Concurrency Control
Lock-Based Protocols

- A lock is a mechanism to control concurrent access to a data item

- Data items can be locked in two modes:
  1. **exclusive (X) mode**: Data item can be both read as well as written.
     - X-lock is requested using `lock-X` instruction.
  2. **shared (S) mode**: Data item can only be read.
     - S-lock is requested using `lock-S` instruction.

- Lock requests are made to concurrency-control manager

- Transaction can proceed **only after request is granted**
Lock-Based Protocols (Cont.)

- Lock-compatibility matrix

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>X</td>
<td>false</td>
<td>false</td>
</tr>
</tbody>
</table>

- A transaction may be granted a lock on an item if the requested lock is compatible with locks already held on the item by other transactions.

- Any number of transactions can hold shared locks on an item, but if any transaction holds an exclusive lock on the item no other transaction may hold any lock on the item.

- If a lock cannot be granted, the requesting transaction is made to wait till all incompatible locks held by other transactions have been released. The lock is then granted.
Lock-Based Protocols (Cont.)

- Example of a transaction performing locking:
  
  \[ T_2: \text{lock-S}(A); \]
  
  \[ \text{read } (A); \]
  
  \[ \text{unlock}(A); \]
  
  \[ \text{lock-S}(B); \]
  
  \[ \text{read } (B); \]
  
  \[ \text{unlock}(B); \]
  
  \[ \text{display}(A+B) \]

- Locking as above is not sufficient to guarantee serializability — if \( A \) and \( B \) get updated in-between the read of \( A \) and \( B \), the displayed sum would be wrong.

- A **locking protocol** is a set of rules followed by all transactions while requesting and releasing locks.

- Locking protocols restrict the set of possible schedules.

- The locking protocol must ensure serializability.
Fig 16.4 Schedule for Transactions: non-serializable schedule

Sample Transactions with Locks

- **T1**: lock-X(B)
  - read(B)
  - B = B - 50;
  - write(B);
  - unlock(B);
  - lock-X(A);
  - unlock(B);
  - read(A);
  - display(A+B);
  - A = A + 50;
  - write(A);
  - unlock(A);

- **T2**: lock-S(A)
  - read(A)
  - lock-S(B)
  - unlock(A)
  - read(B);
  - unlock(B)
  - display(A+B);

- **concurrency-control manager**
  - lock-X(B)
  - grant-X(B, T1)
  - Read(B)
  - B = B - 50;
  - write(B);
  - unlock(B);
  - lock-S(A)
  - grant-S(A, T2)
  - read(A)
  - lock-S(B)
  - unlock(A)
  - lock-S(B)
  - grant-S(B, T2)
  - read(B)
  - unlock(B)
  - display(A+B);
  - lock-X(A)
  - grant-X(A, T2)
  - read(A)
  - A = A + 50;
  - write(A);
  - unlock(A);
Pitfalls of Lock-Based Protocols

- Consider the partial schedule
- Neither $T_3$ nor $T_4$ can make progress
  - Executing `lock-S(B)` causes $T_4$ to wait for $T_3$ to release its lock on $B$, while executing `lock-X(A)` causes $T_3$ to wait for $T_4$ to release its lock on $A$.
- Such a situation is called a **deadlock**.
  - To handle a deadlock one of $T_3$ or $T_4$ must be rolled back and its locks must be released.
- The potential for deadlock exists in most locking protocols.
- **Starvation** is also possible if concurrency control manager is badly designed. For example:
  - A transaction may be waiting for an X-lock on an item, while a sequence of other transactions request and are granted an S-lock on the same item.
  - The same transaction is repeatedly rolled back due to deadlocks.
- Concurrency control manager can be designed to prevent starvation.
The Two-Phase Locking Protocol

- This is a protocol which ensures conflict-serializable schedules.
- **Phase 1: Growing Phase**
  - transaction may obtain locks
  - transaction may not release locks
- **Phase 2: Shrinking Phase**
  - transaction may release locks
  - transaction may not obtain locks
- The protocol assures serializability.
  - It can be proved that the transactions can be serialized in the order of their **lock points** (i.e. the point where a transaction acquired its final lock).
The Two-Phase Locking Protocol

- Two-phase locking does not ensure freedom from deadlocks
- Cascading roll-back is possible under two-phase locking.

