Multiversion Concurrency Control

Multiversion Schemes

- Multiversion schemes keep old versions of data item to increase concurrency. Several variants:
  - **Multiversion Timestamp Ordering**
  - **Multiversion Two-Phase Locking**
  - **Snapshot isolation**
- Key ideas:
  - Each successful **write** results in the creation of a new version of the data item written.
  - Use timestamps to label versions.
  - When a **read**(*Q*) operation is issued, select an appropriate version of *Q* based on the timestamp of the transaction issuing the read request, and return the value of the selected version.
- **read**s never have to wait as an appropriate version is returned immediately.

Multiversion Timestamp Ordering

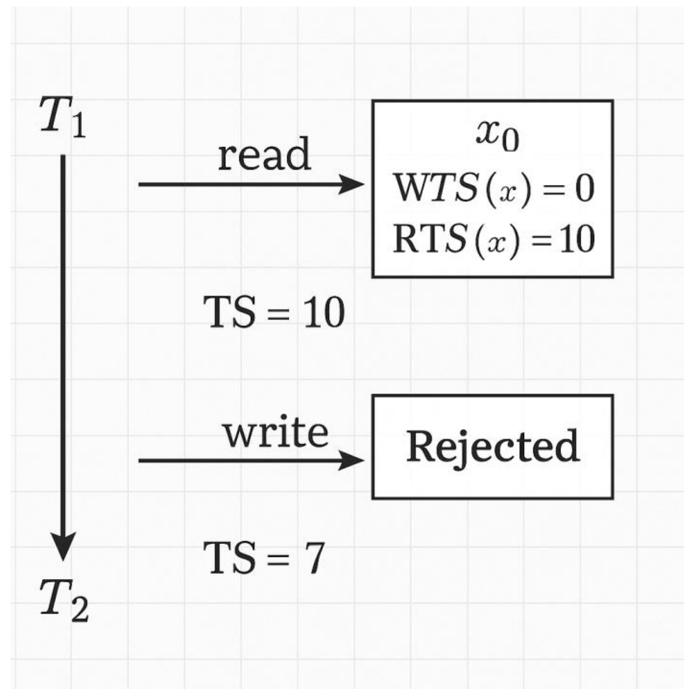
- Each data item Q has a sequence of versions $<Q_1, Q_2, \ldots., Q_m>$. Each version Q_k contains three data fields:
  - **Content** -- the value of version Q_k.
  - **W-timestamp**(Q_k) -- timestamp of the transaction that created (wrote) version Q_k
  - **R-timestamp**(Q_k) -- largest timestamp of a transaction that successfully read version Q_k

Multiversion Timestamp Ordering (Cont)

- Suppose that transaction T_i issues a **read**(Q) or **write**(Q) operation.
- Let Q_k denote the version of Q whose write timestamp is the largest write timestamp less than or equal to TS(T_i).
  1. If **transaction** $\boldsymbol{T_i}$ issues a **read**(Q), then
     - the value returned is the content of version Q_k
     - If R-timestamp(Q_k) < TS(T_i), set R-timestamp(Q_k) = TS(T_i),
  2. If transaction T_i issues a **write**(Q)
     1. if **TS($\boldsymbol{T_i}$)** $<$ **R-timestamp($\boldsymbol{Q_k}$),** then transaction T_i is rolled back.
     2. if **TS($\boldsymbol{T_i}$)** $=$ **W-timestamp($\boldsymbol{Q_k}$),** the contents of Q_k are overwritten
     3. Otherwise, a new version Q_i of Q is created
        - W-timestamp(Q_i) and R-timestamp(Q_i) are initialized to TS(T_i).

