

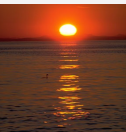


# Chapter 17: Recovery System

Version: Oct 5, 2006

Database System Concepts

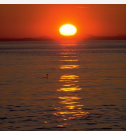
©Silberschatz, Korth and Sudarshan  
See [www.db-book.com](http://www.db-book.com) for conditions on re-use





# Chapter 17: Recovery System

- Failure Classification
- Storage Structure
- Recovery and Atomicity
- Log-Based Recovery
- Shadow Paging
- Recovery With Concurrent Transactions
- Buffer Management
- Failure with Loss of Nonvolatile Storage
- Advanced Recovery Techniques
- ARIES Recovery Algorithm
- Remote Backup Systems





# Failure Classification

## ■ Transaction failure :

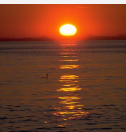
- **Logical errors**: transaction cannot complete due to some internal error condition
- **System errors**: the database system must terminate an active transaction due to an error condition (e.g., deadlock)

## ■ **System crash**: a power failure or other hardware or software failure causes the system to crash.

- **Fail-stop assumption**: non-volatile storage contents are assumed to not be corrupted by system crash
  - ▶ Database systems have numerous integrity checks to prevent corruption of disk data

## ■ **Disk failure**: a head crash or similar disk failure destroys all or part of disk storage

- Destruction is assumed to be detectable: disk drives use checksums to detect failures





# Recovery Algorithms

- Recovery algorithms are techniques to ensure database consistency and transaction atomicity and durability despite failures
  - Focus of this chapter
- Recovery algorithms have two parts
  1. Actions taken during normal transaction processing to ensure enough information exists to recover from failures
  2. Actions taken after a failure to recover the database contents to a state that ensures atomicity, consistency and durability





# Storage Structure

## ■ Volatile storage:

- does not survive system crashes
- examples: main memory, cache memory

## ■ Nonvolatile storage:

- survives system crashes
- examples: disk, tape, flash memory,  
non-volatile (battery backed up) RAM

## ■ Stable storage:

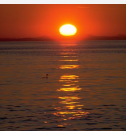
- a mythical form of storage that survives all failures
- approximated by maintaining multiple copies on distinct nonvolatile media





# Stable-Storage Implementation

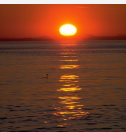
- Maintain multiple copies of each block on separate disks
  - copies can be at remote sites to protect against disasters such as fire or flooding.
- Failure during data transfer can still result in inconsistent copies: Block transfer can result in
  - Successful completion
  - Partial failure: destination block has incorrect information
  - Total failure: destination block was never updated
- Protecting storage media from failure during data transfer (one solution):
  - Execute output operation as follows (assuming two copies of each block):
    1. Write the information onto the first physical block.
    2. When the first write successfully completes, write the same information onto the second physical block.
    3. The output is completed only after the second write successfully completes.





# Stable-Storage Implementation (Cont.)

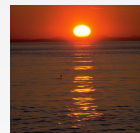
- Protecting storage media from failure during data transfer (cont.):
- Copies of a block may differ due to failure during output operation. To recover from failure:
  1. First find inconsistent blocks:
    1. *Expensive solution:* Compare the two copies of every disk block.
    2. *Better solution:*
      - Record in-progress disk writes on non-volatile storage (Non-volatile RAM or special area of disk).
      - Use this information during recovery to find blocks that may be inconsistent, and only compare copies of these.
      - Used in hardware RAID systems
  2. If either copy of an inconsistent block is detected to have an error (bad checksum), overwrite it by the other copy. If both have no error, but are different, overwrite the second block by the first block.





# Data Access

- **Physical blocks** are those blocks residing on the disk.
- **Buffer blocks** are the blocks residing temporarily in main memory.
- Block movements between disk and main memory are initiated through the following two operations:
  - **input**( $B$ ) transfers the physical block  $B$  to main memory.
  - **output**( $B$ ) transfers the buffer block  $B$  to the disk, and replaces the appropriate physical block there.
- Each transaction  $T_i$  has its private work-area in which local copies of all data items accessed and updated by it are kept.
  - $T_i$ 's local copy of a data item  $X$  is called  $x_i$ .
- We assume, for simplicity, that each data item fits in, and is stored inside, a single block.

