

(Mar 29)

Lecture 18: Case Studies

Burrows

- update test suite
- ES3

Fri

- Practical benefits
- Solutions

Recap

Multi-Version Concurrency Control

- The DBMS maintains multiple physical versions of a single logical object in the database:
 - ▶ When a txn writes to an object, the DBMS creates a new version of that object.
 - ▶ When a txn reads an object, it reads the newest version that existed when the txn started.

Snapshot

Multi-Version Concurrency Control

- Writers don't block readers. Readers don't block writers.
- Read-only txns can read a consistent snapshot without acquiring locks or txn ids.
 - ▶ Use timestamps to determine visibility.
- Easily support time-travel queries.

Today's Agenda

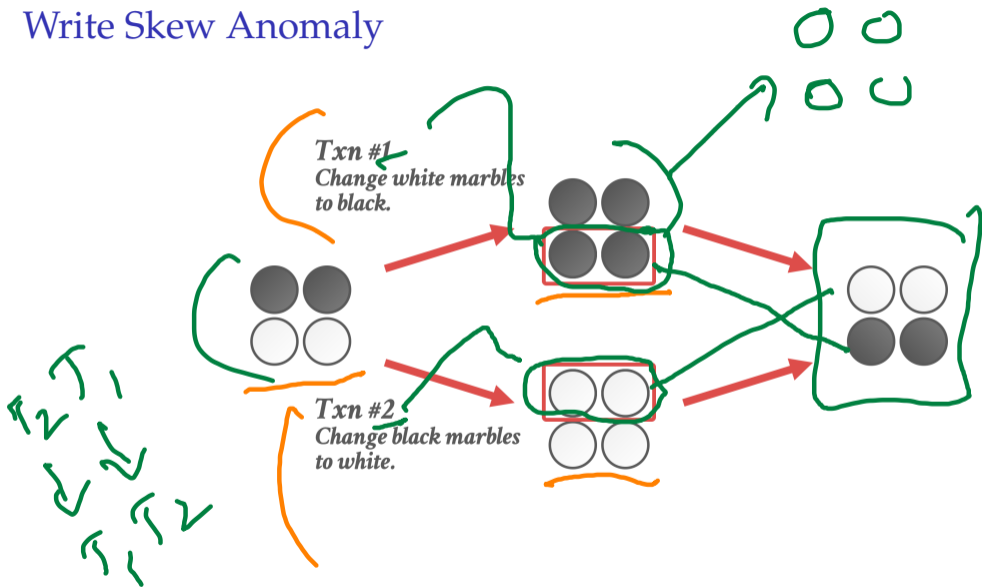
- MVCC Protocols
- Microsoft Hekaton (SQL Server)
- TUM HyPer ✓
- SAP HANA
- CMU Cicada ✓

MVCC Protocols

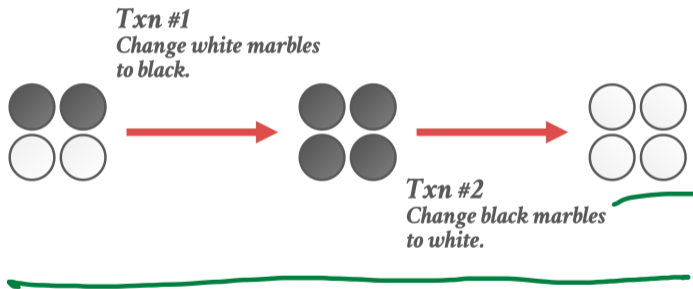
Snapshot Isolation (SI)

- When a txn starts, it sees a consistent snapshot of the database that existed when that the txn started.
 - ▶ No torn writes from active txns.
 - ▶ If two txns update the same object, then first writer wins.
- SI is susceptible to the Write Skew Anomaly.

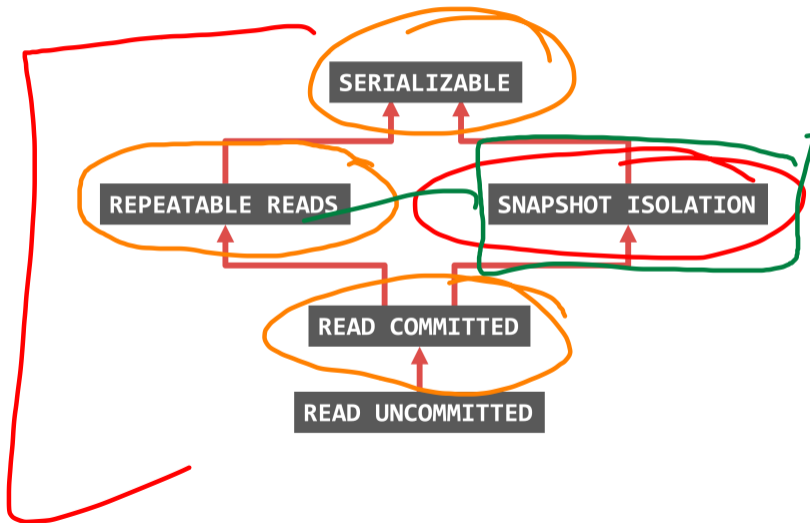
Write Skew Anomaly



Write Skew Anomaly



Isolation Level Hierarchy



Concurrency Control Protocol

MVTO

- Approach 1: Timestamp Ordering

- ▶ Assign txns timestamps that determine serial order.
- ▶ Considered to be original MVCC protocol.

- Approach 2: Optimistic Concurrency Control

- ▶ Three-phase protocol from last class.
- ▶ Use private workspace for new versions.

