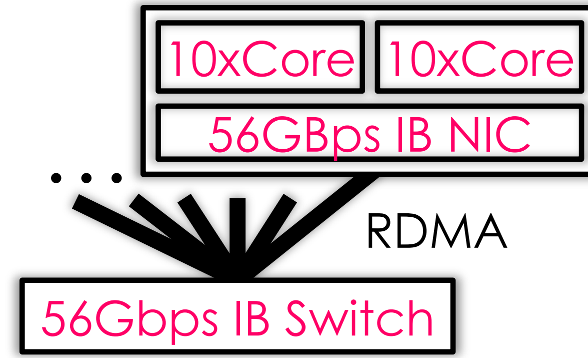
Leverage the Sequence Number (SN)



Leverage the Sequence Number (SN)



Evaluation



Baseline: *Calvin*

Platforms: A small-scale 6-machine cluster

- Each: two 10-cores, RTM-enabled Intel Xeon E5-2650 (disabled HT), 64GB DRAM, Mellanox ConnectX-3 MCX353A 56Gbps InfiniBand NIC w/ RDMA[1]

Benchmarks[2]

- TPC-C
- SmallBank

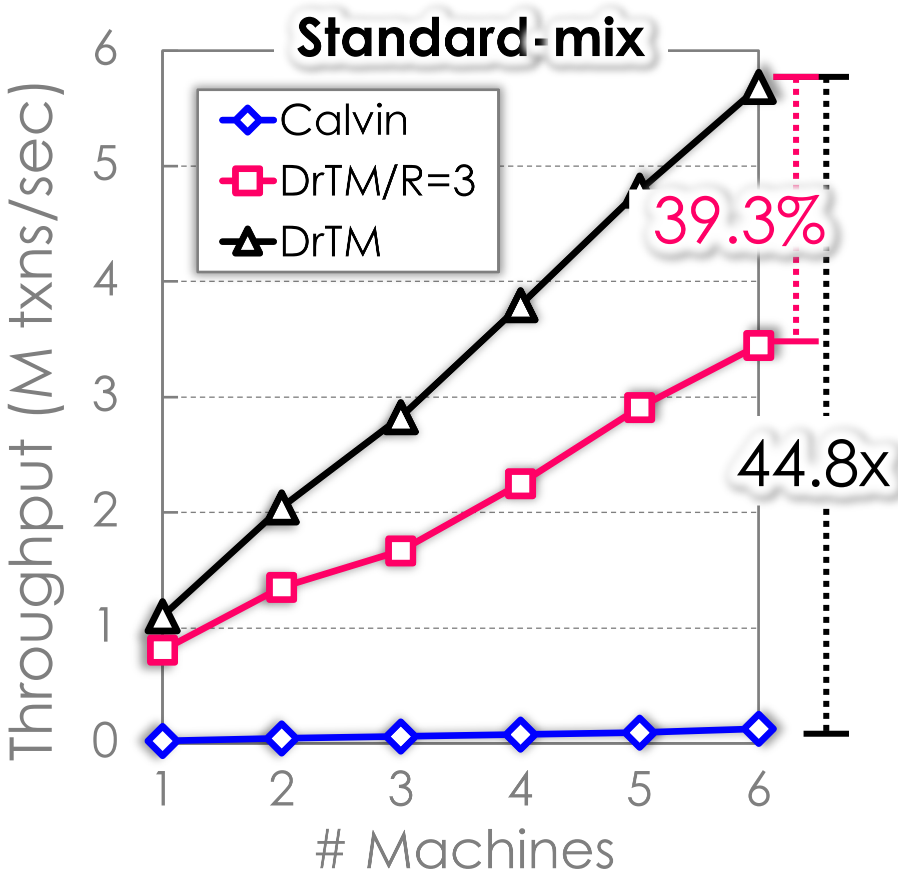
TPC-C	NEW	PAY	DLY	OS	SL
Ratio	45%	43%	4%	4%	4%
Type	d+rw	d+rw	l+rw	l+ro	l+ro

SmallBank	SP	AMG	BAL	DC	WC	TS
Ratio	25%	15%	15%	15%	15%	15%
Type	d+rw	d+rw	l+ro	l+rw	l+rw	l+rw

[1] All machines run Ubuntu 14.04 with Mellanox OFED v3.0-2.0.1 stack.
[2] **d** and **l** stand for distributed and local. **rw** and **ro** stand for read-write and read-only.

