Intro to Database Systems (15-445/645)

FALL 2022

Andy Pavlo

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.
Protocol was first proposed in 1978 MIT PhD dissertation.

First implementations was Rdb/VMS and InterBase at DEC in early 1980s.
→ Both were by Jim Starkey, co-founder of NuoDB.
→ DEC Rdb/VMS is now "Oracle Rdb"
→ InterBase was open-sourced as Firebird.
MULTI-VERSION CONCURRENCY CONTROL

Writers do not block readers. Readers do not block writers.

Read-only txns can read a consistent snapshot without acquiring locks. → Use timestamps to determine visibility.

Easily support time-travel queries.
**MVCC - EXAMPLE #1**

Schedule:

- **$TS(T_1)=1$**
- **$TS(T_2)=2$**

Database:

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$</td>
<td>123</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

BEGIN

- **R(A)**
- **BEGIN**
- **W(A)**
- **BEGIN**
- **COMMIT**
- **COMMIT**
**MVCC - EXAMPLE #1**

Schedule:

- **TS(T₁)=1**
  - **T₁**
    - BEGIN
      - R(A)
    - R(A)
    - COMMIT

- **TS(T₂)=2**
  - **T₂**
    - BEGIN
      - W(A)
    - COMMIT

Database:

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>123</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>A₁</td>
<td>456</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

**T₂ creates version A₁ and sets A₀ End-TS.**
**MVCC – EXAMPLE #1**

**Schedule**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>1</td>
<td>Active</td>
</tr>
<tr>
<td>T₂</td>
<td>2</td>
<td>Active</td>
</tr>
</tbody>
</table>

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>123</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>A₁</td>
<td>456</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>1</td>
<td>Active</td>
</tr>
<tr>
<td>T₂</td>
<td>2</td>
<td>Active</td>
</tr>
</tbody>
</table>

**Example #1**

- **TS(T₁)=1**
- **TS(T₂)=2**

- **T₁**
  - **BEGIN**
  - **R(A)**
  - **COMMIT**

- **T₂**
  - **BEGIN**
  - **W(A)**
  - **COMMIT**

**T₂** creates version **A₁** and sets **A₀ End-TS**.
MVCC - EXAMPLE #1

### Txn Status Table

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>1</td>
<td>Active</td>
</tr>
<tr>
<td>T₂</td>
<td>2</td>
<td>Active</td>
</tr>
</tbody>
</table>

### Schedule

- **T₁**
  - BEGIN
  - R(A)
  - COMMIT
  - **TS(T₁)=1**

- **T₂**
  - BEGIN
  - W(A)
  - COMMIT
  - **TS(T₂)=2**

### Database

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>123</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>A₁</td>
<td>456</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

**T₁ reads version A₀.**
**MVCC - EXAMPLE #2**

**Schedule**
- **TS(T₁)=1**
  - T₁
  - BEGIN
  - R(A)
  - W(A)
  - R(A)
  - COMMIT

- **TS(T₂)=2**
  - T₂
  - BEGIN
  - R(A)
  - W(A)
  - R(A)
  - COMMIT

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>123</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>1</td>
<td>Active</td>
</tr>
</tbody>
</table>
MVCC - EXAMPLE #2

**Schedule**

- $TS(T_1) = 1$
- $TS(T_2) = 2$

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$</td>
<td>123</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$A_1$</td>
<td>456</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>1</td>
<td>Active</td>
</tr>
</tbody>
</table>
**MVCC - EXAMPLE #2**

**Schedule**

- **TS(T₁)=1**
- **TS(T₂)=2**

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>123</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A₁</td>
<td>456</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>1</td>
<td>Active</td>
</tr>
</tbody>
</table>

**TIME**
**MVCC - EXAMPLE #2**

**Schedule**

- **TS(T₁) = 1**
  - T₁
    - BEGIN
    - R(A)
    - W(A)
    - R(A)
    - COMMIT

- **TS(T₂) = 2**
  - T₂
    - BEGIN
    - R(A)
    - W(A)
    - COMMIT

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>123</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A₁</td>
<td>456</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₂</td>
<td>Active</td>
</tr>
</tbody>
</table>