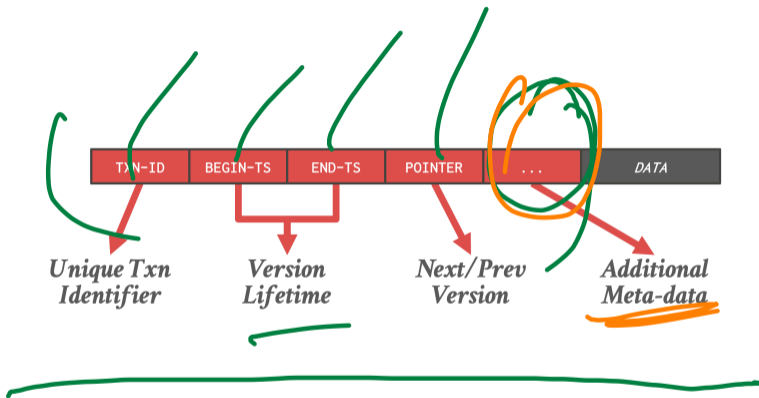
MVOC

- Approach 3: Two-Phase Locking

- ▶ Txns acquire appropriate lock on physical version before they can read/write a logical tuple.

MV2PL

Tuple Format

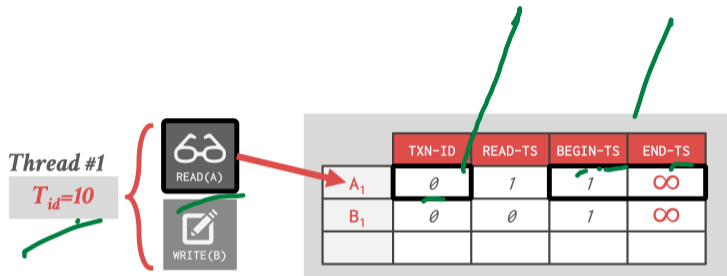


Timestamp Ordering (MVTO)

	TXN-ID	READ-TS	BEGIN-TS	END-TS
A ₁	0	1	1	∞
B ₁	0	0	1	∞

Use *read-ts* field in the header to keep track of the timestamp of the last txn that read it.

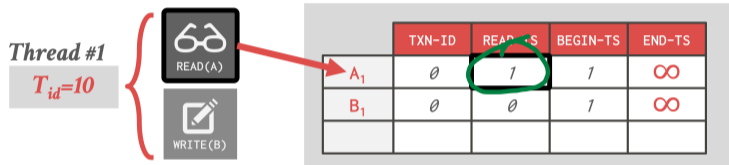
Timestamp Ordering (MVTO)



Use *read-ts* field in the header to keep track of the timestamp of the last txn that read it.

Txn can read version if the latch is unset and its T_{id} is between *begin-ts* and *end-ts*.

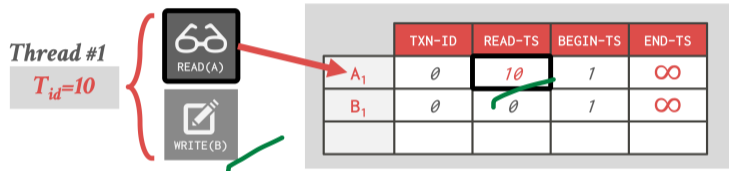
Timestamp Ordering (MVTO)



Use *read-ts* field in the header to keep track of the timestamp of the last txn that read it.

Txn can read version if the latch is unset and its T_{id} is between *begin-ts* and *end-ts*.

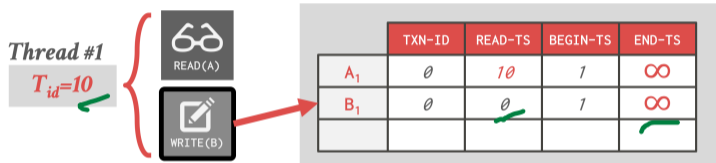
Timestamp Ordering (MVTO)



Use *read-ts* field in the header to keep track of the timestamp of the last txn that read it.

Txn can read version if the latch is unset and its T_{id} is between *begin-ts* and *end-ts*.

Timestamp Ordering (MVTO)



Use *read-ts* field in the header to keep track of the timestamp of the last txn that read it.

Txn can read version if the latch is unset and its T_{id} is between *begin-ts* and *end-ts*.

Txn creates a new version if no other txn holds latch and T_{id} is greater than *read-ts*.

Timestamp Ordering (MVTO)



Use *read-ts* field in the header to keep track of the timestamp of the last txn that read it.

	TXN-ID	READ-TS	BEGIN-TS	END-TS
A ₁	0	10	1	∞
B ₁	10	0	1	∞

A padlock icon is positioned to the left of the table. A green arrow points from the TXN-ID '10' in the B₁ row to the READ-TS '0' in the same row.

Txn can read version if the latch is unset and its T_{id} is between *begin-ts* and *end-ts*.

Txn creates a new version if no other txn holds latch and T_{id} is greater than *read-ts*.

Timestamp Ordering (MVTO)



Use *read-ts* field in the header to keep track of the timestamp of the last txn that read it.

	TXN-ID	READ-TS	BEGIN-TS	END-TS
A ₁	0	10	1	∞
B ₁	10	0	1	10
B ₂	10	0	10	∞

Txn can read version if the latch is unset and its T_{id} is between *begin-ts* and *end-ts*.

Txn creates a new version if no other txn holds latch and T_{id} is greater than *read-ts*.

