Multi-Version Concurrency Control (Protocols)

Carnegie Mellon University
ADVANCED DATABASE SYSTEMS

Lecture #04

@Andy_Pavlo // 15-721 // Spring 2020
We discussed the four major design decisions for building a MVCC DBMS.
→ Concurrency Control Protocol
→ Version Storage
→ Garbage Collection
→ Index Management
TODAY'S AGENDA

Microsoft Hekaton (SQL Server)
TUM HyPer
SAP HANA
CMU Cicada
MICROSOFT HEKATON

Incubator project started in 2008 to create new OLTP engine for MSFT SQL Server (MSSQL).
→ Led by DB ballers Paul Larson and Mike Zwilling

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.
HEKATON 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.
**Thread #1**

*Begin @ 25*

**HEKATON: OPERATIONS**

**Main Data Table**

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
</tr>
<tr>
<td>A_2</td>
<td>20</td>
<td>∞</td>
<td>$200</td>
</tr>
</tbody>
</table>
### HEKATON: OPERATIONS

#### Thread #1

**Begin @ 25**

**Main Data Table**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>$A_2$</td>
<td>20</td>
<td>$\infty$</td>
<td>$200$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>
### HEKATON: OPERATIONS

#### Thread #1
*Begin @ 25*

**Main Data Table**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>∞</td>
<td>$200</td>
<td>Ø</td>
</tr>
</tbody>
</table>

**READ(A)**
Thread #1

Begin @ 25

READ(A)
WRITE(A)

Main Data Table

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>∞</td>
<td>$200</td>
<td>Ø</td>
</tr>
</tbody>
</table>
**Thread #1**

*Begin @ 25*

**HEKATON: OPERATIONS**

**Main Data Table**

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
</tr>
<tr>
<td>A_2</td>
<td>20</td>
<td>∞</td>
<td>$200</td>
</tr>
</tbody>
</table>
HEKATON: OPERATIONS

Thread #1

Begin @ 25

Main Data Table

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>$100</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>$200</td>
<td>∞</td>
<td>$200</td>
<td>ø</td>
</tr>
<tr>
<td>$300</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>
**HEKATON: OPERATIONS**

### Main Data Table

<table>
<thead>
<tr>
<th>Thread</th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>∞</td>
<td>$200</td>
<td>Ø</td>
</tr>
<tr>
<td>A₃</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>

*Thread #1*

*Begin @ 25*

### Example Operation

**Txn@25 → 10000000...00000000 00011001**
Thread #1

Begin @ 25

READ(A)

WRITE(A)

HEKATON: OPERATIONS

Main Data Table

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
</tr>
<tr>
<td>A_2</td>
<td>20</td>
<td>∞</td>
<td>$200</td>
</tr>
<tr>
<td>A_3</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
</tr>
</tbody>
</table>
### Thread #1

**Begin @ 25**

#### HEKATON: OPERATIONS

#### Main Data Table

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>Txn@25</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>

**READ(A)**

**WRITE(A)**
**HEKATON: OPERATIONS**

**Thread #1**

*Begin @ 25*

*Commit @ 35*

---

**Main Data Table**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>20</td>
<td>Txn@25</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>
HEKATON: OPERATIONS

Thread #1

Begin @ 25
Commit @ 35

Main Data Table

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
</tr>
<tr>
<td>A_2</td>
<td>20</td>
<td>35</td>
<td>$200</td>
</tr>
<tr>
<td>A_3</td>
<td>35</td>
<td>∞</td>
<td>$300</td>
</tr>
</tbody>
</table>
HEKATON: OPERATIONS

Thread #1

**Begin @ 25**
**Commit @ 35**

![Read(A) and Write(A) icons]

**Main Data Table**

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>20</td>
<td>$100</td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>35</td>
<td>$200</td>
</tr>
<tr>
<td>A₃</td>
<td>35</td>
<td>∞</td>
<td>$300</td>
</tr>
</tbody>
</table>

REWIND
**HEKATON: OPERATIONS**

### Main Data Table

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>$A_2$</td>
<td>20</td>
<td>$Txn@25$</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>$A_3$</td>
<td>$Txn@25$</td>
<td>$\infty$</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>

**Thread #1**

*Begin @ 25*

- **READ(A)**
- **WRITE(A)**

**Thread #2**

*Begin @ 30*
HEKATON: OPERATIONS

Thread #1
Begin @ 25

**Main Data Table**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>20</td>
<td>Txn@25</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>

Thread #2
Begin @ 30

**Operations**

- **READ(A)**
- **WRITE(A)**
HEKATON: OPERATIONS

Thread #1

Begin @ 25

READ(A)