**Explanation:**

T₂ reads version A₀ because T₁ has not committed yet.
**MVCC - EXAMPLE #2**

**Schedule**

- $TS(T_1)=1$
- $TS(T_2)=2$

**Txn Status Table**

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>Active</td>
</tr>
<tr>
<td>$T_2$</td>
<td>Active</td>
</tr>
</tbody>
</table>

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$</td>
<td>123</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$A_1$</td>
<td>456</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

**Txn Execution**

- $T_1$: `BEGIN R(A) W(A) COMMIT`
- $T_2$: `BEGIN R(A) W(A) COMMIT`

$T_2$ must stall until $T_1$ commits.
**MVCC - EXAMPLE #2**

**Schedule**

<table>
<thead>
<tr>
<th>Time</th>
<th>Transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T₁</td>
</tr>
<tr>
<td></td>
<td>T₂</td>
</tr>
</tbody>
</table>

- **TS(T₁)=1**
  - T₁: BEGIN, R(A), W(A), COMMIT

- **TS(T₂)=2**
  - T₂: BEGIN, R(A), W(A), COMMIT

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₀</td>
<td>123</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A₁</td>
<td>456</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>1</td>
<td>Active</td>
</tr>
<tr>
<td>T₂</td>
<td>2</td>
<td>Active</td>
</tr>
</tbody>
</table>
MVCC - EXAMPLE #2

**Schedule**

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>( TS(T_1) = 1 )</td>
<td>BEGIN</td>
</tr>
<tr>
<td>( T_1 )</td>
<td>R(A)</td>
</tr>
<tr>
<td></td>
<td>W(A)</td>
</tr>
<tr>
<td></td>
<td>R(A)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>( T_2 )</td>
<td></td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_1 )</td>
<td>1</td>
<td>Committed</td>
</tr>
<tr>
<td>( T_2 )</td>
<td>2</td>
<td>Active</td>
</tr>
</tbody>
</table>

**Version**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_0 )</td>
<td>123</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( A_1 )</td>
<td>456</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>
**MVCC - EXAMPLE #2**

**Schedule**

- $TS(T_1) = 1$
- $TS(T_2) = 2$

**Database**

<table>
<thead>
<tr>
<th>Version</th>
<th>Value</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$</td>
<td>123</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$A_1$</td>
<td>456</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$A_2$</td>
<td>789</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Txn Status Table**

<table>
<thead>
<tr>
<th>TxnId</th>
<th>Timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>1</td>
<td>Committed</td>
</tr>
<tr>
<td>$T_2$</td>
<td></td>
<td>Active</td>
</tr>
</tbody>
</table>

Now $T_2$ can create the new version.
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

**Txn #1**
Change white marbles to black.

**Txn #2**
Change black marbles to white.
WRITE SKEW ANOMALY

*txn #1*
Change white marbles to black.

 txn #2
Change black marbles to white.
**WRITE SKEW ANOMALY**

*txn #1*
Change white marbles to black.

*txn #2*
Change black marbles to white.
WRITE SKEW ANOMALY

**Txn #1**
Change white marbles to black.

**Txn #2**
Change black marbles to white.
MVCC is more than just a concurrency control protocol. It completely affects how the DBMS manages transactions and the database.
MVCC DESIGN DECISIONS

Concurrency Control Protocol
Version Storage
Garbage Collection
Index Management
Deletes
CONCURRENCY CONTROL PROTOCOL

Approach #1: Timestamp Ordering
→ Assign txns timestamps that determine serial order.

Approach #2: Optimistic Concurrency Control
→ Three-phase protocol from last class.
→ Use private workspace for new versions.