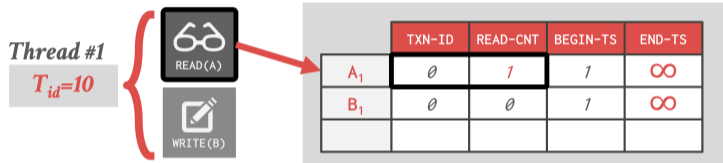
Two-Phase Locking (MV2PL)



Txns use the tuple's **read-cnt** field as SHARED lock.
Use **txn-id** and **read-cnt** together as EXCLUSIVE lock.

	TXN-ID	READ-CNT	BEGIN-TS	END-TS
A ₁	0	0	1	∞
B ₁	0	0	1	∞

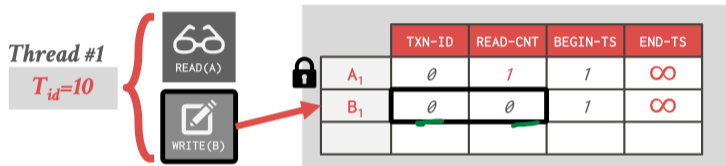
Two-Phase Locking (MV2PL)



Txns use the tuple's *read-cnt* field as SHARED lock. Use *txn-id* and *read-cnt* together as EXCLUSIVE lock.

If *txn-id* is zero, then the txn acquires the SHARED lock by incrementing the *read-cnt* field.

Two-Phase Locking (MV2PL)



Txns use the tuple's **read-cnt** field as SHARED lock. Use **txn-id** and **read-cnt** together as EXCLUSIVE lock.

If **txn-id** is zero, then the txn acquires the SHARED lock by incrementing the **read-cnt** field.

If both **txn-id** and **read-cnt** are zero, then txn acquires the EXCLUSIVE lock by setting both of them.

Two-Phase Locking (MV2PL)



Txns use the tuple's *read-cnt* field as SHARED lock. Use *txn-id* and *read-cnt* together as EXCLUSIVE lock.

	TXN-ID	READ-CNT	BEGIN-TS	END-TS
A ₁	0	1	1	∞
B ₁	10	1	1	∞

The table is shown with a lock icon on the left. The row for B₁ is highlighted with a thick black border, and the values '10' and '1' in its TXN-ID and READ-CNT columns are underlined in green.

If *txn-id* is zero, then the txn acquires the SHARED lock by incrementing the *read-cnt* field.

If both *txn-id* and *read-cnt* are zero, then txn acquires the EXCLUSIVE lock by setting both of them.

Two-Phase Locking (MV2PL)



Txns use the tuple's *read-cnt* field as SHARED lock. Use *txn-id* and *read-cnt* together as EXCLUSIVE lock.

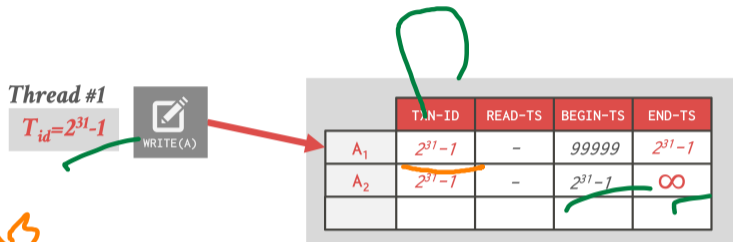
	TXN-ID	READ-CNT	BEGIN-TS	END-TS
A ₁	0	1	1	∞
B ₁	10	1	1	10
B ₂	10	0	10	∞

Two padlock icons are shown to the left of the table, indicating locks on the rows. Green checkmarks are present under the TXN-ID and READ-CNT fields of rows B₁ and B₂, and under the END-TS field of row B₁.

If *txn-id* is zero, then the txn acquires the SHARED lock by incrementing the *read-cnt* field.

If both *txn-id* and *read-cnt* are zero, then txn acquires the EXCLUSIVE lock by setting both of them.

Observation



If the DBMS reaches the max value for its timestamps, it will have to wrap around and restart at one. This will make all previous versions be in the "future" from new transactions.

32 bits

128 bits

1 B bytes

Observation

Thread #1

$T_{id}=2^{31}-1$



	TXN-ID	READ-TS	BEGIN-TS	END-TS
A_1	0	-	99999	$2^{31}-1$
A_2	0	-	$2^{31}-1$	∞

If the DBMS reaches the max value for its timestamps, it will have to wrap around and restart at one. This will make all previous versions be in the "future" from new transactions.

Observation

Thread #1

$T_{id}=2^{31}-1$

Thread #2

$T_{id}=1$



	TXN-ID	READ-TS	BEGIN-TS	END-TS
A ₁	0	-	99999	$2^{31}-1$
A ₂	1	-	$2^{31}-1$	1
A ₃	1	-	1	∞

If the DBMS reaches the max value for its timestamps, it will have to wrap around and restart at one. This will make all previous versions be in the "future" from new transactions.

Observation

Thread #1

$$T_{id}=2^{31}-1$$

Thread #2

$$T_{id}=1$$

	TXN-ID	READ-TS	BEGIN-TS	END-TS
A_1	\emptyset	-	99999	$2^{31}-1$
A_2	\emptyset	-	$2^{31}-1$	1
A_3	\emptyset	-	1	∞

If the DBMS reaches the max value for its timestamps, it will have to wrap around and restart at one. This will make all previous versions be in the "future" from new transactions.