**Strict two-phase locking.**
- Here a transaction must hold *all its exclusive locks* till it commits/aborts.
- No cascading rollback

**Rigorous two-phase locking** is even stricter:
- Here *all locks (shared and exclusive)* are held till commit/abort.
- No cascading rollback (of course)
- In this protocol transactions can be serialized in the order in which they commit.
### Fig 16.8 Partial Schedule Under 2PL

<table>
<thead>
<tr>
<th>$T_5$</th>
<th>$T_6$</th>
<th>$T_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>lock-x($A$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>read($A$)</td>
<td>lock-x($A$)</td>
<td></td>
</tr>
<tr>
<td>lock-S($B$)</td>
<td>read($A$)</td>
<td></td>
</tr>
<tr>
<td>read($B$)</td>
<td>write($A$)</td>
<td></td>
</tr>
<tr>
<td>write($A$)</td>
<td>unlock($A$)</td>
<td>lock-S($A$)</td>
</tr>
<tr>
<td>unlock($A$)</td>
<td></td>
<td>read($A$)</td>
</tr>
</tbody>
</table>
The Two-Phase Locking Protocol

- There can be conflict serializable schedules that cannot be obtained if 2PL is used.
- However, in the absence of extra information (e.g., ordering of access to data), 2PL is needed for conflict serializability in the following sense: Given a transaction $T_i$ that does not follow 2PL, we can find a transaction $T_j$ that uses 2PL, and a schedule for $T_i$ and $T_j$ that is not conflict serializable.
2PL with Lock Conversions

- The original lock mode with (lock-X, lock-S)
  - assign lock-X on a data D when D is both read and written

- Two-phase locking with lock conversions:
  - **First Phase**:
    - can acquire a lock-S on item
    - can acquire a lock-X on item
    - can convert a lock-S to a lock-X (upgrade)
  - **Second Phase**:
    - can release a lock-S
    - can release a lock-X
    - can convert a lock-X to a lock-S (downgrade)

- This protocol assures serializability.

- The refined 2PL gets more concurrency than the original 2PL
** Unless lock conversion, the first step “lock-S (a1)” in T8 should be “lock-X (a1)”, then T8 and T9 in the following schedule cannot be run concurrently in the original 2PL.

** However, T8 and T9 can run concurrently in the refined 2PL owing to the lock conversion

<table>
<thead>
<tr>
<th>T8</th>
<th>T9</th>
</tr>
</thead>
<tbody>
<tr>
<td>read (a1)</td>
<td>read (a1)</td>
</tr>
<tr>
<td>read (a2)</td>
<td>read (a2)</td>
</tr>
<tr>
<td>.....</td>
<td>.....</td>
</tr>
<tr>
<td>read (an)</td>
<td>read (an)</td>
</tr>
<tr>
<td>a1 = a1 + a2 ... an</td>
<td>a3 = a2 – a1</td>
</tr>
<tr>
<td>write(a1)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T8</th>
<th>T9</th>
</tr>
</thead>
<tbody>
<tr>
<td>lock-X (a1)</td>
<td>lock-S (a1)</td>
</tr>
<tr>
<td>lock –S (a2)</td>
<td>lock-S (a2)</td>
</tr>
<tr>
<td>.....</td>
<td>lock-S (a3)</td>
</tr>
<tr>
<td>lock-S (an)</td>
<td>lock-S (a4)</td>
</tr>
<tr>
<td>unlock (a1)</td>
<td>unlock(a1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T8</th>
<th>T9</th>
</tr>
</thead>
<tbody>
<tr>
<td>lock-S (a1)</td>
<td>lock-S (a1)</td>
</tr>
<tr>
<td>lock-S (a2)</td>
<td>lock-S (a2)</td>
</tr>
<tr>
<td>lock-S (a3)</td>
<td>lock-S (a3)</td>
</tr>
<tr>
<td>lock-S (a4)</td>
<td>unlock(a1)</td>
</tr>
<tr>
<td>lock-S (an)</td>
<td>unlock(a2)</td>
</tr>
<tr>
<td>upgrade (a1)</td>
<td></td>
</tr>
</tbody>
</table>

- **Strict 2PL (with Lock conversions) & Rigorous 2PL (with Lock conversions)** are used extensively in commercial DBMSs
Implementation of Locking