Approach #3: Two-Phase Locking
→ Txns acquire appropriate lock on physical version before they can read/write a logical tuple.
The DBMS uses the tuples' pointer field to create a version chain per logical tuple.

→ This allows the DBMS to find the version that is visible to a particular txn at runtime.
→ Indexes always point to the "head" of the chain.

Different storage schemes determine where/what to store for each version.
VERSION STORAGE

Approach #1: Append-Only Storage
→ New versions are appended to the same table space.

Approach #2: Time-Travel Storage
→ Old versions are copied to separate table space.

Approach #3: Delta Storage
→ The original values of the modified attributes are copied into a separate delta record space.
All the physical versions of a logical tuple are stored in the same table space. The versions are inter-mixed.

On every update, append a new version of the tuple into an empty space in the table.
APPEND-ONLY STORAGE

All the physical versions of a logical tuple are stored in the same table space. The versions are inter-mixed.

On every update, append a new version of the tuple into an empty space in the table.
APPEND-ONLY STORAGE

All the physical versions of a logical tuple are stored in the same table space. The versions are inter-mixed.

On every update, append a new version of the tuple into an empty space in the table.
VERSION CHAIN ORDERING

Approach #1: Oldest-to-Newest (O2N)
→ Append new version to end of the chain.
→ Must traverse chain on look-ups.

Approach #2: Newest-to-Oldest (N2O)
→ Must update index pointers for every new version.
→ Do not have to traverse chain on look-ups.
On every update, copy the current version to the time-travel table. Update pointers.
TIME-TRAVEL STORAGE

Main Table

<table>
<thead>
<tr>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_2</td>
<td>$222</td>
</tr>
<tr>
<td>B_1</td>
<td>$10</td>
</tr>
</tbody>
</table>

Time-Travel Table

<table>
<thead>
<tr>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_1</td>
<td>$111</td>
</tr>
<tr>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>A_2</td>
<td>$222</td>
</tr>
</tbody>
</table>

On every update, copy the current version to the time-travel table. Update pointers.
On every update, copy the current version to the time-travel table. Update pointers.

Overwrite master version in the main table and update pointers.
On every update, copy the current version to the time-travel table. Update pointers.

Overwrite master version in the main table and update pointers.
### TIME-TRAVEL STORAGE

#### Main Table

<table>
<thead>
<tr>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_3</td>
<td>$333</td>
</tr>
<tr>
<td>B_1</td>
<td>$10</td>
</tr>
</tbody>
</table>

#### Time-Travel Table

<table>
<thead>
<tr>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_1</td>
<td>$111</td>
</tr>
<tr>
<td>A_2</td>
<td>$222</td>
</tr>
</tbody>
</table>

On every update, copy the current version to the time-travel table. Update pointers.

Overwrite master version in the main table and update pointers.
On every update, copy only the values that were modified to the delta storage and overwrite the master version.
On every update, copy only the values that were modified to the delta storage and overwrite the master version.
On every update, copy only the values that were modified to the delta storage and overwrite the master version.
DELTA STORAGE

Main Table

<table>
<thead>
<tr>
<th>VALUE</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>$333</td>
<td></td>
</tr>
<tr>
<td>$10</td>
<td></td>
</tr>
</tbody>
</table>

Delta Storage Segment

<table>
<thead>
<tr>
<th>DELTA</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>(VALUE→$111)</td>
<td>Ø</td>
</tr>
<tr>
<td>(VALUE→$222)</td>
<td>Ø</td>
</tr>
</tbody>
</table>

On every update, copy only the values that were modified to the delta storage and overwrite the master version.

Txns can recreate old versions by applying the delta in reverse order.
The DBMS needs to remove reclaimable physical versions from the database over time.

→ No active txn in the DBMS can "see" that version (SI).
→ The version was created by an aborted txn.