PostgreSQL: Txn Id Wraparound

Version Visibility

- Set a flag in each tuple header that says that it is frozen in the past. Any new txn id will always be newer than a frozen version.
- Runs the vacuum before the system gets close to this upper limit.
- Otherwise it must stop accepting new commands when the system gets close to the max txn id.

GC

Microsoft Hekaton



Microsoft Hekaton

sybase

- Incubator project started in 2008 to create new OLTP engine for MSFT SQL Server (MSSQL).
 - ▶ Reference
- Had to integrate with MSSQL ecosystem.
- Had to support all possible OLTP workloads with predictable performance.
 - ▶ Single-threaded partitioning (e.g., H-Store/VoltDB) works well for some applications but terrible for others.

MV + OLTP + RPL

Hekaton MVCC

MVCC

- Each txn is assigned a timestamp when they begin (BeginTS) and when they commit (CommitTS).
- Each tuple contains two timestamps that represents their visibility and current state:
 - ▶ BEGIN-TS: The BeginTS of the active txn or the CommitTS of the committed txn that created it.
 - ▶ END-TS: The BeginTS of the active txn that created the next version or infinity or the CommitTS of the committed txn that created it.

Hekaton: Operations

Thread #1

Begin @ 25



READ(A)

Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
A ₁	10	20	\$100	●
A ₂	20	∞	\$200	∅

Hekaton: Operations

Thread #1

Begin @ 25



READ(A)

Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
A ₁	10	20	\$100	●
A ₂	20	∞	\$200	∅

Hekaton: Operations

Thread #1

Begin @ 25



Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
A ₁	10	20	\$100	●
A ₂	20	∞	\$200	∅
A ₃	Txn@25	∞	\$300	

Txn@25
 → 10000000...00000000 00011001

64-bit
 98-bit

not committed

Hekaton: Operations

Thread #1

Begin @ 25



Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
<i>A₁</i>	<i>10</i>	<i>20</i>	\$100	
<i>A₂</i>	<i>20</i>	∞	\$200	\emptyset
<i>A₃</i>	<i>Txn@25</i>	∞	\$300	

Hekaton: Operations

Thread #1

Begin @ 25



READ(A)



WRITE(A)

Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
<i>A₁</i>	<i>10</i>	<i>20</i>	<i>\$100</i>	
<i>A₂</i>	<i>20</i>	<i>Txn@25</i>	<i>\$200</i>	
<i>A₃</i>	<i>Txn@25</i>	<i>∞</i>	<i>\$300</i>	

Hekaton: Operations

Thread #1

Begin @ 25

Commit @ 35



READ(A)



WRITE(A)

Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
A ₁	10	20	\$100	●
A ₂	20	Txn@25	\$200	●
A ₃	Txn@25	∞	\$300	

Diagram illustrating the Main Data Table structure and thread operations. The table has columns: BEGIN-TS, END-TS, VALUE, and POINTER. The rows represent data items A₁, A₂, and A₃. The BEGIN-TS and END-TS columns are highlighted in red. The END-TS for A₂ and the BEGIN-TS for A₃ are both Txn@25. The END-TS for A₃ is ∞. The POINTER column contains red dots with arrows pointing to the right. Orange arrows point from the text 'Begin @ 25' and 'Commit @ 35' to the BEGIN-TS and END-TS columns respectively. Another orange arrow points from 'Txn@25' in the END-TS column of A₂ to the BEGIN-TS column of A₃.

Hekaton: Operations

Thread #1

Begin @ 25

Commit @ 35



READ(A)



WRITE(A)

Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
A ₁	10	20	\$100	●
A ₂	20	35	\$200	●
A ₃	35	∞	\$300	

Hekaton: Operations

Thread #1

Begin @ 25

Commit @ 35



READ(A)



WRITE(A)