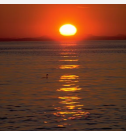






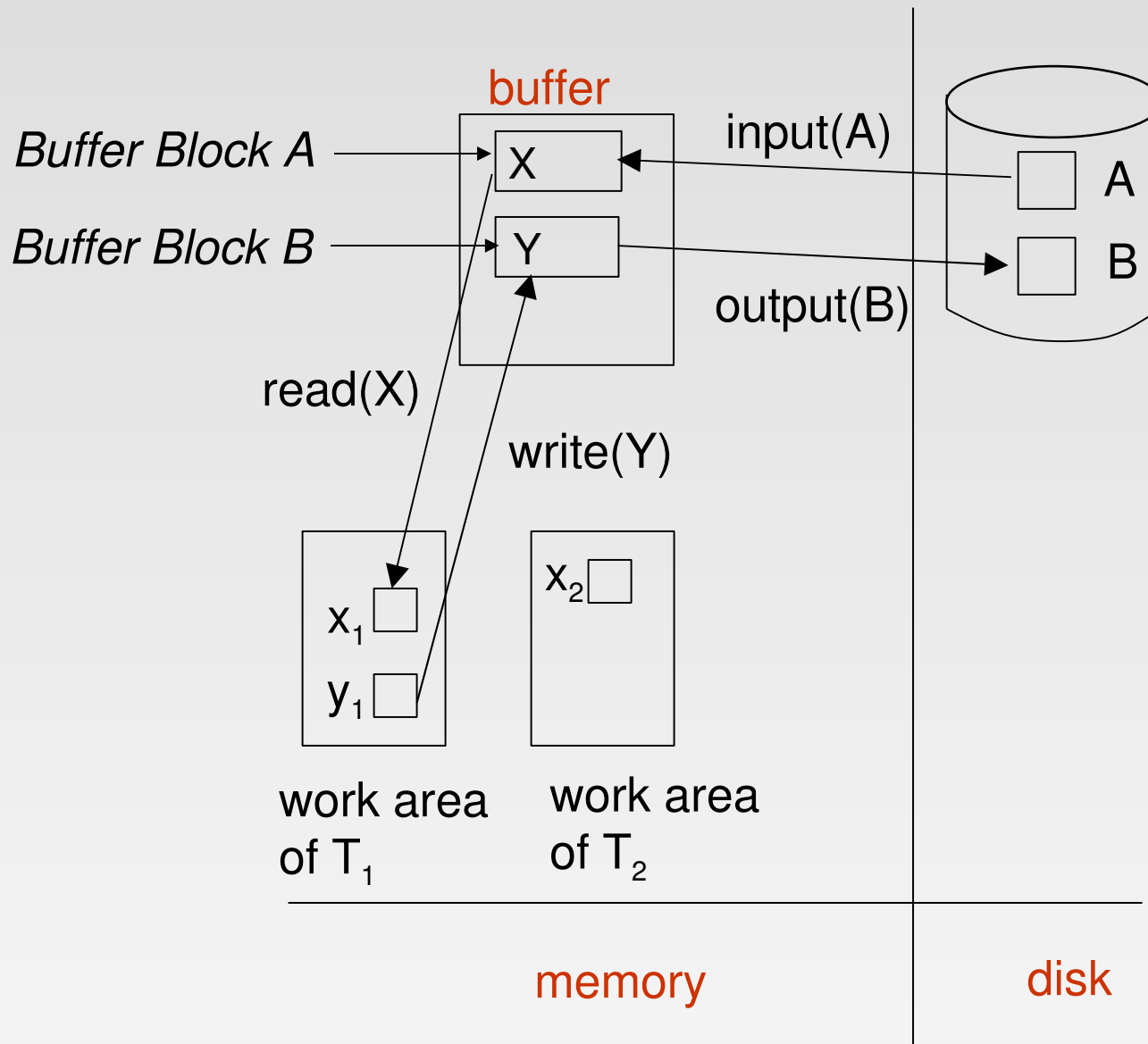
# Data Access (Cont.)

- Transaction transfers data items between system buffer blocks and its private work-area using the following operations :
  - **read**( $X$ ) assigns the value of data item  $X$  to the local variable  $x_i$ .
  - **write**( $X$ ) assigns the value of local variable  $x_i$  to data item  $\{X\}$  in the buffer block.
  - both these commands may necessitate the issue of an **input**( $B_x$ ) instruction before the assignment, if the block  $B_x$  in which  $X$  resides is not already in memory.
- Transactions
  - Perform **read**( $X$ ) while accessing  $X$  for the first time;
  - All subsequent accesses are to the local copy.
  - After last access, transaction executes **write**( $X$ ).
- **output**( $B_x$ ) need not immediately follow **write**( $X$ ). System can perform the **output** operation when it deems fit.





# Example of Data Access





# Recovery and Atomicity

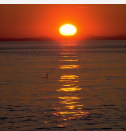
- Modifying the database without ensuring that the transaction will commit may leave the database in an inconsistent state.
- Consider transaction  $T_i$  that transfers \$50 from account  $A$  to account  $B$ ; goal is either to perform all database modifications made by  $T_i$  or none at all.
- Several output operations may be required for  $T_i$  (to output  $A$  and  $B$ ). A failure may occur after one of these modifications have been made but before all of them are made.