Multiversion Timestamp Ordering (Cont)

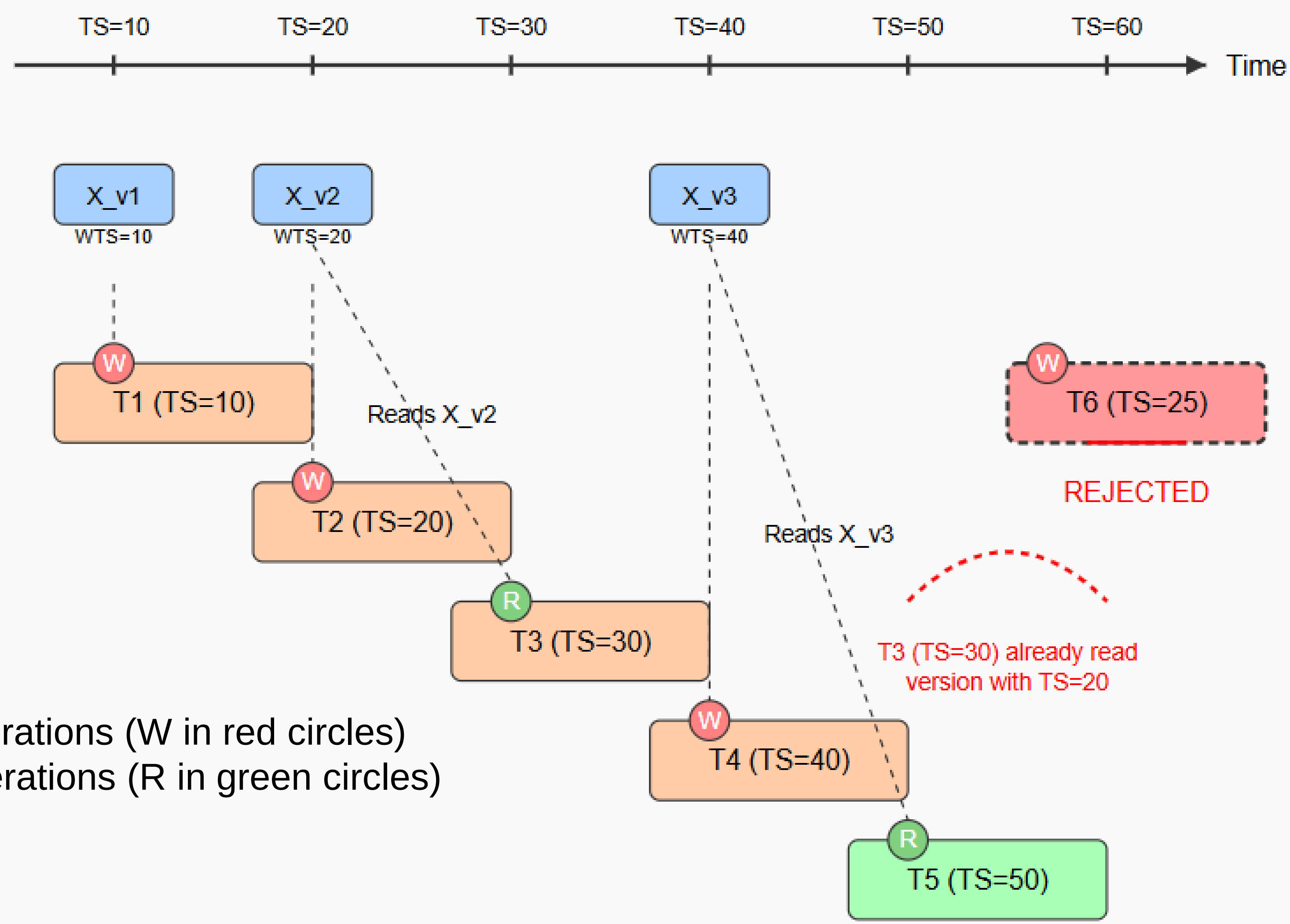
- Observations
  - Reads always succeed
  - A write by T_i is rejected if some other transaction T_j that (in the serialization order defined by the timestamp values) should read T_i's write, has already read a version created by a transaction older than T_i.
- Protocol guarantees serializability