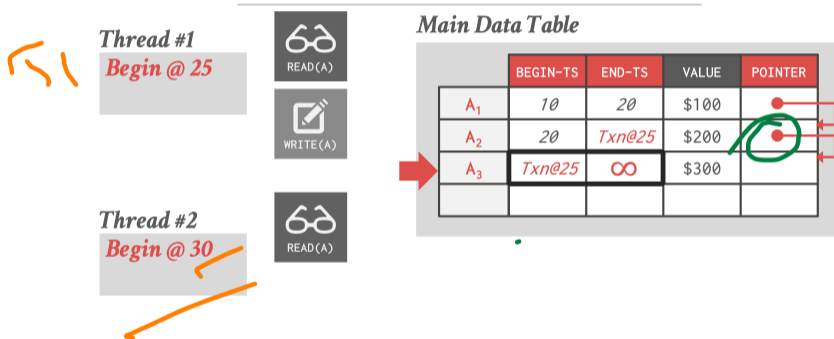
Main Data Table

	BEGIN-TS	END-TS	VALUE	POINTER
A ₁	10	20	\$100	●
A ₂	20	35	\$200	●
A ₃	35	∞	\$300	

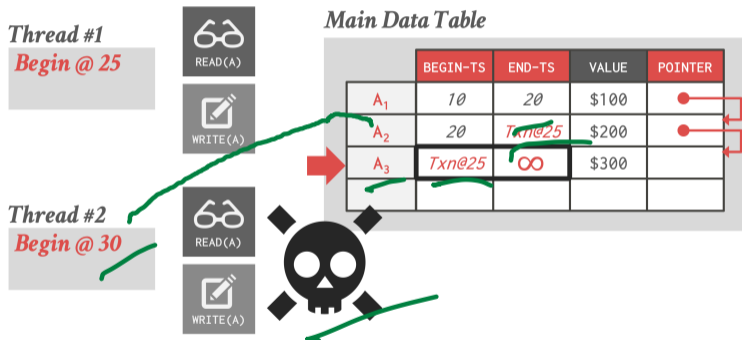


REWIND

Hekaton: Operations



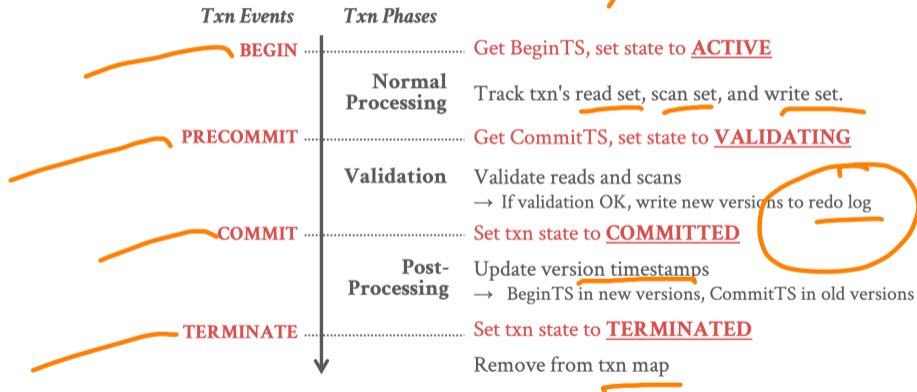
Hekaton: Operations



Hekaton: Transaction State Map

- Global map of all txns' states in the system:
 - ▶ **ACTIVE**: The txn is executing read/write operations.
 - ▶ **VALIDATING**: The txn has invoked commit and the DBMS is checking whether it is valid.
 - ▶ **COMMITTED**: The txn is finished but may have not updated its versions' TS.
 - ▶ **TERMINATED**: The txn has updated the TS for all of the versions that it created.

Hekaton: Transaction Lifecycle



Hekaton: Transaction Meta-Data

OLTP

- Read Set

- ▶ Pointers to physical versions returned to access method.

- Write Set

- ▶ Pointers to versions updated (old and new), versions deleted (old), and version inserted (new).

- Scan Set

- ▶ Stores enough information needed to perform each scan operation again to check result.

- Commit Dependencies

- ▶ List of txns that are waiting for this txn to finish.

primary

secondary

Hekaton: Transaction Validation

- Read Stability

- ▶ Check that each version read is still visible as of the end of the txn.

- Phantom Avoidance

- ▶ Repeat each scan to check whether new versions have become visible since the txn began.

- Extent of validation depends on isolation level:

- ▶ SERIALIZABLE: Read Stability + Phantom Avoidance
- ▶ REPEATABLE READS: Read Stability
- ▶ SNAPSHOT ISOLATION: None
- ▶ READ COMMITTED: None

Hekaton: Optimistic vs. Pessimistic

- **Optimistic Txns:**

- ▶ Check whether a version read is still visible at the end of the txn.
- ▶ Repeat all index scans to check for phantoms.

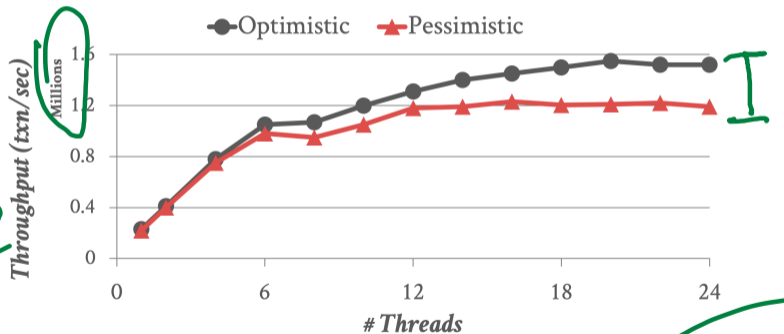
- **Pessimistic Txns:**

- ▶ Use shared & exclusive locks on records and buckets.
- ▶ No validation is needed.
- ▶ Separate background thread to detect deadlocks.

MVCC

Hekaton: Optimistic vs. Pessimistic

Database: Single table with 1000 tuples
Workload: 80% read-only txns + 20% update txns
Processor: 2 sockets, 12 cores



933,000

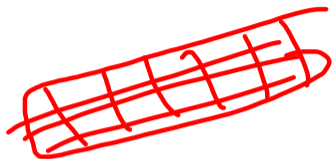
Hekaton: Lessons

- Use only lock-free data structures
 - ▶ No latches, spin locks, or critical sections
 - ▶ Indexes, txn map, memory alloc, garbage collector
 - ▶ Example: Bw-Trees
- Only one single serialization point in the DBMS to get the txn's begin and commit timestamp
 - ▶ Atomic Addition (CAS)

CAS

latch-free B+Tree

Observation



OLTP (OLAP, HTAP)

- Read/scan set validations are expensive if the txns access a lot of data.
- Appending new versions hurts the performance of OLAP scans due to pointer chasing & branching.
- Record-level conflict checks may be too coarse-grained and incur false positives.