# Recovery and Atomicity (Cont.)

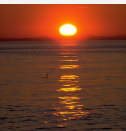
- To ensure atomicity despite failures, we first output information describing the modifications to stable storage without modifying the database itself.
- We study two approaches:
  - **log-based recovery**, and
  - **shadow-paging**
- We assume (initially) that transactions run serially, that is, one after the other.





# Log-Based Recovery

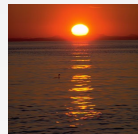
- A **log** is kept on stable storage.
  - The log is a sequence of **log records**, and maintains a record of update activities on the database.
- When transaction  $T_i$  starts, it registers itself by writing a  $\langle T_i \text{ start} \rangle$  log record
- Before  $T_i$  executes **write**( $X$ ), a log record  $\langle T_i, X, V_1, V_2 \rangle$  is written, where  $V_1$  is the value of  $X$  before the write, and  $V_2$  is the value to be written to  $X$ .
  - Log record notes that  $T_i$  has performed a write on data item  $X_j$ .  $X_j$  had value  $V_1$  before the write, and will have value  $V_2$  after the write.
- When  $T_i$  finishes its last statement, the log record  $\langle T_i \text{ commit} \rangle$  is written.
- We assume for now that log records are written directly to stable storage (that is, they are not buffered)
- Two approaches using logs
  - Deferred database modification
  - Immediate database modification





# Deferred Database Modification

- The **deferred database modification** scheme records all modifications to the log, but defers all the **writes** to after partial commit.
- Assume that transactions execute serially
- Transaction starts by writing  $\langle T_i, \text{start} \rangle$  record to log.
- A **write**( $X$ ) operation results in a log record  $\langle T_i, X, V \rangle$  being written, where  $V$  is the new value for  $X$ 
  - Note: old value is not needed for this scheme
- The write is not performed on  $X$  at this time, but is deferred.
- When  $T_i$  partially commits,  $\langle T_i, \text{commit} \rangle$  is written to the log
- Finally, the log records are read and used to actually execute the previously deferred writes.





# Deferred Database Modification (Cont.)

- During recovery after a crash, a transaction needs to be redone if and only if both  $\langle T_i \text{ start} \rangle$  and  $\langle T_i \text{ commit} \rangle$  are there in the log.
- Redoing a transaction  $T_i$  (**redo**  $T_i$ ) sets the value of all data items updated by the transaction to the new values.
- Crashes can occur while
  - the transaction is executing the original updates, or
  - while recovery action is being taken
- example transactions  $T_0$  and  $T_1$  ( $T_0$  executes before  $T_1$ ):

$T_0$ : **read** ( $A$ )

$A$ : -  $A - 50$

**Write** ( $A$ )

**read** ( $B$ )

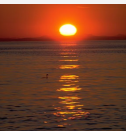
$B$ :-  $B + 50$

**write** ( $B$ )

$T_1$  : **read** ( $C$ )

$C$ :-  $C - 100$

**write** ( $C$ )



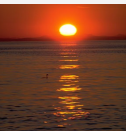


# Deferred Database Modification (Cont.)

- Below we show the log as it appears at three instances of time.

$\langle T_0 \text{ start} \rangle$	$\langle T_0 \text{ start} \rangle$	$\langle T_0 \text{ start} \rangle$
$\langle T_0, A, 950 \rangle$	$\langle T_0, A, 950 \rangle$	$\langle T_0, A, 950 \rangle$
$\langle T_0, B, 2050 \rangle$	$\langle T_0, B, 2050 \rangle$	$\langle T_0, B, 2050 \rangle$
	$\langle T_0 \text{ commit} \rangle$	$\langle T_0 \text{ commit} \rangle$
	$\langle T_1 \text{ start} \rangle$	$\langle T_1 \text{ start} \rangle$
	$\langle T_1, C, 600 \rangle$	$\langle T_1, C, 600 \rangle$
		$\langle T_1 \text{ commit} \rangle$
(a)	(b)	(c)

- If log on stable storage at time of crash is as in case:
  - (a) No redo actions need to be taken
  - (b) redo( $T_0$ ) must be performed since  $\langle T_0 \text{ commit} \rangle$  is present
  - (c) redo( $T_0$ ) must be performed followed by redo( $T_1$ ) since  $\langle T_0 \text{ commit} \rangle$  and  $\langle T_1 \text{ commit} \rangle$  are present

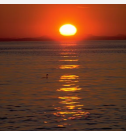






# Immediate Database Modification

- The **immediate database modification** scheme allows database updates of an uncommitted transaction to be made as the writes are issued
  - since undoing may be needed, update logs must have both old value and new value
- Update log record must be written *before* database item is written
  - We assume that the log record is output directly to stable storage
  - Can be extended to postpone log record output, so long as prior to execution of an **output**( $B$ ) operation for a data block  $B$ , all log records corresponding to items  $B$  must be flushed to stable storage
- Output of updated blocks can take place at any time before or after transaction commit
- Order in which blocks are output can be different from the order in which they are written.