Multiversion Timestamp Ordering (Cont)

```
procedure MVTO_Write(x, T_i):
    // Find the version x_j with the largest
  WTS(x_j) < TS(T_i)
    x_j ← latest_version_of(x) such that
  WTS(x_j) ≤ TS(T_i)
    if RTS(x_j) > TS(T_i) then
        // Conflict: a future-timestamped
  transaction already read the old value
        reject T_i's write on x
    else
        // Safe: install new version
        create version x_i with
            WTS(x_i) ← TS(T_i)
            RTS(x_i) ← TS(T_i)
```

Multiversion Timestamp Ordering



1. Write operations (W in red circles)
2. Read operations (R in green circles)

T6's write (TS=25) is rejected because T3 (TS=30) has already read X_v2 (TS=20)

In the serialization order, T3 should have read T6's write, but it already read an older version

Multiversion Two-Phase Locking

- **Differentiates** between **read-only transactions** and update transactions
- **Update transactions** acquire read and write locks, **and hold all locks up to the end of the transaction**. That is, **update transactions follow rigorous two-phase locking.**
  - Read of a data item returns the latest version of the item
  - The first **write** of Q by T_i results in the creation of a new version Q_i of the data item Q written
    - W-timestamp(Q_i) set to ∞ **initially to not allow other writes**
  - When **update** transaction T_i **completes**, **commit** processing occurs:
    - Value **ts-counter** stored in the database is used to assign timestamps
      - **ts-counter** is locked in two-phase manner
    - Set **W-timestamp(Q_i)** = (**ts-counter** + 1) for all versions Q_i that it creates
    - **ts-counter** = **ts-counter +** 1
    - Thereby, those transactions that start before $\boldsymbol{T_i}$ commits will see the value before the updates by T_i .

Multiversion Two-Phase Locking

Imagine an online store has a product **P** with a price history. Transactions update or read the price.

Initial State:

- The product **P** has an initial price: **$100,** The **timestamp counter (ts-counter) = 10**.

Transactions:

Transaction T1 (Update Transaction)

- **Starts** at **ts = 11**. **Reads** the latest price (**$100**).
- **First Write**: Creates a **new version P1** with price **$120.**
  - The **write timestamp W-ts(P1) is set to ∞** (blocking other writes).
- **Commit Process:**
- Locks **ts-counter**. Sets **W-ts(P1) = ts-counter + 1** → **W-ts(P1) = 11**. Updates **ts-counter = 11**.

Transaction T2 (Read-Only Transaction)

- **Starts at ts = 10** (before T1 commits).
- **Reads product P**, but since **T1 hasn't committed**, it **sees the older price ($100)**.

Multiversion Two-Phase Locking (Cont.)

- **Read-only transactions**
  - are assigned a **timestamp** = **ts-counter** when they start execution
  - follow the multiversion timestamp-ordering protocol for performing reads
    - Do not obtain any locks
- Read-only transactions that start after T_i increments **ts-counter** will see the values updated by T_i.
- Read-only transactions that start before T_i increments the **ts-counter** will see the value before the updates by T_i.
- Only serializable schedules are produced.

Multiversion Two-Phase Locking (Cont.)

- Example
- One product P with initial price: **$100**
- A **global timestamp counter (ts-counter) = 10**

Transaction T1 (Update Transaction)

- **Starts**: ts(T1) = 11
- **Reads**: version P0 (price = $100, W-ts = 10)
- **Tentatively creates** a new version: P1 with price = $120
  - W-ts(P1) = ∞ (not committed yet)
- **At this moment:**
  - `P0`: price = $100, `W-ts = 10`
  - `P1`: price = $120, `W-ts = ∞` (invisible to others)

Transaction	Start Time	Sees Version	W-ts of Version	Observed Price
T2 (read-only)	**10**	P0	10	$100
T1 (update)	**11**	—	—	— (writes P1)
T3 (read-only)	**11**	P1	11	$120

MVCC: Implementation Issues

- Creation of multiple versions increases **storage overhead**
  - **Extra tuples**
  - **Extra space** in each tuple for storing version information
- Versions can, however, be **garbage collected**
  - E.g., if Q has two versions Q5 and Q9, and the oldest active transaction has timestamp > 9, than Q5 will never be required again
- Issues with
  - **primary** key and **foreign key constraint** checking
  - **Indexing** of records with multiple versions