- A **Lock manager** can be implemented as a separate process to which transactions send lock and unlock requests.
- The lock manager replies to a lock request by sending a lock grant messages (or a message asking the transaction to roll back, in case of a deadlock).
- The requesting transaction waits until its request is answered.
- The lock manager maintains a data structure called a **lock table** to record granted locks and pending requests.
- The lock table is usually implemented as an in-memory hash table indexed on the name of the data item being locked.
  - Hashing with overflow chaining
  - Lock request on item I ➔ Hashing (I) & enqueue & wait
  - Lock granted
  - Unlock request ➔ dequeue
Lock Table

- Black rectangles indicate **granted locks**, white ones indicate **waiting requests**
- Lock table also records the type of lock granted or requested
- New request is added to the end of the queue of requests for the data item, and granted if it is compatible with all earlier locks
- Unlock requests result in the request being deleted, and later requests are checked to see if they can now be granted
- If transaction aborts, all waiting or granted requests of the transaction are deleted
  - lock manager may keep a list of locks held by each transaction, to implement this efficiently
Graph-Based Protocols

- Graph-based protocols are an alternative to two-phase locking.
- Impose a partial ordering \( \rightarrow \) on the set \( D = \{d_1, d_2, ..., d_h\} \) of all data items.
  - If \( d_i \rightarrow d_j \) then any transaction accessing both \( d_i \) and \( d_j \) must access \( d_i \) before accessing \( d_j \).
  - Implies that the set \( D \) may now be viewed as a directed acyclic graph (DAG), called a database graph.

- The tree-protocol is a simple kind of graph protocol.
  - Only exclusive locks are allowed.
  - The first lock by \( T_i \) may be on any data item if there is no lock on the data item.
  - Subsequently, a data \( Q \) can be locked by \( T_i \) only if the parent of \( Q \) is currently locked by \( T_i \).
  - Data items may be unlocked at any time.
### Fig 16.12 Serializable Schedule Under the Tree Protocol

<table>
<thead>
<tr>
<th>$T_{10}$</th>
<th>$T_{11}$</th>
<th>$T_{12}$</th>
<th>$T_{13}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>lock-$x(B)$</td>
<td>lock-$x(D)$</td>
<td>lock-$x(B)$</td>
<td>lock-$x(D)$</td>
</tr>
<tr>
<td>lock-$x(E)$</td>
<td>lock-$x(H)$</td>
<td>lock-$x(E)$</td>
<td>lock-$x(H)$</td>
</tr>
<tr>
<td>lock-$x(D)$</td>
<td>unlock($D$)</td>
<td>unlock($D$)</td>
<td>unlock($D$)</td>
</tr>
<tr>
<td>unlock($B$)</td>
<td>unlock($E$)</td>
<td>unlock($E$)</td>
<td>unlock($E$)</td>
</tr>
<tr>
<td>unlock($E$)</td>
<td>unlock($H$)</td>
<td>unlock($E$)</td>
<td>unlock($H$)</td>
</tr>
<tr>
<td>lock-$x(G)$</td>
<td>unlock($D$)</td>
<td>lock-$x(D)$</td>
<td>lock-$x(D)$</td>
</tr>
<tr>
<td>unlock($D$)</td>
<td>unlock($B$)</td>
<td>unlock($B$)</td>
<td>unlock($B$)</td>
</tr>
</tbody>
</table>
Graph-Based Protocols (Cont.)

- The tree protocol ensures conflict serializability as well as freedom from deadlock.
- Unlocking may occur earlier in the tree-locking protocol than in the two-phase locking protocol.
  - shorter waiting times, and increase in concurrency
  - protocol is deadlock-free, no rollbacks are required
  - the abort of a transaction can still lead to cascading rollbacks.
    (this correction has to be made in the book also.)
- However, in the tree-locking protocol, a transaction may have to lock data items that it does not access.
  - increased locking overhead and additional waiting time
  - potential decrease in concurrency
- Some schedules not possible under 2PL are possible under tree protocol, and vice versa.
Timestamp-Based Protocols

- Each transaction is issued a timestamp when it enters the system.
- If an old transaction $T_i$ has time-stamp $TS(T_i)$, a new transaction $T_j$ is assigned time-stamp $TS(T_j)$ such that $TS(T_i) < TS(T_j)$.
- The protocol manages concurrent execution such that the time-stamps determine the serializability order.
- In order to assure such behavior, the protocol maintains for each data $Q$ two timestamp values:
  - **W-timestamp**($Q$) is the largest time-stamp of any transaction that executed **write**($Q$) successfully.
  - **R-timestamp**($Q$) is the largest time-stamp of any transaction that executed **read**($Q$) successfully.
Timestamp-Based Protocols (Cont.)