# Immediate Database Modification Example

Log	Write	Output
<hr/>		
$\langle T_0 \text{ start} \rangle$		
$\langle T_0, A, 1000, 950 \rangle$		
$T_0, B, 2000, 2050$		
	$A = 950$	
	$B = 2050$	
$\langle T_0 \text{ commit} \rangle$		
$\langle T_1 \text{ start} \rangle$		
$\langle T_1, C, 700, \overset{x}{600} \rangle$		
	$C = 600$	
		$B_B, B_C$
$\langle T_1 \text{ commit} \rangle$		$B_A$

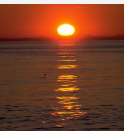
■ Note:  $B_x$  denotes block containing  $X$ .





# Immediate Database Modification (Cont.)

- Recovery procedure has two operations instead of one:
  - **undo**( $T_i$ ) restores the value of all data items updated by  $T_i$  to their old values, going backwards from the last log record for  $T_i$
  - **redo**( $T_i$ ) sets the value of all data items updated by  $T_i$  to the new values, going forward from the first log record for  $T_i$
- Both operations must be **idempotent**
  - That is, even if the operation is executed multiple times the effect is the same as if it is executed once
    - ▶ Needed since operations may get re-executed during recovery
- When recovering after failure:
  - Transaction  $T_i$  needs to be undone if the log contains the record  $\langle T_i \text{ start} \rangle$ , but does not contain the record  $\langle T_i \text{ commit} \rangle$ .
  - Transaction  $T_i$  needs to be redone if the log contains both the record  $\langle T_i \text{ start} \rangle$  and the record  $\langle T_i \text{ commit} \rangle$ .
- Undo operations are performed first, then redo operations.





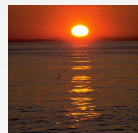
# Immediate DB Modification Recovery Example

Below we show the log as it appears at three instances of time.

$\langle T_0 \text{ start} \rangle$ $\langle T_0, A, 1000, 950 \rangle$ $\langle T_0, B, 2000, 2050 \rangle$	$\langle T_0 \text{ start} \rangle$ $\langle T_0, A, 1000, 950 \rangle$ $\langle T_0, B, 2000, 2050 \rangle$ $\langle T_0 \text{ commit} \rangle$ $\langle T_1 \text{ start} \rangle$ $\langle T_1, C, 700, 600 \rangle$	$\langle T_0 \text{ start} \rangle$ $\langle T_0, A, 1000, 950 \rangle$ $\langle T_0, B, 2000, 2050 \rangle$ $\langle T_0 \text{ commit} \rangle$ $\langle T_1 \text{ start} \rangle$ $\langle T_1, C, 700, 600 \rangle$ $\langle T_1 \text{ commit} \rangle$
(a)	(b)	(c)

Recovery actions in each case above are:

- (a) undo ( $T_0$ ): B is restored to 2000 and A to 1000.
- (b) undo ( $T_1$ ) and redo ( $T_0$ ): C is restored to 700, and then A and B are set to 950 and 2050 respectively.
- (c) redo ( $T_0$ ) and redo ( $T_1$ ): A and B are set to 950 and 2050 respectively. Then C is set to 600





# Checkpoints

- Problems in recovery procedure as discussed earlier :
  1. searching the entire log is time-consuming
  2. we might unnecessarily redo transactions which have already
  3. output their updates to the database.
- Streamline recovery procedure by periodically performing **checkpointing**
  1. Output all log records currently residing in main memory onto stable storage.
  2. Output all modified buffer blocks to the disk.
  3. Write a log record < **checkpoint** > onto stable storage.





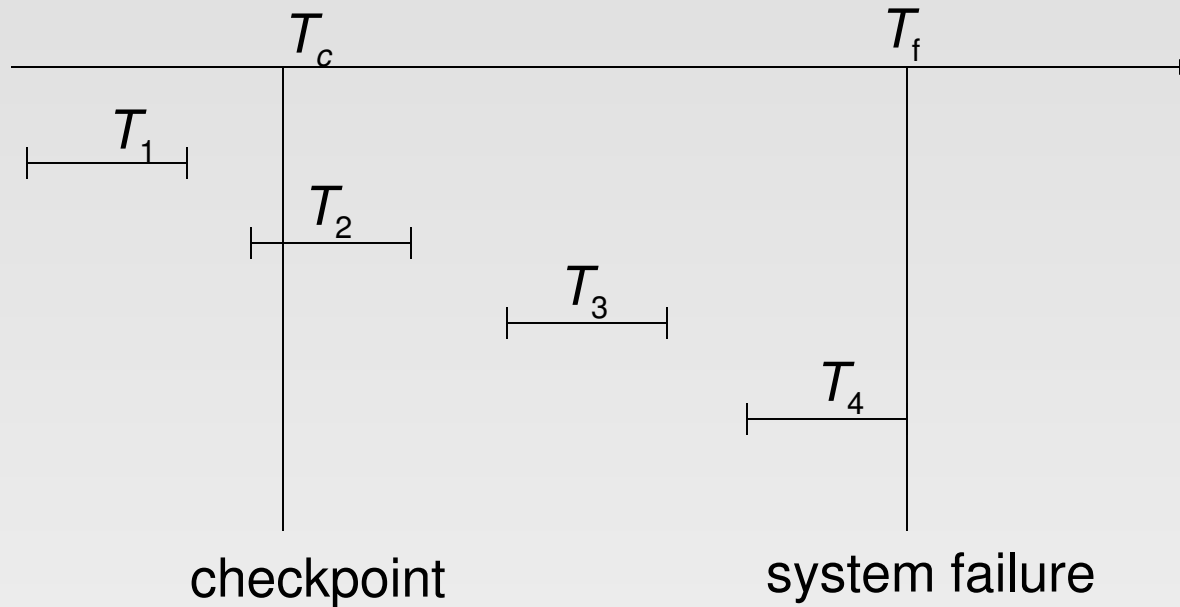
# Checkpoints (Cont.)