See textbook for details

Snapshot Isolation

- Motivation: Decision support queries that read large amounts of data have concurrency conflicts with **OLTP** transactions that update a few rows
  - Poor performance results
- Solution 1: Use multiversion 2-phase locking
  - Give logical "snapshot" of database state to read only transaction
    - Reads performed on snapshot
  - **Update** (read-write) **transactions** use **normal locking**
  - **Works** well, but how does system know a transaction is **read only**?
- Solution 2 (partial): **Give snapshot of database state** to every transaction
  - Reads performed on snapshot
  - Use 2-phase locking on updated data items
  - Problem: variety of anomalies such as lost update can result
  - Better solution: snapshot isolation level (next slide)

Snapshot Isolation

- A transaction T1 executing with Snapshot Isolation
  - Takes **snapshot** of **committed** data at **start**
  - Always **reads/modifies data** in its **own snapshot**
  - **Updates** of **concurrent transactions** are not **visible** to **T1**
  - **Writes** of T1 **complete** when it **commits**
  - **First-committer-wins rule**:
    - Commits only if no other concurrent transaction has already written data that T1 intends to write.

T1	T2	T3
W(Y := 1) Commit		
	Start R(X) → 0 R(Y)→ 1	
		W(X:=2) W(Z:=3) Commit
	R(Z) → 0 R(Y) → 1 W(X:=3) Commit-Req Abort	

Concurrent updates not visible
Own updates are visible
Not first-committer of X
Serialization error, T2 is rolled back

Snapshot Read

- Concurrent updates invisible to snapshot read

$X_0 = 100,\ Y_0 = 0$

T_1 deposits 50 in Y	T_2 withdraws 50 from X
$r_1(X_0, 100)$	
$r_1(Y_0, 0)$	
	$r_2(Y_0, 0)$
	$r_2(X_0, 100)$
	$w_2(X_2, 50)$
$w_1(Y_1, 50)$	
$r_1(X_0, 100)$ (update by T_2 not seen)	
$r_1(Y_1, 50)$ (can see its own updates)	
	$r_2(Y_0, 0)$ (update by T_1 not seen)

$X_2 = 50,\ Y_1 = 50$

Snapshot Write: First Committer Wins

$X_0 = 100$

T_1 deposits 50 in X	T_2 withdraws 50 from X
$r_1(X_0, 100)$	
	$r_2(X_0, 100)$
	$w_2(X_2, 50)$
$w_1(X_1, 150)$	
$commit_1$	
	$commit_2$ (Serialization Error T_2 is rolled back)

$X_1 = 150$

- Variant: "**First-updater-wins**"
  - Check for concurrent updates when write occurs by locking item
    - But lock should be held till all concurrent transactions have finished
  - (Oracle uses **this plus some extra features**)
  - Differs **only in when abort occurs**, otherwise equivalent

Benefits of SI

- **Reads are *never* blocked**,
  - and also don't block other txns activities
- Performance similar to **Read Committed**
- Avoids **several anomalies**
  - **No** dirty read, i.e. no read of **uncommitted data**
  - **No lost** update
    - I.e., update made by a transaction is overwritten by another transaction that did not see the update)
  - No non-repeatable read
    - I.e., if read is executed again, it will see the same value
- Problems with SI
  - SI does not always give **serializable** executions
    - Serializable: among two concurrent txns, one sees the effects of the other
    - In SI: neither sees the effects of the other
  - Result: Integrity constraints can be violated

Snapshot Isolation

- Example of problem with SI
  - Initially A = 3 and B = 17
    - Serial execution: A = ??, B = ??
    - if both transactions start at the same time, with snapshot isolation: A = ?? , B = ??
- Called **skew write**
- Skew also occurs with inserts
  - E.g:
    - Find max order number among all orders
    - Create a new order with order number = previous max + 1
    - Two transaction can both create order with same number
      - Is an example of phantom phenomenon

T_i	T_j
read(A)	
read(B)	
	read(A)
	read(B)
A=B	
	B=A
write(A)	
	write(B)