Thread #2

Begin @ 30

READ(A)

Main Data Table

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>20</td>
<td>Txn@25</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>
**HEKATON: OPERATIONS**

**Main Data Table**

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>20</td>
<td>$100</td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>Txn@25</td>
<td>$200</td>
</tr>
<tr>
<td>A₃</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
</tr>
</tbody>
</table>

**Thread #1**

Begin @ 25

Thread #2

Begin @ 30
HEKATON: OPERATIONS

Thread #1
Begin @ 25

READ(A)
WRITE(A)

Thread #2
Begin @ 30

READ(A)
WRITE(A)

Main Data Table

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td>Txn@25</td>
</tr>
<tr>
<td>A2</td>
<td>20</td>
<td>Txn@25</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>
HEKATON: OPERATIONS

Thread #1

Begin @ 25

Thread #2

Begin @ 30

Main Data Table

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>Pointer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>10-20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>20-Txn@25</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Txn@25-∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>
HEKATON: OPERATIONS

Main Data Table

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>20</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>Txn@25</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>A₃</td>
<td>Txn@25</td>
<td>∞</td>
<td>$300</td>
<td></td>
</tr>
</tbody>
</table>

Thread #1
Begin @ 25

Thread #2
Begin @ 30
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

**Txn Events** | **Txn Phases**
---|---
BEGIN  | Get BeginTS, set state to **ACTIVE**
**Normal Processing**  | Track txn's read set, scan set, and write set.
PRECOMMIT  | Get CommitTS, set state to **VALIDATING**
**Validation**  | Validate reads and scans
→ If validation OK, write new versions to redo log
COMMIT  | Set txn state to **COMMITTED**
**Post-Processing**  | Update version timestamps
→ BeginTS in new versions, CommitTS in old versions
TERMINATE  | Set txn state to **TERMINATED**
  | Remove from txn map
HEKATON: TRANSACTION META-DATA

**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.
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.
HEKATON: OPTIMISTIC VS. PESSIONISTIC

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

Source: Paul Larson
HEKATON: LESSONS

Use only lock-free data structures
→ No latches, spin locks, or critical sections
→ Indexes,txn map, memory alloc, garbage collector
→ We will discuss Bw-Trees + Skip Lists later...

Only one single serialization point in the DBMS to get the txn’s begin and commit timestamp
→ Atomic Addition (CAS)
Observations

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.
HYPER MVCC

Column-store with delta record versioning.
→ 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.
## HYPER: STORAGE ARCHITECTURE

### Main Data Table

<table>
<thead>
<tr>
<th>ATTR1</th>
<th>ATTR2</th>
<th>Version Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tupac</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>IceT</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>B.I.G.</td>
<td>$150</td>
<td>Ø</td>
</tr>
<tr>
<td>DrDre</td>
<td>$139</td>
<td>Ø</td>
</tr>
</tbody>
</table>

### Delta Storage (Per Txn)

- **Txn #1**
  - (ATTR2→$199)
  - Ø

- **Txn #2**
  - (ATTR2→$122)
  - Ø
HYPER: STORAGE ARCHITECTURE

**Main Data Table**

<table>
<thead>
<tr>
<th>ATTR1</th>
<th>ATTR2</th>
<th>Version Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tupac</td>
<td>$100</td>
<td></td>
</tr>
<tr>
<td>IceT</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>B.I.G.</td>
<td>$150</td>
<td>$122</td>
</tr>
<tr>
<td>DrDre</td>
<td>$139</td>
<td>$199 (Txn #2)</td>
</tr>
</tbody>
</table>

**Delta Storage (Per Txn)**

- **Txn #1**: \((\text{ATTR2} \rightarrow \$199)\)  
  - Version Vector: \(\varnothing\)

- **Txn #2**: \((\text{ATTR2} \rightarrow \$122)\)  
  - Version Vector: \(\varnothing\)

- **Txn #3**: \((\text{ATTR2} \rightarrow \$100)\)
**HYPER: STORAGE ARCHITECTURE**

### Main Data Table

<table>
<thead>
<tr>
<th>ATTR1</th>
<th>ATTR2</th>
<th>Version Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tupac</td>
<td>$200</td>
<td>$200</td>
</tr>
<tr>
<td>IceT</td>
<td>$200</td>
<td>$200</td>
</tr>
<tr>
<td>B.I.G.</td>
<td>$150</td>
<td>$150</td>
</tr>
<tr>
<td>DrDre</td>
<td>$139</td>
<td>$139</td>
</tr>
</tbody>
</table>