- During recovery we need to consider only the most recent transaction  $T_i$  that started before the checkpoint, and transactions that started after  $T_i$ .
  1. Scan backwards from end of log to find the most recent **<checkpoint>** record
  2. Continue scanning backwards till a record **< $T_i$  start>** is found.
  3. Need only consider the part of log following above **start** record. Earlier part of log can be ignored during recovery, and can be erased whenever desired.
  4. For all transactions (starting from  $T_i$  or later) with no **< $T_i$  commit>**, execute **undo( $T_i$ )**. (Done only in case of immediate modification.)
  5. Scanning forward in the log, for all transactions starting from  $T_i$  or later with a **< $T_i$  commit>**, execute **redo( $T_i$ )**.

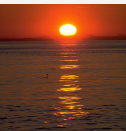




# Example of Checkpoints



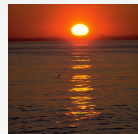
- $T_1$  can be ignored (updates already output to disk due to checkpoint)
- $T_2$  and  $T_3$  redone.
- $T_4$  undone





# Recovery With Concurrent Transactions

- We modify the log-based recovery schemes to allow multiple transactions to execute concurrently.
  - All transactions share a single disk buffer and a single log
  - A buffer block can have data items updated by one or more transactions
- We assume concurrency control using strict two-phase locking;
  - i.e. the updates of uncommitted transactions should not be visible to other transactions
    - ▶ Otherwise how to perform undo if T1 updates A, then T2 updates A and commits, and finally T1 has to abort?
- Logging is done as described earlier.
  - Log records of different transactions may be interspersed in the log.
- The checkpointing technique and actions taken on recovery have to be changed
  - since several transactions may be active when a checkpoint is performed.

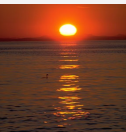






# Recovery With Concurrent Transactions (Cont.)

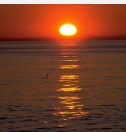
- Checkpoints are performed as before, except that the checkpoint log record is now of the form  
    < **checkpoint**  $L$  >  
where  $L$  is the list of transactions active at the time of the checkpoint
  - We assume no updates are in progress while the checkpoint is carried out (will relax this later)
- When the system recovers from a crash, it first does the following:
  1. Initialize *undo-list* and *redo-list* to empty
  2. Scan the log backwards from the end, stopping when the first <**checkpoint**  $L$ > record is found.  
For each record found during the backward scan:
    - ★ if the record is < $T_i$  **commit**>, add  $T_i$  to *redo-list*
    - ★ if the record is < $T_i$  **start**>, then if  $T_i$  is not in *redo-list*, add  $T_i$  to *undo-list*
  3. For every  $T_i$  in  $L$ , if  $T_i$  is not in *redo-list*, add  $T_i$  to *undo-list*





# Recovery With Concurrent Transactions (Cont.)

- At this point *undo-list* consists of incomplete transactions which must be undone, and *redo-list* consists of finished transactions that must be redone.
- Recovery now continues as follows:
  1. Scan log backwards from most recent record, stopping when  $\langle T_i \text{ start} \rangle$  records have been encountered for every  $T_i$  in *undo-list*.
    - During the scan, perform **undo** for each log record that belongs to a transaction in *undo-list*.
  2. Locate the most recent **<checkpoint L>** record.
  3. Scan log forwards from the **<checkpoint L>** record till the end of the log.
    - During the scan, perform **redo** for each log record that belongs to a transaction on *redo-list*





# Example of Recovery

- Go over the steps of the recovery algorithm on the following log:

$\langle T_0 \text{ start} \rangle$

$\langle T_0, A, 0, 10 \rangle$

$\langle T_0 \text{ commit} \rangle$

$\langle T_1 \text{ start} \rangle$  /\* Scan at step 1 comes up to here \*/

$\langle T_1, B, 0, 10 \rangle$

$\langle T_2 \text{ start} \rangle$

$\langle T_2, C, 0, 10 \rangle$

$\langle T_2, C, 10, 20 \rangle$

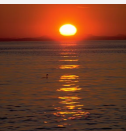
$\langle \text{checkpoint } \{T_1, T_2\} \rangle$

$\langle T_3 \text{ start} \rangle$

$\langle T_3, A, 10, 20 \rangle$

$\langle T_3, D, 0, 10 \rangle$

$\langle T_3 \text{ commit} \rangle$





# Log Record Buffering

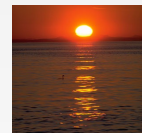
- **Log record buffering**: log records are buffered in main memory, instead of of being output directly to stable storage.
  - Log records are output to stable storage when a block of log records in the buffer is full, or a **log force** operation is executed.
- Log force is performed to commit a transaction by forcing all its log records (including the commit record) to stable storage.
- Several log records can thus be output using a single output operation, reducing the I/O cost.





# Log Record Buffering (Cont.)

- The rules below must be followed if log records are buffered:
  - Log records are output to stable storage in the order in which they are created.
  - Transaction  $T_i$  enters the commit state only when the log record  $\langle T_i, \text{commit} \rangle$  has been output to stable storage.
  - Before a block of data in main memory is output to the database, all log records pertaining to data in that block must have been output to stable storage.
    - ▶ This rule is called the **write-ahead logging** or **WAL** rule
      - Strictly speaking WAL only requires undo information to be output





# Database Buffering

- Database maintains an in-memory buffer of data blocks
  - When a new block is needed, if buffer is full an existing block needs to be removed from buffer
  - If the block chosen for removal has been updated, it must be output to disk
- If a block with uncommitted updates is output to disk, log records with undo information for the updates are output to the log on stable storage first
  - (Write ahead logging)
- No updates should be in progress on a block when it is output to disk. Can be ensured as follows.
  - Before writing a data item, transaction acquires exclusive lock on block containing the data item
  - Lock can be released once the write is completed.
    - ▶ Such locks held for short duration are called **latches**.
  - Before a block is output to disk, the system acquires an exclusive latch on the block
    - ▶ Ensures no update can be in progress on the block





# Buffer Management (Cont.)

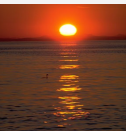
- Database buffer can be implemented either
  - in an area of real main-memory reserved for the database, or
  - in virtual memory
- Implementing buffer in reserved main-memory has drawbacks:
  - Memory is partitioned before-hand between database buffer and applications, limiting flexibility.
  - Needs may change, and although operating system knows best how memory should be divided up at any time, it cannot change the partitioning of memory.





# Buffer Management (Cont.)

- Database buffers are generally implemented in virtual memory in spite of some drawbacks:
    - When operating system needs to evict a page that has been modified, the page is written to swap space on disk.
    - When database decides to write buffer page to disk, buffer page may be in swap space, and may have to be read from swap space on disk and output to the database on disk, resulting in extra I/O!
      - ▶ Known as **dual paging** problem.
    - Ideally when OS needs to evict a page from the buffer, it should pass control to database, which in turn should
      1. Output the page to database instead of to swap space (making sure to output log records first), if it is modified
      2. Release the page from the buffer, for the OS to use
- Dual paging can thus be avoided, but common operating systems do not support such functionality.

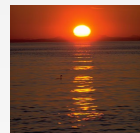






# Failure with Loss of Nonvolatile Storage

- So far we assumed no loss of non-volatile storage
- Technique similar to checkpointing used to deal with loss of non-volatile storage
  - Periodically **dump** the entire content of the database to stable storage
  - No transaction may be active during the dump procedure; a procedure similar to checkpointing must take place
    - ▶ Output all log records currently residing in main memory onto stable storage.
    - ▶ Output all buffer blocks onto the disk.
    - ▶ Copy the contents of the database to stable storage.
    - ▶ Output a record **<dump>** to log on stable storage.





# Recovering from Failure of Non-Volatile Storage

- To recover from disk failure
  - restore database from most recent dump.
  - Consult the log and redo all transactions that committed after the dump
- Can be extended to allow transactions to be active during dump; known as **fuzzy dump** or **online dump**
  - Will study fuzzy checkpointing later





# **Advanced Recovery Algorithm**

**Database System Concepts**

©Silberschatz, Korth and Sudarshan

See [www.db-book.com](http://www.db-book.com) for conditions on re-use





# Advanced Recovery: Key Features

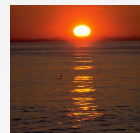
- Support for high-concurrency locking techniques, such as those used for B<sup>+</sup>-tree concurrency control, which release locks early
  - Supports “logical undo”
- Recovery based on “**repeating history**”, whereby recovery executes exactly the same actions as normal processing
  - including redo of log records of incomplete transactions, followed by subsequent undo
  - Key benefits
    - ▶ supports logical undo
    - ▶ easier to understand/show correctness