- The timestamp ordering protocol ensures that any conflicting read and write operations are executed in timestamp order.
- Suppose a transaction $T_i$ issues a read($Q$)
  1. If $TS(T_i) < W$-timestamp($Q$),
     - $T_i$ needs to read a value of $Q$ that was already overwritten.
     - Hence, the read operation is rejected, and $T_i$ is rolled back.
  2. If $TS(T_i) \geq W$-timestamp($Q$),
     - The read operation is executed, and R-timestamp($Q$) is set to the maximum of R-timestamp($Q$) and $TS(T_i)$. 
Suppose that transaction $T_i$ issues a \texttt{write}(Q).

1. If $\text{TS}(T_i) < R$-timestamp(Q),
   - The value of $Q$ that $T_i$ is producing was needed previously, and the system assumed that that value would never be produced.
   - Hence, the \texttt{write} operation is rejected, and $T_i$ is rolled back.

2. If $\text{TS}(T_i) < W$-timestamp(Q),
   - Then $T_i$ is attempting to write an obsolete value of $Q$.
   - Hence, this \texttt{write} operation is rejected, and $T_i$ is rolled back.

3. Otherwise, ($\text{TS}(T_i) \geq R$-timestamp(Q) and $\text{TS}(T_i) \geq W$-timestamp(Q))
   - The \texttt{write} operation is executed, and $W$-timestamp(Q) is set to $\text{TS}(T_i)$. 
Example Use of the Timestamp Protocol

A partial schedule for several data items for transactions with timestamps 1, 2, 3, 4, 5

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$T_4$</th>
<th>$T_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>read($Y$)</td>
<td>read($Y$)</td>
<td>write($Y$) write($Z$)</td>
<td>write($Y$)</td>
<td>read($X$)</td>
</tr>
<tr>
<td>$T_1$</td>
<td>read($X$)</td>
<td>abort</td>
<td>abort</td>
<td></td>
<td>write($Z$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>write($Y$) write($Z$)</td>
</tr>
</tbody>
</table>
Correctness of Timestamp-Ordering Protocol

- The timestamp-ordering protocol guarantees **serializability** since all the arcs in the precedence graph are of the form:

```
transaction with smaller timestamp  transaction with larger timestamp
```

Thus, there will be **no cycles** in the precedence graph.

- Timestamp protocol ensures **freedom from deadlock** as no transaction ever waits.

- But the schedule may not be **cascade-free**, and may not even be **recoverable**.
Fig 16.13 Schedule 3: Not possible under 2PL, but possible under the time stamping protocol

<table>
<thead>
<tr>
<th>$T_{14}$</th>
<th>$T_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>read($B$)</td>
<td>read($B$)</td>
</tr>
<tr>
<td></td>
<td>$B := B - 50$</td>
</tr>
<tr>
<td>read($A$)</td>
<td>write($B$)</td>
</tr>
<tr>
<td>display($A + B$)</td>
<td>read($A$)</td>
</tr>
<tr>
<td></td>
<td>$A := A + 50$</td>
</tr>
<tr>
<td></td>
<td>write($A$)</td>
</tr>
<tr>
<td></td>
<td>display($A + B$)</td>
</tr>
</tbody>
</table>

Should wait in 2PL
Recoverability and Cascade Freedom

- Problem with timestamp-ordering protocol:
  - Suppose $T_i$ aborts, but $T_j$ has read a data item written by $T_i$
  - Then $T_j$ must abort; if $T_j$ had been allowed to commit earlier, the schedule is not recoverable.
  - Further, any transaction that has read a data item written by $T_j$ must abort
  - This can lead to cascading rollback --- that is, a chain of rollbacks

- Solution:
  - A transaction is structured such that its writes are all performed at the end of its processing
  - All writes of a transaction form an atomic action; no transaction may execute while a transaction is being written

- Starvation problem
  - A transaction that aborts is restarted with a new timestamp
Time Stamping with Thomas’ Write Rule

- Modified version of the timestamp-ordering protocol in which obsolete write operations may be ignored under certain circumstances.