### Delta Storage (Per Txn)

- **Txn #1**
  - (ATTR2→$199) → Ø

- **Txn #2**
  - (ATTR2→$122) → Ø

- **Txn #3**
  - (ATTR2→$100)
HYPER: STORAGE ARCHITECTURE

**Main Data Table**

<table>
<thead>
<tr>
<th>ATTR1</th>
<th>ATTR2</th>
<th>Version Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tupac</td>
<td>$200</td>
<td>$199</td>
</tr>
<tr>
<td>IceT</td>
<td>$200</td>
<td>$122</td>
</tr>
<tr>
<td>B.I.G.</td>
<td>$150</td>
<td>Ø</td>
</tr>
<tr>
<td>DrDre</td>
<td>$139</td>
<td>Ø</td>
</tr>
</tbody>
</table>

**Delta Storage (Per Txn)**

- **Txn #1**: (ATTR2→$199) Ø
- **Txn #2**: (ATTR2→$122) Ø
- **Txn #3**: (ATTR2→$100) Ø, (ATTR2→$139) Ø
**HYPER: STORAGE ARCHITECTURE**

**Main Data Table**

<table>
<thead>
<tr>
<th>ATTR1</th>
<th>ATTR2</th>
<th>Version Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tupac</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>IceT</td>
<td>$200</td>
<td></td>
</tr>
<tr>
<td>B.I.G.</td>
<td>$150</td>
<td>Ø</td>
</tr>
<tr>
<td>DrDre</td>
<td>$200</td>
<td></td>
</tr>
</tbody>
</table>

**Delta Storage (Per Txn)**

- **Txn #1**
  - (ATTR2→$199) Ø

- **Txn #2**
  - (ATTR2→$122) Ø

- **Txn #3**
  - (ATTR2→$100)
  - (ATTR2→$139) Ø
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.
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)

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)
99 IN (10,20,30)
FALSE
(attr2→33)
33 IN (10,20,30)

Txn #1002
(attr2→122)

Txn #1003
(attr1→'IceCube', attr2→199)
HYPER: PRECISION LOCKING

Validating Txn

<table>
<thead>
<tr>
<th>SQL Query</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>SELECT * FROM foo WHERE attr2 &gt; 20 AND attr2 &lt; 30</code></td>
<td></td>
</tr>
<tr>
<td><code>SELECT COUNT(attr1) FROM foo WHERE attr2 IN (10,20,30)</code></td>
<td></td>
</tr>
<tr>
<td><code>SELECT attr1, AVG(attr2) FROM foo WHERE attr1 LIKE '%Ice%' GROUP BY attr1 HAVING AVG(attr2) &gt; 100</code></td>
<td></td>
</tr>
</tbody>
</table>

Delta Storage (Per Txn)

- **Txn #1001**
  - `attr2 = 99`
- **Txn #1002**
  - `attr2 = 33`
- **Txn #1003**
  - `attr1 = 'IceCube', attr2 = 199`

- **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)`

TRUE

'IceCube' LIKE '%Ice%'

15-721 (Spring 2020)
### HYPER: VERSION SYNOPSIS

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.

#### Main Data Table

<table>
<thead>
<tr>
<th>Version Synopsis</th>
<th>ATTR1</th>
<th>ATTR2</th>
<th>Version Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2, 5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tupac</td>
<td>$100</td>
<td>Ø</td>
<td></td>
</tr>
<tr>
<td>IceT</td>
<td>$200</td>
<td>Ø</td>
<td></td>
</tr>
<tr>
<td>B.I.G.</td>
<td>$150</td>
<td>Ø</td>
<td></td>
</tr>
<tr>
<td>DrDre</td>
<td>$99</td>
<td>Ø</td>
<td></td>
</tr>
<tr>
<td>RZA</td>
<td>$300</td>
<td>Ø</td>
<td></td>
</tr>
<tr>
<td>GZA</td>
<td>$300</td>
<td>Ø</td>
<td></td>
</tr>
<tr>
<td>ODB</td>
<td>$0</td>
<td>Ø</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Version Synopsis</th>
<th>ATTR1</th>
<th>ATTR2</th>
<th>Version Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2, 5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tupac</td>
<td>$100</td>
<td></td>
<td>Ø</td>
</tr>
<tr>
<td>IceT</td>
<td>$200</td>
<td></td>
<td>Ø</td>
</tr>
<tr>
<td>B.I.G.</td>
<td>$150</td>
<td></td>
<td>Ø</td>
</tr>
<tr>
<td>DrDre</td>
<td>$99</td>
<td></td>
<td>Ø</td>
</tr>
<tr>
<td>RZA</td>
<td>$300</td>
<td></td>
<td>Ø</td>
</tr>
<tr>
<td>GZA</td>
<td>$300</td>
<td></td>
<td>Ø</td>
</tr>
<tr>
<td>ODB</td>
<td>$0</td>
<td></td>
<td>Ø</td>
</tr>
</tbody>
</table>