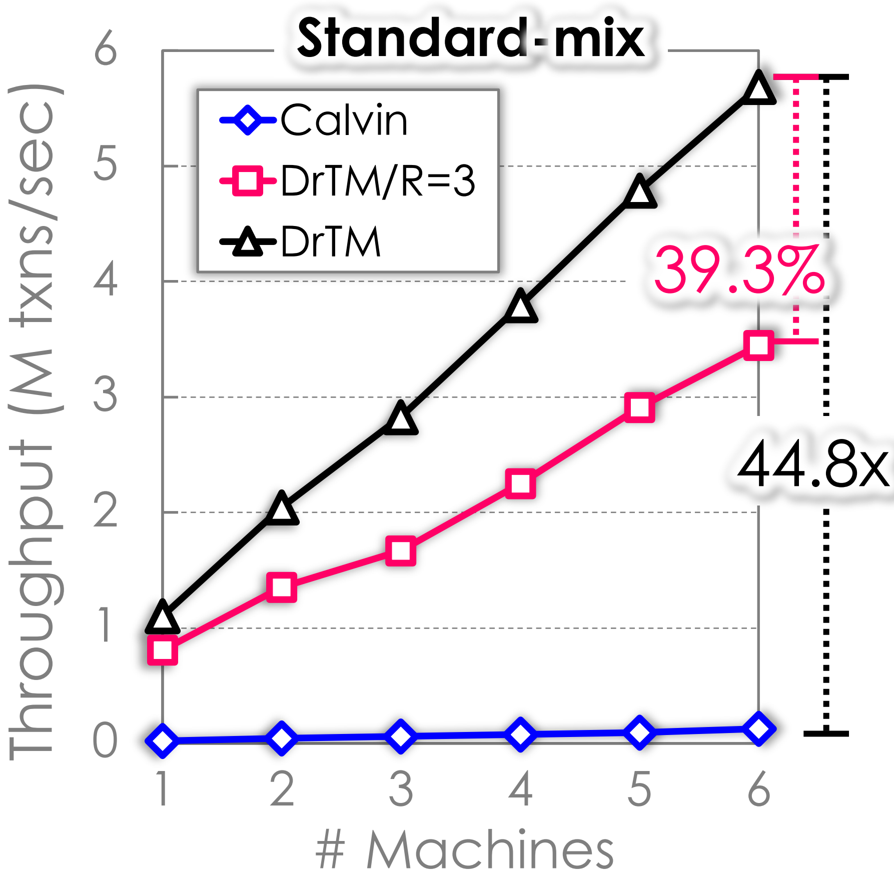
Performance on TPC-C

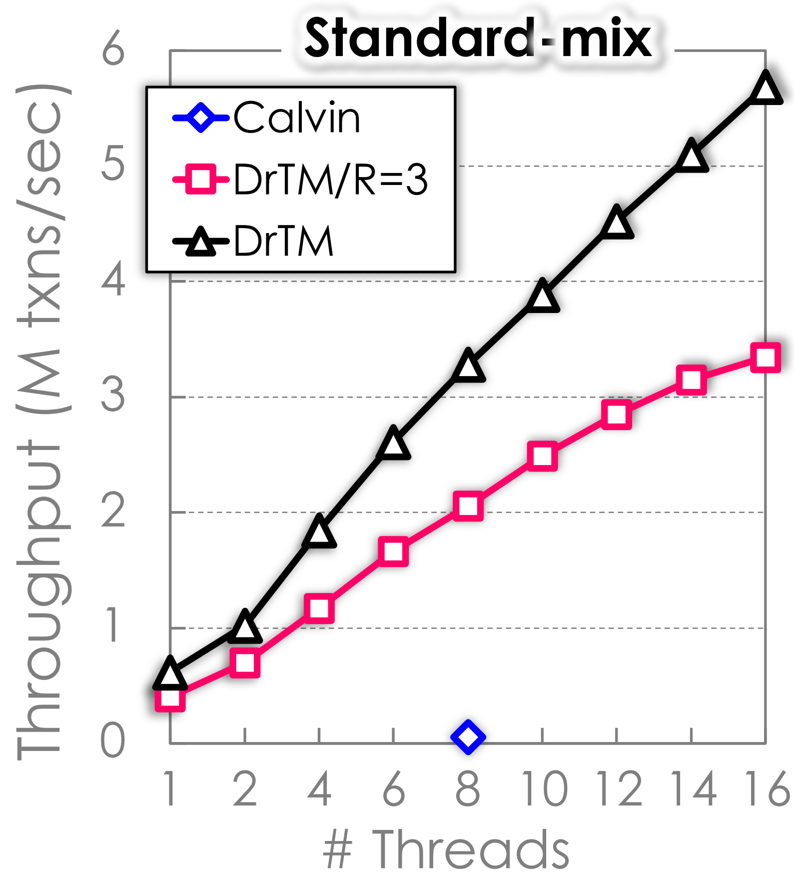
New-order TX
≈ Standard-mix x45%



Performance on TPC-C

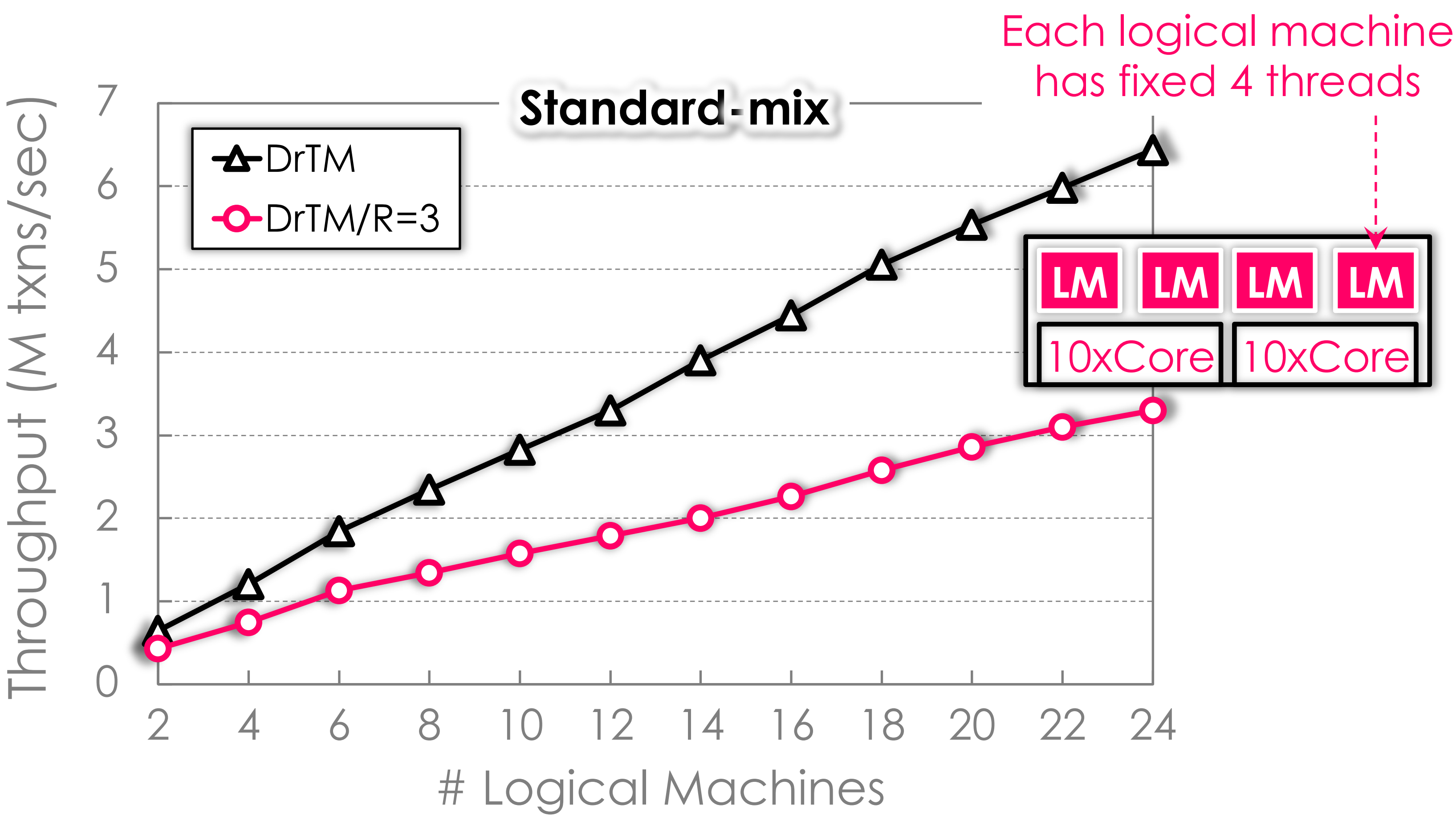
New-order TX
≈ Standard-mix x45%





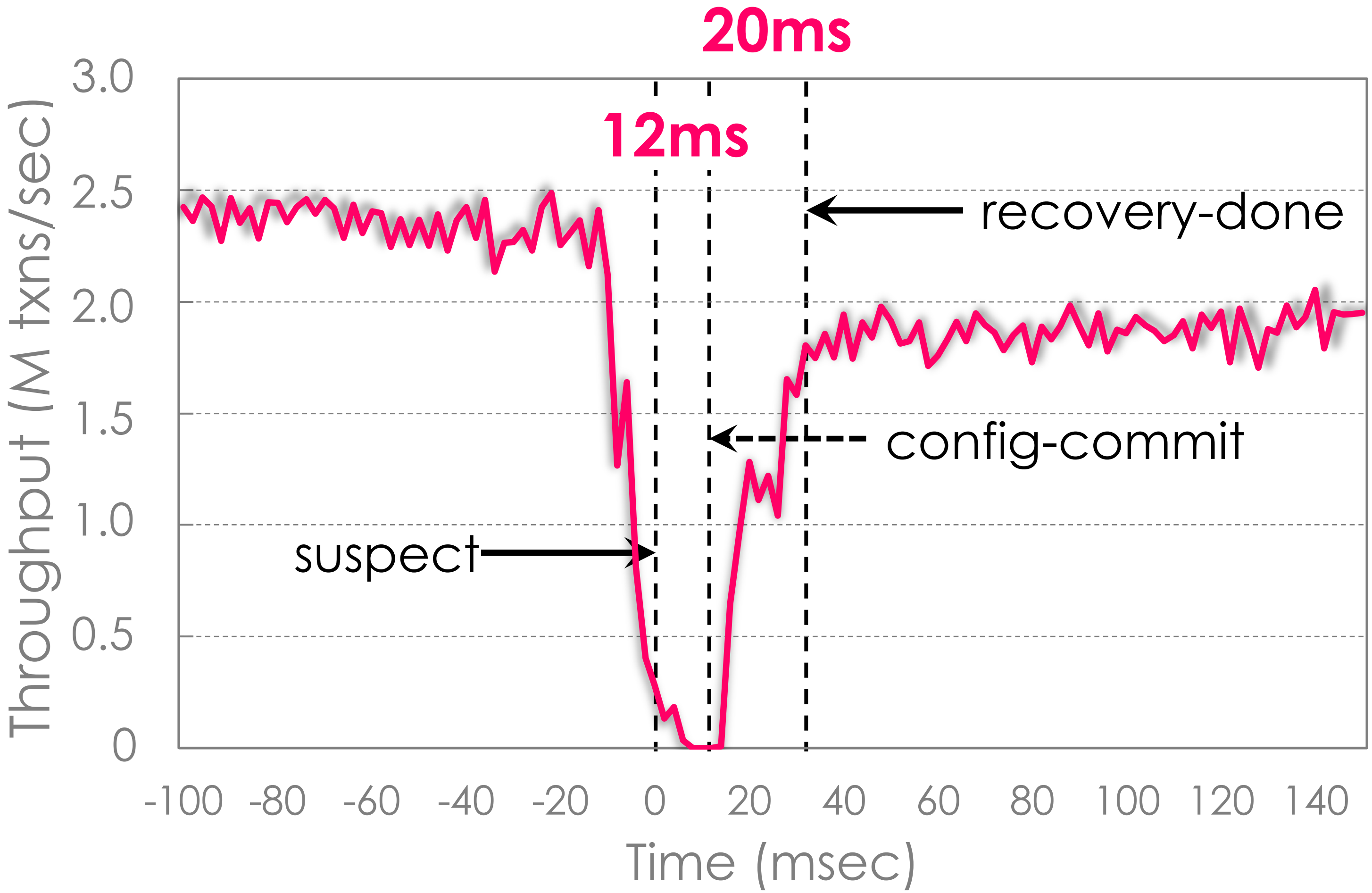
Scalability on TPC-C

New-order TX