# Advanced Recovery: Logical Undo Logging

- Operations like B<sup>+</sup>-tree insertions and deletions release locks early.
  - They cannot be undone by restoring old values (**physical undo**), since once a lock is released, other transactions may have updated the B<sup>+</sup>-tree.
  - Instead, insertions (resp. deletions) are undone by executing a deletion (resp. insertion) operation (known as **logical undo**).
- For such operations, undo log records should contain the undo operation to be executed
  - Such logging is called **logical undo logging**, in contrast to **physical undo logging**
    - ▶ Operations are called **logical operations**
  - Other examples:
    - ▶ delete of tuple, to undo insert of tuple
      - allows early lock release on space allocation information
    - ▶ subtract amount deposited, to undo deposit
      - allows early lock release on bank balance





# Advanced Recovery: Physical Redo

- Redo information is logged **physically** (that is, new value for each write) even for operations with logical undo
  - Logical redo is very complicated since database state on disk may not be “operation consistent” when recovery starts
  - Physical redo logging does not conflict with early lock release





# Advanced Recovery: Operation Logging

■ Operation logging is done as follows:

1. When operation starts, log  $\langle T_i, O_j, \text{operation-begin} \rangle$ . Here  $O_j$  is a unique identifier of the operation instance.
2. While operation is executing, normal log records with physical redo and physical undo information are logged.
3. When operation completes,  $\langle T_i, O_j, \text{operation-end}, U \rangle$  is logged, where  $U$  contains information needed to perform a logical undo information.

Example: insert of (key, record-id) pair (K5, RID7) into index I9

$\langle T1, O1, \text{operation-begin} \rangle$

....

$\langle T1, X, 10, K5 \rangle$

$\langle T1, Y, 45, \text{RID7} \rangle$

$\langle T1, O1, \text{operation-end}, (\text{delete I9, K5, RID7}) \rangle$

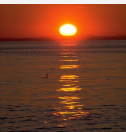
} Physical redo of steps in insert





# Advanced Recovery: Operation Logging (Cont.)

- If crash/rollback occurs before operation completes:
  - the **operation-end** log record is not found, and
  - the physical undo information is used to undo operation.
- If crash/rollback occurs after the operation completes:
  - the **operation-end** log record is found, and in this case
  - logical undo is performed using  $U$ ; the physical undo information for the operation is ignored.
- Redo of operation (after crash) still uses physical redo information.





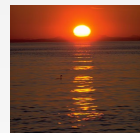


# Advanced Recovery: Txn Rollback

Rollback of transaction  $T_i$  is done as follows:

- Scan the log backwards

1. If a log record  $\langle T_i, X, V_1, V_2 \rangle$  is found, perform the undo and log a special **redo-only log record**  $\langle T_i, X, V_1 \rangle$ .
2. If a  $\langle T_i, O_j, \text{operation-end}, U \rangle$  record is found
  - ▶ Rollback the operation logically using the undo information  $U$ .
    - Updates performed during roll back are logged just like during normal operation execution.
    - At the end of the operation rollback, instead of logging an **operation-end** record, generate a record  $\langle T_i, O_j, \text{operation-abort} \rangle$ .
  - ▶ Skip all preceding log records for  $T_i$  until the record  $\langle T_i, O_j, \text{operation-begin} \rangle$  is found



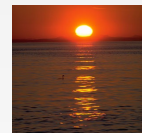


# Advanced Recovery: Txn Rollback (Cont.)

- Scan the log backwards (cont.):
  1. If a redo-only record is found ignore it
  2. If a  $\langle T_i, O_j, \text{operation-abort} \rangle$  record is found:
    - ★ skip all preceding log records for  $T_i$  until the record  $\langle T_i, O_j, \text{operation-begin} \rangle$  is found.
  3. Stop the scan when the record  $\langle T_i, \text{start} \rangle$  is found
  4. Add a  $\langle T_i, \text{abort} \rangle$  record to the log

Some points to note:

- Cases 3 and 4 above can occur only if the database crashes while a transaction is being rolled back.
- Skipping of log records as in case 4 is important to prevent multiple rollback of the same operation.





# Advanced Recovery: Txn Rollback Example

- Example with a complete and an incomplete operation

<T1, start>

<T1, O1, operation-begin>

....

<T1, X, 10, K5>

<T1, Y, 45, RID7>

<T1, O1, operation-end, (delete I9, K5, RID7)>

<T1, O2, operation-begin>

<T1, Z, 45, 70>

← T1 Rollback begins here

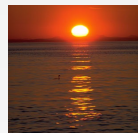
<T1, Z, 45> ← redo-only log record during physical undo (of incomplete O2)

<T1, Y, ..., ...> ← Normal redo records for logical undo of O1

...