Two additional design decisions:
→ How to look for expired versions?
→ How to decide when it is safe to reclaim memory?
GARBAGE COLLECTION

Approach #1: Tuple-level
→ Find old versions by examining tuples directly.
→ Background Vacuuming vs. Cooperative Cleaning

Approach #2: Transaction-level
→ Txns keep track of their old versions so the DBMS does not have to scan tuples to determine visibility.
**TUPLE-LEVEL GC**

**Txn #1**
- \( T_{id} = 12 \)

**Txn #2**
- \( T_{id} = 25 \)

**Background Vacuuming:**
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_{100} )</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>( B_{100} )</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>( B_{101} )</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>
Background Vacuuming:
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.
**Background Vacuuming:**
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.
Background Vacuuming:
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.
TUPLE-LEVEL GC

Background Vacuuming:
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.
**Background Vacuuming:** Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.

**Cooperative Cleaning:** Worker threads identify reclaimable versions as they traverse version chain. Only works with O2N.
Background Vacuuming:
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.

Cooperative Cleaning:
Worker threads identify reclaimable versions as they traverse version chain. Only works with O2N.
Background Vacuuming:
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.

Cooperative Cleaning:
Worker threads identify reclaimable versions as they traverse version chain. Only works with O2N.
Background Vacuuming:
Separate thread(s) periodically scan the table and look for reclaimable versions. Works with any storage.

Cooperative Cleaning:
Worker threads identify reclaimable versions as they traverse version version chain. Only works with O2N.
Each txn keeps track of its read/write set. On commit/abort, the txn provides this information to a centralized vacuum worker.

The DBMS periodically determines when all versions created by a finished txn are no longer visible.
### Transaction-Level GC

**Txn #1**

**BEGIN @ 10**

```
UPDATE(A)
```

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₂</td>
<td>1</td>
<td>∞</td>
<td>-</td>
</tr>
<tr>
<td>B₆</td>
<td>8</td>
<td>∞</td>
<td>-</td>
</tr>
</tbody>
</table>
### Transaction-Level GC

**Txn #1**

**BEGIN @ 10**

**UPDATE(A)**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>1</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>B6</td>
<td>8</td>
<td>∞</td>
<td>-</td>
</tr>
<tr>
<td>A3</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
</tbody>
</table>
# Transaction-Level GC

**Txn #1**

**BEGIN @ 10**

**UPDATE(A)**

Old Versions

| A₂ | 2 | - |
| B₆ | 6 | - |
| A₃ | 10 | - |

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₂</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>B₆</td>
<td>8</td>
<td>∞</td>
</tr>
<tr>
<td>A₃</td>
<td>10</td>
<td>∞</td>
</tr>
</tbody>
</table>
**Transaction-Level GC**

**Txn #1**

**BEGIN @ 10**

**Old Versions**

- A₂

**UPDATE(A)**

**UPDATE(B)**

**Table:**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₂</td>
<td>1</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>B₆</td>
<td>8</td>
<td>∞</td>
<td>-</td>
</tr>
<tr>
<td>A₃</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
</tbody>
</table>
**TRANSACTION-LEVEL GC**

**Txn #1**

BEGIN @ 10

**Old Versions**

- A₂

**UPDATE (A)**

**UPDATE (B)**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₂</td>
<td>1</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>B₆</td>
<td>8</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>A₃</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
<tr>
<td>B₇</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
</tbody>
</table>
**TRANSACTION-LEVEL GC**

**Txn #1**

BEGIN @ 10

**Old Versions**

- A₂
- B₆

**BEGIN (A)**

- UPDATE(A)

**BEGIN (B)**

- UPDATE(B)

**BEGIN-TS** | **END-TS** | **DATA**
--- | --- | ---
A₂ | 1 | 10 | -
B₆ | 8 | 10 | -
A₃ | 10 | ∞ | -
B₇ | 10 | ∞ | -
**Transaction-Level GC**

### Txn #1

**BEGIN @ 10**

**COMMIT @ 15**

**Old Versions**

- A₂
- B₆

**Updates**

- UPDATE(A)
- UPDATE(B)