Begin, End

Emp: 20 attributes
1000 attributes

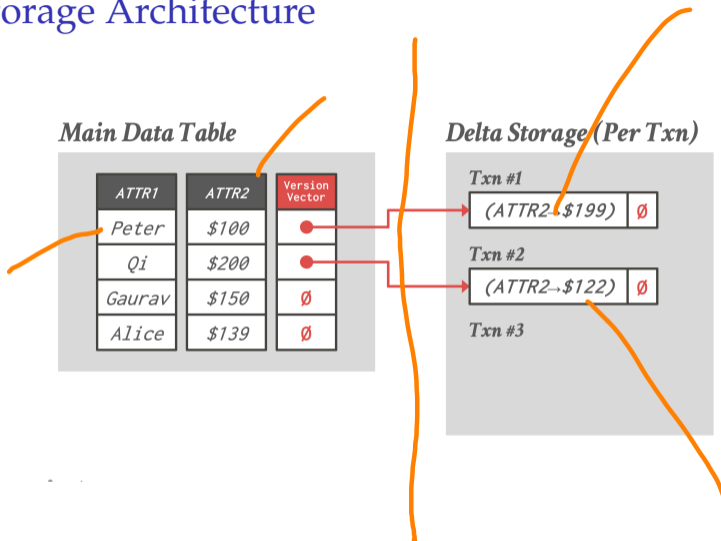
Hyper

TU Munich

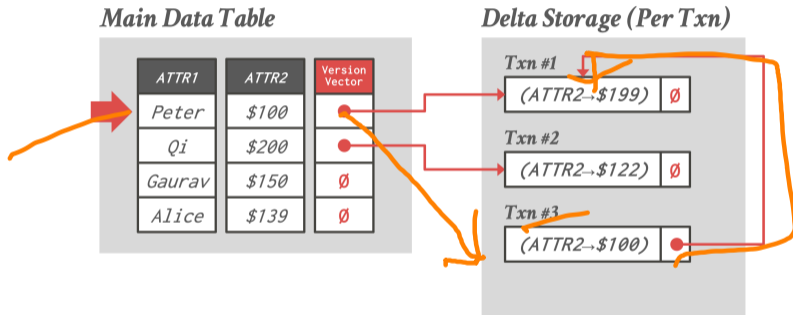
Hyper MVCC

- Column store with delta record versioning.
 - Reference
 - ▶ In-Place updates for non-indexed attributes
 - ▶ Delete/Insert updates for indexed attributes.
 - ▶ Newest-to-Oldest Version Chains
 - ▶ No Predicate Locks / No Scan Checks
 - Avoids write-write conflicts by aborting txns that try to update an uncommitted object.
- Previsibly Locking*

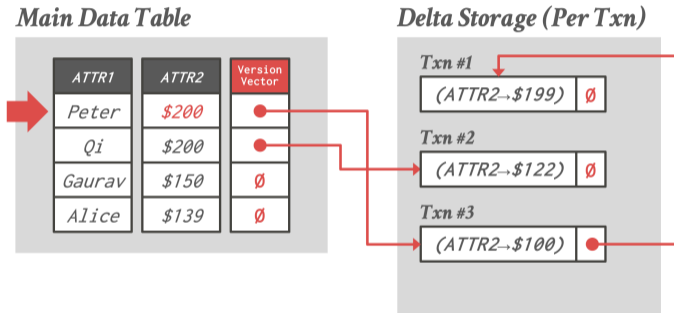
Hyper: Storage Architecture



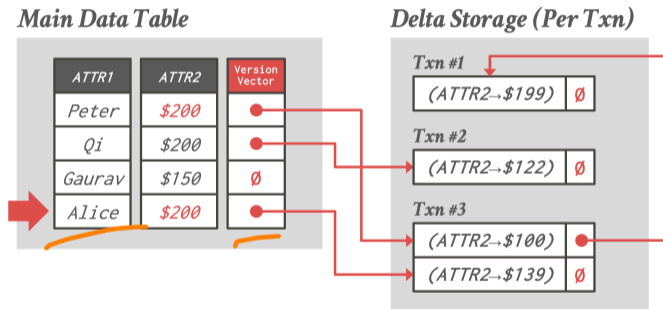
Hyper: Storage Architecture



Hyper: Storage Architecture



Hyper: Storage Architecture



Hyper: Validation

- First-Writer Wins
 - ▶ If version vector is not null, then it always points to the last committed version.
 - ▶ Do not need to check whether write-sets overlap.
- Check the redo buffers of txns that committed after the validating txn started.
 - ▶ Compare the committed txn's write set for phantoms using Precision Locking.
 - ▶ Only need to store the txn's read predicates and not its entire read set.

NZO

Hyper: Precision Locking



Validating Txn

```
SELECT * FROM foo
WHERE attr2 > 20
  AND attr2 < 30
```

```
SELECT COUNT(attr1)
FROM foo
WHERE attr2 IN (10,20,30)
```

```
SELECT attr1, AVG(attr2)
FROM foo
WHERE attr1 LIKE '%Ice%'
GROUP BY attr1
HAVING AVG(attr2) > 100
```

attr2
E [20,30]

99 > 20 AND 99 < 30

FALSE

33 > 20 AND 33 < 30

Delta Storage (Per Txn)

Txn #1001

(attr2 → 99)

(attr2 → 33)

Txn #1002

(attr2 → 122)

Txn #1003

(attr1 → 'IceCube',
attr2 → 199)

Hyper: Precision Locking

Validating Txn

```
SELECT * FROM foo
WHERE attr2 > 20
AND attr2 < 30
```

```
SELECT COUNT(attr1)
FROM foo
WHERE attr2 IN (10,20,30)
```

```
SELECT attr1, AVG(attr2)
FROM foo
WHERE attr1 LIKE 'Ice%'
GROUP BY attr1
HAVING AVG(attr2) > 100
```

Delta Storage (Per Txn)