≈ Standard-mix x45%



NOTE: the interaction btw. two logical nodes sharing the same machine still uses our RDMA-friendly OCC protocol

Recovery

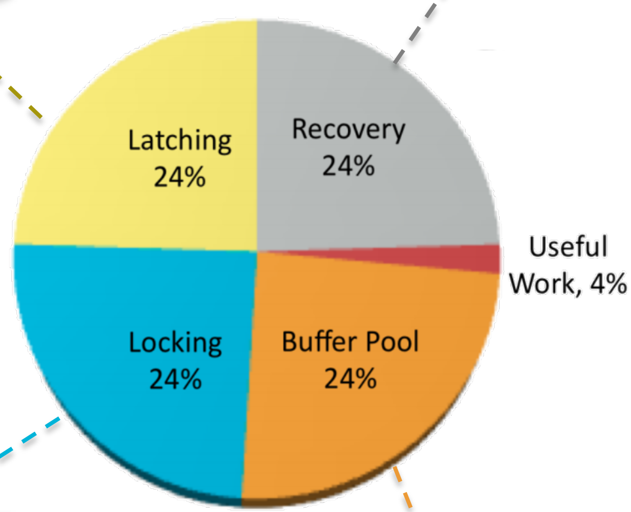


20ms
12ms
recovery-done
config-commit
suspect
Throughput (M txns/sec)
3.0
2.5
2.0
1.5
1.0
0.5
0
-100
-80
-60
-40
-20
0
20
40
60
80
100
120
140
Time (msec)

Revisit M. Stonebraker's View

In-memory Replication with NVM

Use HTM's ACI features



Michael Stonebraker
2014 Turing Award

One-sided RDMA Ops

In-memory Store

Conclusion

New hardware technologies open opportunities

DrTM : The first design and impl. of combining HTM, RDMA and NVM to boost in-memory transaction system with high availability

Achieving orders-of-magnitude higher throughput and lower latency than prior general designs

Thanks

DrTM

Questions

http://ipads.se.sjtu.edu.cn/drtm

http://ipads.se.sjtu.edu.cn/wukong

Institute of Parallel and Distributed Systems