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₂</td>
<td>1</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>B₆</td>
<td>8</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>A₃</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
<tr>
<td>B₇</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
</tbody>
</table>
**TRANSACTION-LEVEL GC**

**Txn #1**

- **BEGIN @ 10**
- **UPDATE(A)**
- **COMMIT @ 15**

**Old Versions**

- **UPDATE(B)**

### Vacuum

**TS<10**

- **A₂**
- **B₆**

### Table

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₂</td>
<td>1</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>B₆</td>
<td>8</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>A₃</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
<tr>
<td>B₇</td>
<td>10</td>
<td>∞</td>
<td>-</td>
</tr>
</tbody>
</table>
Primary key indexes point to version chain head.
→ How often the DBMS must update the pkey index depends on whether the system creates new versions when a tuple is updated.
→ If a txn updates a tuple's pkey attribute(s), then this is treated as a **DELETE** followed by an **INSERT**.

Secondary indexes are more complicated...
Primary key indexes point to version chain head. How often the DBMS must update the primary key index depends on whether the system creates new versions when a tuple is updated. If a transaction updates a tuple’s primary key attribute(s), then this is treated as a DELETE followed by an INSERT.

Secondary indexes are more complicated...
SECONDARY INDEXES

Approach #1: Logical Pointers
→ Use a fixed identifier per tuple that does not change.
→ Requires an extra indirection layer.
→ Primary Key vs. Tuple Id

Approach #2: Physical Pointers
→ Use the physical address to the version chain head.
INDEX POINTERS

GET(A)

PRIMARY INDEX

SECONDARY INDEX

Record Id

A4 → A3 → A2 → A1

Append-Only
Newest-to-Oldest
INDEX POINTERS

PRIMARY INDEX

SECONDARY INDEX

GET(A)

Record Id

{ Append-Only
Newest-to-Oldest

A_4 \rightarrow A_3 \rightarrow A_2 \rightarrow A_1
INDEX POINTERS

PRIMARY INDEX

SECONDARY INDEX

SECONDARY INDEX

SECONDARY INDEX

SECONDARY INDEX

GET(A)

Append-Only
Newest-to-Oldest
INDEX POINTERS

GET(A)

PRIMARY INDEX

SECONDARY INDEX

Record Id

Primary Key

A₄ → A₃ → A₂ → A₁

Append-Only
Newest-to-Oldest

Primary Key
MVCC INDEXES

MVCC DBMS indexes (usually) do not store version information about tuples with their keys.
→ Exception: Index-organized tables (e.g., MySQL)

Every index must support duplicate keys from different snapshots:
→ The same key may point to different logical tuples in different snapshots.
**MVCC DUPLICATE KEY PROBLEM**

**Txn #1**

BEGIN @ 10

![Diagram](image.png)
**MVCC DUPLICATE KEY PROBLEM**

*Txn #1*
BEGIN @ 10

**Txn #2**
BEGIN @ 20

---

**Index**

---

**Table**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>1</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>20</td>
<td>∞</td>
<td>Ø</td>
</tr>
</tbody>
</table>
**MVCC DUPLICATE KEY PROBLEM**

**Txn #1**
BEGIN @ 10

READ(A)

**Txn #2**
BEGIN @ 20

UPDATE(A)
DELETE(A)

**Index**

```
<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>1</td>
<td>20</td>
<td>ø</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>∞</td>
<td>ø</td>
</tr>
</tbody>
</table>
```

---

15-445/645 (Fall 2022)
**MVCC DUPLICATE KEY PROBLEM**

**Txn #1**
BEGIN @ 10
READ(A)

**Txn #2**
BEGIN @ 20
COMMIT @ 25
UPDATE(A)
DELETE(A)

---

**Index**

**Table**

<table>
<thead>
<tr>
<th></th>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>1</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
<td>Ø</td>
</tr>
</tbody>
</table>