- When \( T_i \) attempts to write data item \( Q \), if \( TS(T_i) < W\text{-timestamp}(Q) \), then \( T_i \) is attempting to write an obsolete value of \( \{Q\} \).
  - Hence, rather than rolling back \( T_i \) as the timestamp ordering protocol would have done, this “write” operation can be ignored.

- Otherwise this protocol is the same as the timestamp ordering protocol.

- Thomas' Write Rule allows greater potential concurrency.
  - Unlike previous protocols, it allows some view-serializable schedules that are not conflict-serializable.

**Fig 16.14 Schedule 4:**
Not possible under 2PL, but possible in the timestamping with the Thomas’s write rule

<table>
<thead>
<tr>
<th></th>
<th>( T_{16} )</th>
<th>( T_{17} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>read(( Q ))</td>
<td>write(( Q ))</td>
<td>write(( Q ))</td>
</tr>
</tbody>
</table>
**Validation-Based Protocol**

- Execution of transaction $T_i$ is done in three phases.
  
  1. **Read and execution phase**
     - Transaction $T_i$ writes only to temporary local variables
  
  2. **Validation phase**
     - Transaction $T_i$ performs a "validation test" to determine if local variables can be written without violating serializability.
  
  3. **Write phase**
     - If $T_i$ is validated, the updates are applied to the database; otherwise, $T_i$ is rolled back.

- The three phases of concurrently executing transactions can be interleaved, but each transaction must go through the three phases in that order.

- Also called as **optimistic concurrency control** since transaction executes fully in the hope that all will go well during validation.
Each transaction $T_i$ has 3 timestamps
- **Start**($T_i$): the time when $T_i$ started its execution
- **Validation**($T_i$): the time when $T_i$ entered its validation phase
- **Finish**($T_i$): the time when $T_i$ finished its write phase

Serializability order is determined by **timestamp given at validation time**, to increase concurrency.
- Thus $TS(T_i)$ is given the value of **Validation**($T_i$).

This protocol is useful and gives greater degree of concurrency if probability of conflicts is low.
- That is because the serializability order is not pre-decided and relatively less transactions will have to be rolled back.
Validation Test for Transaction $T_j$

- If for all $T_i$ with $TS(T_i) < TS(T_j)$ either one of the following conditions holds:
  - $\text{finish}(T_i) < \text{start}(T_j)$
  - $\text{start}(T_j) < \text{finish}(T_i) < \text{validation}(T_j)$ and the set of data items written by $T_i$ does not intersect with the set of data items read by $T_j$

  then validation succeeds and $T_j$ can be committed.

  Otherwise, validation fails and $T_j$ is aborted.

<table>
<thead>
<tr>
<th>Ti:</th>
<th>start $\rightarrow$ validation $\rightarrow$ finish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tj:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ti:</th>
<th>start $\rightarrow$ validation $\rightarrow$ finish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tj:</td>
<td>start $\rightarrow$ validation $\rightarrow$ finish</td>
</tr>
</tbody>
</table>

**Justification:** Either first condition is satisfied, and there is no overlapped execution, or second condition is satisfied and

1. the writes of $T_j$ do not affect reads of $T_i$ since they occur after $T_i$ has finished its reads.

2. the writes of $T_i$ do not affect reads of $T_j$ since $T_j$ does not read any item written by $T_i$. 
**Schedule Produced by Validation**

- Example of schedule produced using validation

<table>
<thead>
<tr>
<th>$T_{14}$</th>
<th>$T_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>read($B$)</td>
<td>read($B$)</td>
</tr>
<tr>
<td>read($A$)</td>
<td>$B$: - $B$-50</td>
</tr>
<tr>
<td>(validate)</td>
<td>read($A$)</td>
</tr>
<tr>
<td>display ($A+B$)</td>
<td>$A$: - $A$+50</td>
</tr>
<tr>
<td>(validate)</td>
<td>write ($B$)</td>
</tr>
<tr>
<td></td>
<td>write ($A$)</td>
</tr>
</tbody>
</table>

NOT POSSIBLE IN 2PL
Multiple Granularity

- Allow data items to be of various sizes and define a hierarchy of data granularities, where the small granularities are nested within larger ones
  - Can be represented graphically as a tree (but don't confuse with tree-locking protocol)
  - When a transaction locks a node in the tree explicitly, it implicitly locks all the node's descendents in the same mode.