Txn #1001

(attr2→99)

(attr2→33)

Txn #1002

(attr2→122)

Txn #1003

(attr1→'IceCube',
attr2→199)

99 IN (10,20,30)
33 IN (10,20,30)

FALSE

Hyper: Precision Locking

Validating Txn

```
SELECT * FROM foo
WHERE attr2 > 20
AND attr2 < 30
```

```
SELECT COUNT(attr1)
FROM foo
WHERE attr2 IN (10,20,30)
```

```
SELECT attr1, AVG(attr2)
FROM foo
WHERE attr1 LIKE '%Ice%'
GROUP BY attr1
HAVING AVG(attr2) > 100
```

Delta Storage (Per Txn)

Txn #1001

(attr2→99)

(attr2→33)

Txn #1002

(attr2→122)

Txn #1003

(attr1→'IceCube',
attr2→199)

NULL LIKE '%Ice%'
NULL LIKE '%Ice%'

FALSE

Hyper: Precision Locking

19705

Validating Txn

```
SELECT * FROM foo
WHERE attr2 > 20
AND attr2 < 30
```

```
SELECT COUNT(attr1)
FROM foo
WHERE attr2 IN (10,20,30)
```

```
SELECT attr1, AVG(attr2)
FROM foo
WHERE attr1 LIKE '%Ice%'
GROUP BY attr1
HAVING AVG(attr2) > 100
```



Delta Storage (Per Txn)

Txn #1001

(attr2→99)

(attr2→33)

Txn #1002

(attr2→122)

Txn #1003

(attr1→'IceCube',
attr2→199)

TRUE

'IceCube' LIKE '%Ice%'

Logical

Hyper: Version Synopses

OLAP + Column Store

Main Data Table

Version Synopsis	ATTR1	ATTR2	Version Vector
(2, 5)	0 Peter	\$100	∅
	1 Qi	\$200	∅
	2 Gaurav	\$150	• →
	3 Alice	\$99	∅
	4 Mark	\$300	• →
	5 Rahul	\$300	∅
	6 Alex	\$0	∅

Offsets

Store a separate column that tracks the position of the first and last versioned tuple in a block of tuples.

When scanning tuples, the DBMS can check for strides of tuples without older versions and execute more efficiently.

Hyper: Version Synopses

Main Data Table

Zone *pred*

Version Synopsis	ATTR1	ATTR2	Version Vector
[2, 5)	Peter	\$100	∅
	Qi	\$200	∅
	Gaurav	\$150	• →
	Alice	\$99	∅
	Mark	\$300	• →
	Rahul	\$300	∅
	Alex	\$0	∅

Zone *pred*

Store a separate column that tracks the position of the first and last versioned tuple in a block of tuples.

When scanning tuples, the DBMS can check for strides of tuples without older versions and execute more efficiently.

SAP HANA



SAP HANA

- In-memory HTAP DBMS with time-travel version storage (N2O).
- Reference
 - ▶ Supports both optimistic and pessimistic MVCC.
 - ▶ Latest versions are stored in time-travel space.
 - ▶ Hybrid storage layout (row + columnar).
- Based on P*TIME, TREX, and MaxDB.
- First released in 2012.

ERP

SAP HANA: Version Storage

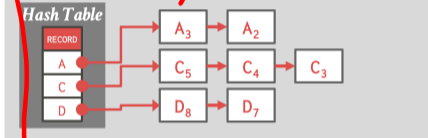
- Store the oldest version in the main data table.
- Each tuple maintains a flag to denote whether there exists newer versions in the version space.
- Maintain a separate hash table that maps record identifiers to the head of version chain.

SAP HANA: Version Storage

Main Data Table

RID	VERS?	VERSION	DATA
A	True	A ₁	-
B	False	B ₃	-
C	True	C ₂	-
D	True	D ₆	-

Version Storage



SAP HANA: Transactions

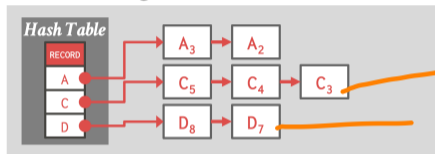
- Instead of embedding meta-data about the txn that created a version with the data, store a pointer to a context object.
 - ▶ Reads are slower because you must follow pointers.
- ▶ Large updates are faster because it's a single write to update the status of all tuples.
- Store meta-data about whether a txn has committed in a separate object as well.

SAP HANA: Version Storage

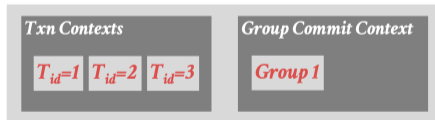
Main Data Table

RID	VERS?	VERSION	DATA
A	True	A ₁	-
B	False	B ₃	-
C	True	C ₂	-
D	True	D ₆	-

Version Storage



Txn Meta-Data



Thread #1

T_{id} = 3



SAP HANA: Version Storage

Main Data Table

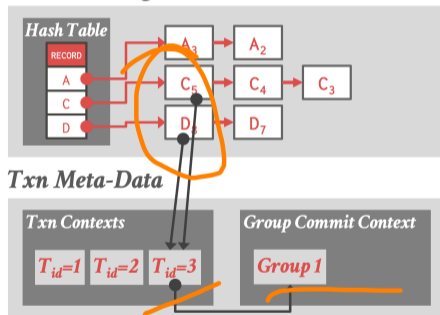
RID	VERS?	VERSION	DATA
A	True	A ₁	-
B	False	B ₃	-
C	True	C ₂	-
D	True	D ₆	-

Thread #1

T_{id} = 3



Version Storage



Cicada — Cmd

MVCC Limitations

- Computation & Storage Overhead

- ▶ Most MVCC schemes use indirection to search a tuple's version chain. This increases CPU cache misses.
- ▶ Also requires frequent garbage collection to minimize the number versions that a thread must evaluate.