---

15-445/645 (Fall 2022)
MVCC DUPLICATE KEY PROBLEM

**Txn #1**
BEGIN @ 10
READ(A)

**Index**

**Txn #2**
BEGIN @ 20
COMMIT @ 25
UPDATE(A)
DELETE(A)

**Txn #3**
BEGIN @ 30
INSERT(A)

**Table**

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>∞</td>
</tr>
</tbody>
</table>

CMU-DB 15-445/645 (Fall 2022)
MVCC DUPLICATE KEY PROBLEM

**Txn #1**
BEGIN @ 10
READ(A)
READ(A)

**Txn #2**
BEGIN @ 20
UPDATE(A)
DELETE(A)

**Txn #3**
BEGIN @ 30
INSERT(A)

---

**Index**

<table>
<thead>
<tr>
<th>BEGIN-TS</th>
<th>END-TS</th>
<th>POINTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_1</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>∞</td>
<td>Ø</td>
</tr>
<tr>
<td>A_1</td>
<td>30</td>
<td>∞</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ø</td>
</tr>
</tbody>
</table>
MVCC INDEXES

Each index's underlying data structure must support the storage of non-unique keys.

Use additional execution logic to perform conditional inserts for pkey / unique indexes.
→ Atomically check whether the key exists and then insert.

Workers may get back multiple entries for a single fetch. They then must follow the pointers to find the proper physical version.
The DBMS physically deletes a tuple from the database only when all versions of a logically deleted tuple are not visible.

→ If a tuple is deleted, then there cannot be a new version of that tuple after the newest version.
→ No write-write conflicts / first-writer wins

We need a way to denote that tuple has been logically delete at some point in time.
Approach #1: Deleted Flag
→ Maintain a flag to indicate that the logical tuple has been deleted after the newest physical version.
→ Can either be in tuple header or a separate column.

Approach #2: Tombstone Tuple
→ Create an empty physical version to indicate that a logical tuple is deleted.
→ Use a separate pool for tombstone tuples with only a special bit pattern in version chain pointer to reduce the storage overhead.
## MVCC Implementations

<table>
<thead>
<tr>
<th>Database</th>
<th>Protocol</th>
<th>Version Storage</th>
<th>Garbage Collection</th>
<th>Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>MV2PL</td>
<td>Delta</td>
<td>Vacuum</td>
<td>Logical</td>
</tr>
<tr>
<td>Postgres</td>
<td>MV-2PL/MV-T0</td>
<td>Append-Only</td>
<td>Vacuum</td>
<td>Physical</td>
</tr>
<tr>
<td>MySQL-InnoDB</td>
<td>MV-2PL</td>
<td>Delta</td>
<td>Vacuum</td>
<td>Logical</td>
</tr>
<tr>
<td>HYRISE</td>
<td>MV-OCC</td>
<td>Append-Only</td>
<td>–</td>
<td>Physical</td>
</tr>
<tr>
<td>Hekaton</td>
<td>MV-OCC</td>
<td>Append-Only</td>
<td>Cooperative</td>
<td>Physical</td>
</tr>
<tr>
<td>SAP HANA</td>
<td>MV-2PL</td>
<td>Time-travel</td>
<td>Hybrid</td>
<td>Logical</td>
</tr>
<tr>
<td>NuoDB</td>
<td>MV-2PL</td>
<td>Append-Only</td>
<td>Vacuum</td>
<td>Logical</td>
</tr>
<tr>
<td>HyPer</td>
<td>MV-OCC</td>
<td>Delta</td>
<td>Txn-level</td>
<td>Logical</td>
</tr>
<tr>
<td>CockroachDB</td>
<td>MV-2PL</td>
<td>Delta (LSM)</td>
<td>Compaction</td>
<td>Logical</td>
</tr>
</tbody>
</table>
CONCLUSION

MVCC is the widely used scheme in DBMSs. Even systems that do not support multi-statement txns (e.g., NoSQL) use it.
NEXT CLASS

Logging and recovery!