- Granularity of locking (level in tree where locking is done):
  - fine granularity (lower in tree): high concurrency, high locking overhead
  - coarse granularity (higher in tree): low locking overhead, low concurrency

The highest level is the entire database and then area, file and record.
In addition to S and X lock modes, there are three additional lock modes with multiple granularity:

- **intention-shared** (IS): indicates explicit locking at a lower level of the tree but only with shared locks.

- **intention-exclusive** (IX): indicates explicit locking at a lower level with exclusive or shared locks

- **shared and intention-exclusive** (SIX): the subtree rooted by that node is locked explicitly in shared mode and explicit locking is being done at a lower level with exclusive-mode locks.

Intention locks allow a higher level node to be locked in S or X mode without having to check all descendant nodes.
The compatibility matrix for all lock modes is:

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>IX</th>
<th>S</th>
<th>S IX</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>IX</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>S</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>S IX</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>X</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Multiple Granularity Locking Scheme

Transaction $T_i$ can lock a node $Q$, using the following rules:

1. The lock compatibility matrix must be observed.
2. The root of the tree must be locked first, and may be locked in any mode.
3. A node $Q$ can be locked by $T_i$ in S or IS mode only if the parent of $Q$ is currently locked by $T_i$ in either IX or IS mode.
4. A node $Q$ can be locked by $T_i$ in X, SIX, or IX mode only if the parent of $Q$ is currently locked by $T_i$ in either IX or SIX mode.
5. $T_i$ can lock a node only if it has not previously unlocked any node (that is, $T_i$ is two-phase).
6. $T_i$ can unlock a node $Q$ only if none of the children of $Q$ are currently locked by $T_i$.

Observe that locks are acquired in root-to-leaf order, whereas they are released in leaf-to-root order.
Multiversion Schemes

- Multiversion schemes keep old versions of data item to increase concurrency.
  - Multiversion Timestamp Ordering
  - Multiversion Two-Phase Locking

- 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, and return the value of the selected version.
- reads 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 $Q_k$
- When a transaction $T_i$ creates a new version $Q_k$ of $Q$, $Q_k$'s W-timestamp and R-timestamp are initialized to $TS(T_i)$.
- R-timestamp of $Q_k$ is updated whenever a transaction $T_j$ reads $Q_k$, and $TS(T_j) > R$-timestamp($Q_k$).
Multiversion Timestamp Ordering

- The multiversion timestamp scheme presented next ensures serializability.

- Suppose that transaction \( T_i \) issues a \textit{read}(Q) or \textit{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 \( T_i \) issues a \textit{read}(Q), then the value returned is the content of version \( Q_k \).
  2. If transaction \( T_i \) issues a \textit{write}(Q),
     - if \( TS(T_i) < R\text{-timestamp}(Q_k) \), then transaction \( T_i \) is rolled back.
     - Otherwise if \( TS(T_i) = W\text{-timestamp}(Q_k) \), the contents of \( Q_k \) are overwritten, otherwise a new version of \( Q \) is created.

- Note that Reads always succeed

- But 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 \( T_i \).
Multiversion Two-Phase Locking

- Differentiates between read-only transactions and update transactions
- *Ts-counter* is a global time-stamp clock
- *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.
  - Each successful write results in the creation of a new version of the data item written.
  - Each version of a data item has a single timestamp whose value is obtained from a counter *ts-counter* that is incremented during commit processing.
- *Read-only transactions* are assigned a timestamp by reading the current value of *ts-counter* before they start execution
  - they follow the multiversion timestamp-ordering protocol for performing reads.
Multiversion Two-Phase Locking

- When an update transaction wants to read a data item,
  - it obtains a shared lock on it,
  - and reads the latest version.

- When an update transaction wants to write an item,
  - it obtains X lock on;
  - it then creates a new version of the item and sets this version's timestamp to $\infty$ (infinity).