- Shared Memory Writes

- ▶ Most MVCC schemes store versions in "global" memory in the heap without considering locality.

- Timestamp Allocation

- ▶ All threads access single shared counter.

OCC Limitations

MVCC

- Frequent Aborts

- ▶ Txns will abort too quickly under high contention, causing high churn.

- Extra Reads & Writes

- ▶ Each txn must copy tuples into their private workspace to ensure repeatable reads. It then has to check whether it read consistent data when it commits.

- Index Contention

- ▶ Txns install "virtual" index entries to ensure unique-key invariants.

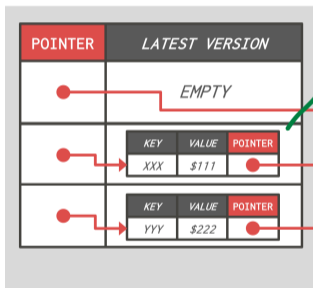
CMU Cicada

pointer change

- In-memory OLTP engine based on optimistic MVCC with append-only storage (N2O).
- **Reference**
 - ▶ Best-effort Inlining
 - ▶ Loosely Synchronized Clocks
 - ▶ Contention-Aware Validation
 - ▶ Index Nodes Stored in Tables
- Designed to be scalable for both low- and high-contention workloads.

Cicada: Best-Effort Inlining

Record Meta-data



- Record Meta-data
- Record meta-data is stored in a fixed location.
- Threads will attempt to inline read-mostly version within this meta-data to reduce version chain traversals.

Cicada: Fast Validation

frequently

- Contention-aware Validation

- ▶ Validate access to recently modified records first.

- Early Consistency Check

- ▶ Pre-validate access set before making global writes.
- ▶ Skip if all recent txns committed successfully.

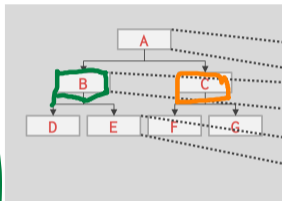
- Incremental Version Search

- ▶ Resume from last search location in version list.

Cicada: Index Storage

SQLite

Index



Index Node Table

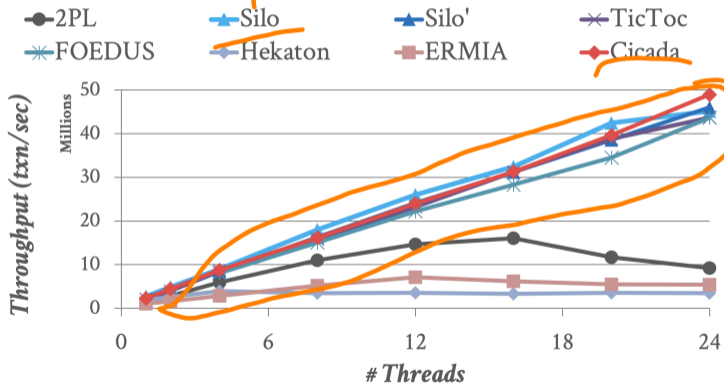
	NODE DATA	POINTER
A ₁	Keys-[100, 200] Pointers-[B, C]	∅
B ₂	Keys-[50, 70] Pointers-[D, E]	● ←
B ₁	Keys-[52, 70] Pointers-[D, E]	∅
E ₃	Keys-[10, 30] Pointers-[RID, RID]	● ←
E ₂	Keys-[11, 30] Pointers-[RID, RID]	● ←
E ₁	Keys-[12, 30] Pointers-[RID, RID]	● ←

auxiliary

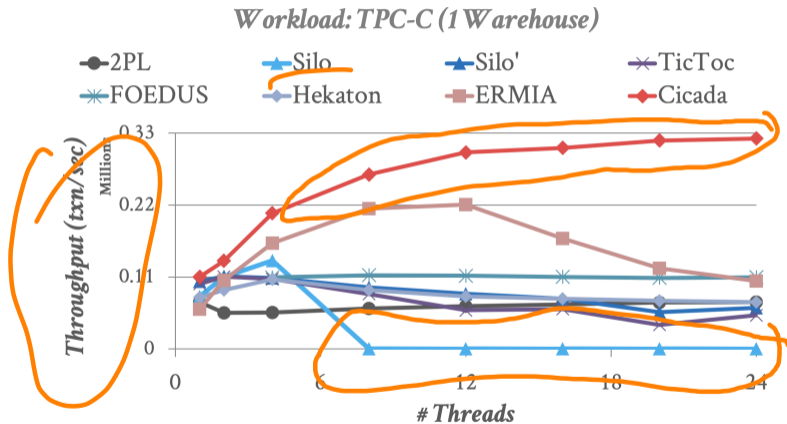
primary

Cicada: Low Contention

DCC

Workload: YCSB (95% read / 5% write) - 1 op per txn

Cicada: High Contention



Conclusion

Parting Thoughts

- There are several other implementation factors for an MVCC DBMS beyond the four main design decisions that we discussed last class.
- Need to balance the trade-offs between indirection and performance.

Next Class

- Wellness Day
- Midterm Exam
- Query Optimization

Mar 29
Mar 26 (Fri)