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.
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

In-memory HTAP DBMS with time-travel version storage (N2O).
→ 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.
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**

<table>
<thead>
<tr>
<th>RID</th>
<th>VERS?</th>
<th>VERSION</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>True</td>
<td>A₁</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>False</td>
<td>B₃</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>True</td>
<td>C₂</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>True</td>
<td>D₆</td>
<td>-</td>
</tr>
</tbody>
</table>

**Version Storage**

**Hash Table**

- RECORD
  - A
    - A₃
  - C
    - C₅
    - C₄
    - C₃
  - D
    - D₈
    - D₇
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**

<table>
<thead>
<tr>
<th>RID</th>
<th>VERS?</th>
<th>VERSION</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>True</td>
<td>A₁</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>False</td>
<td>B₃</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>True</td>
<td>C₂</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>True</td>
<td>D₆</td>
<td>-</td>
</tr>
</tbody>
</table>

**Version Storage**

**Hash Table**

- A → A₃
- C → C₅
- D → D₆

**Thread #1**

- $T_{id} = 3$

**Txn Meta-Data**

**Txn Contexts**

- $T_{id}=1$
- $T_{id}=2$
- $T_{id}=3$

**Group Commit Context**

- Group 1

**Group Commit Context**

- Group 1

Thread #1

- WRITE(C)
- WRITE(D)
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.

Source: Hyeontaek Lim
OCC LIMITATIONS

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.

Source: Hyeontaek Lim
In-memory OLTP engine based on optimistic MVCC with append-only storage (N2O).

→ 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 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

Contention-aware Validation
→ Validate access to recently modified records first.

Early Consistency Check
→ Pre-validate access set before making global writes.

Incremental Version Search
→ Resume from last search location in version list.

Skip if all recent txns committed successfully.

Source: Hyeontaek Lim
CICADA: INDEX STORAGE

**Index Node Table**

<table>
<thead>
<tr>
<th>NODE DATA</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keys-&gt;[100, 200]</td>
<td>Ø</td>
</tr>
<tr>
<td>Pointers-&gt;[B, C]</td>
<td></td>
</tr>
<tr>
<td>Keys-&gt;[50, 70]</td>
<td>Ø</td>
</tr>
<tr>
<td>Pointers-&gt;[D, E]</td>
<td></td>
</tr>
<tr>
<td>Keys-&gt;[52, 70]</td>
<td>Ø</td>
</tr>
<tr>
<td>Pointers-&gt;[D, E]</td>
<td></td>
</tr>
<tr>
<td>Keys-&gt;[10, 30]</td>
<td>Ø</td>
</tr>
<tr>
<td>Pointers-&gt;[RID, RID]</td>
<td></td>
</tr>
<tr>
<td>Keys-&gt;[11, 30]</td>
<td>Ø</td>
</tr>
<tr>
<td>Pointers-&gt;[RID, RID]</td>
<td></td>
</tr>
<tr>
<td>Keys-&gt;[12, 30]</td>
<td>Ø</td>
</tr>
<tr>
<td>Pointers-&gt;[RID, RID]</td>
<td></td>
</tr>
</tbody>
</table>
CICADA: INDEX STORAGE

Index Node Table

<table>
<thead>
<tr>
<th>NODE DATA</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1: Keys([100, 200]) Pointers([B, C])</td>
<td>Ø</td>
</tr>
<tr>
<td>B2: Keys([50, 70]) Pointers([D, E])</td>
<td>Ø</td>
</tr>
<tr>
<td>B1: Keys([52, 70]) Pointers([D, E])</td>
<td></td>
</tr>
<tr>
<td>E3: Keys([10, 30]) Pointers([\text{RID, RID}])</td>
<td></td>
</tr>
<tr>
<td>E2: Keys([11, 30]) Pointers([\text{RID, RID}])</td>
<td></td>
</tr>
<tr>
<td>E1: Keys([12, 30]) Pointers([\text{RID, RID}])</td>
<td></td>
</tr>
</tbody>
</table>
CICADA: LOW CONTENTION

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

Throughput (txn/sec)

# Threads

Source: Hyeontaek Lim
CICADA: HIGH CONTENTION

Workload: TPC-C (1 Warehouse)

Throughput (txn/sec)

# Threads

Source: Hyeontaek Lim
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

MVCC Garbage Collection
Perf Tutorial for Project #1