- When an update transaction $T_i$ completes, commit processing occurs:
  - $T_i$ sets timestamp on the versions it has created to $\text{ts-counter} + 1$
  - $T_i$ increments $\text{ts-counter}$ by 1

- Read-only transactions that start after $T_i$ increments $\text{ts-counter}$ will see the values updated by $T_i$.

- Read-only transactions that start before $T_i$ increments the $\text{ts-counter}$ will see the value before the updates by $T_i$.

- Only serializable schedules are produced.
Deadlock Handling

- Consider the following two transactions:
  \[ T_1: \quad \text{write}(X) \quad T_2: \quad \text{write}(Y) \]
  \[ \text{write}(Y) \quad \text{write}(X) \]

- Schedule with deadlock

<table>
<thead>
<tr>
<th>( T_1 )</th>
<th>( T_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textbf{lock-X} on ( X ) \text{ write } (X) \</td>
<td>\textbf{lock-X} on ( Y ) \text{ write } (Y) \text{ wait for } \textbf{lock-X} \text{ on } X \text{ write}(X)</td>
</tr>
<tr>
<td>\text{wait for } \textbf{lock-X} \text{ on } Y \text{ write}(Y)</td>
<td></td>
</tr>
</tbody>
</table>
Deadlock Handling

- System is deadlocked if there is a set of transactions such that every transaction in the set is waiting for another transaction in the set.
- **Deadlock prevention protocols** ensure that the system will *never* enter into a deadlock state.
- Some prevention strategies:
  - Require that each transaction locks all its data items before it begins execution (predeclaration).
  - Impose **partial ordering of all data items**
    - require that a transaction can lock data items only in the order specified by the partial order (graph-based protocol).
  - **Timeout-Based Schemes**:
    - a transaction waits for a lock only for a specified amount of time.
      - After the wait time is out and the transaction is rolled back. (No deadlock!)
    - simple to implement; but **starvation is possible**
    - Also difficult to determine good value of the timeout interval.
  - Use **timestamping** (in the next slide)
Deadlock Prevention with Timestamps

- Following schemes use transaction timestamps for the sake of deadlock prevention alone.

- **Wait-die** scheme — non-preemptive
  - Older transaction may wait for younger one to release data item.
  - Younger transactions never wait for older ones; they are rolled back instead.
  - A transaction may die several times before acquiring needed data item.

- **Wound-wait** scheme — preemptive
  - Older transaction wounds (forces rollback of) younger transaction instead of waiting for it.
  - Younger transactions may wait for older ones.
  - May be fewer rollbacks than *wait-die* scheme.

- Both in *wait-die* and in *wound-wait* schemes, a rolled back transaction is restarted with its original timestamp.
  - Older transactions thus have precedence over newer ones in these schemes, and starvation is hence avoided.
Deadlock Detection

- Deadlocks can be described as a *wait-for graph*, which consists of a pair $G = (V,E)$,
  - $V$ is a set of vertices (all the transactions in the system)
  - $E$ is a set of edges; each element is an ordered pair $T_i \rightarrow T_j$.
- If $T_i \rightarrow T_j$ is in $E$, then there is a directed edge from $T_i$ to $T_j$
  - implying that $T_i$ is waiting for $T_j$ to release a data item.
- When $T_i$ requests a data item held by $T_j$, then $T_i \rightarrow T_j$ is inserted in the wait-for graph.
  - This edge is removed only when $T_j$ is no longer holding a data item needed by $T_i$.
- The system is in a deadlock state if and only if the wait-for graph has a cycle.
- The system invokes a deadlock-detection algorithm *periodically* to look for cycles.

**Wait-for graph without a cycle**

**Wait-for graph with a cycle**
Deadlock Recovery

- When deadlock is detected:
  - Some transaction will have to rolled back (made a victim) to break deadlock.
    - Select that transaction as **victim that will incur minimum cost**.
  - Rollback -- determine how far to roll back transaction
    - **Total rollback**: Abort the transaction and then restart it.
    - **Partial rollback**: More effective to roll back transaction only as far as necessary to break deadlock.
  - Starvation happens if same transaction is always chosen as victim.
    - The system may include the number of rollbacks in the cost factor to avoid starvation.