<T1, O1, operation-abort> ← What if crash occurred immediately after this?

<T1, abort>





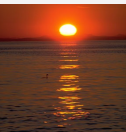
# Advanced Recovery: Crash Recovery

The following actions are taken when recovering from system crash

2. (**Redo phase**): Scan log forward from last < **checkpoint**  $L$  > record till end of log
  1. **Repeat history** by physically redoing all updates of all transactions,
  2. Create an undo-list during the scan as follows
    - ▶ *undo-list* is set to  $L$  initially
    - ▶ Whenever <  $T_i$  **start** > is found  $T_i$  is added to *undo-list*
    - ▶ Whenever <  $T_i$  **commit** > or <  $T_i$  **abort** > is found,  $T_i$  is deleted from *undo-list*

This brings database to state as of crash, with committed as well as uncommitted transactions having been redone.

Now *undo-list* contains transactions that are **incomplete**, that is, have neither committed nor been fully rolled back.

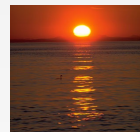




# Advanced Recovery: Crash Recovery (Cont.)

Recovery from system crash (cont.)

2. (**Undo phase**): Scan log backwards, performing undo on log records of transactions found in *undo-list*.
  - Log records of transactions being rolled back are processed as described earlier, as they are found
    - ▶ Single shared scan for all transactions being undone
  - When  $\langle T_i \text{ start} \rangle$  is found for a transaction  $T_i$  in *undo-list*, write a  $\langle T_i \text{ abort} \rangle$  log record.
  - Stop scan when  $\langle T_i \text{ start} \rangle$  records have been found for all  $T_i$  in *undo-list*
- This undoes the effects of incomplete transactions (those with neither **commit** nor **abort** log records). Recovery is now complete.





# Advanced Recovery: Checkpointing

- **Checkpointing** is done as follows:
  1. Output all log records in memory to stable storage
  2. Output to disk all modified buffer blocks
  3. Output to log on stable storage a < **checkpoint**  $L$  > record.

Transactions are not allowed to perform any actions while checkpointing is in progress.

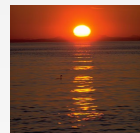
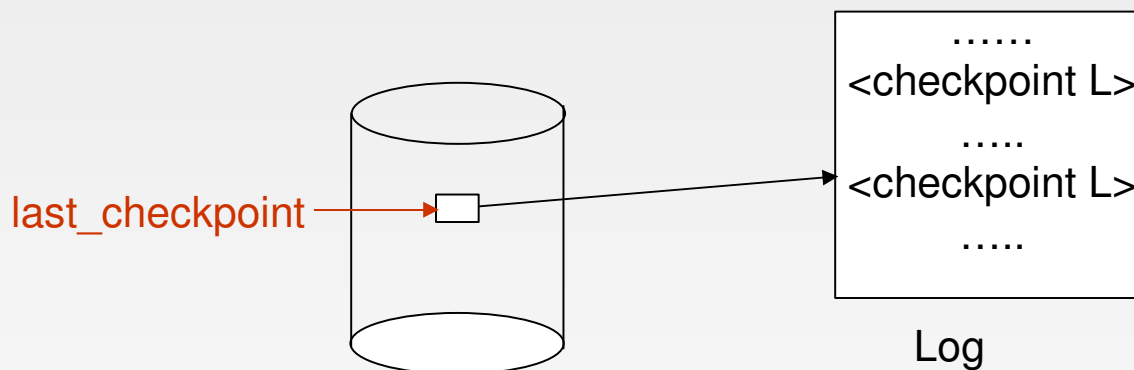
- Fuzzy checkpointing allows transactions to progress while the most time consuming parts of checkpointing are in progress
  - Performed as described on next slide





# Advanced Recovery: Fuzzy Checkpointing

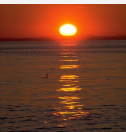
- **Fuzzy checkpointing** is done as follows:
  1. Temporarily stop all updates by transactions
  2. Write a **<checkpoint L>** log record and force log to stable storage
  3. Note list *M* of modified buffer blocks
  4. Now permit transactions to proceed with their actions
  5. Output to disk all modified buffer blocks in list *M*
    - ★ blocks should not be updated while being output
    - ★ Follow WAL: all log records pertaining to a block must be output before the block is output
  6. Store a pointer to the **checkpoint** record in a fixed position **last\_checkpoint** on disk





# Advanced Rec: Fuzzy Checkpointing (Cont.)

- When recovering using a fuzzy checkpoint, start scan from the **checkpoint** record pointed to by **last\_checkpoint**
  - Log records before **last\_checkpoint** have their updates reflected in database on disk, and need not be redone.
  - Incomplete checkpoints, where system had crashed while performing checkpoint, are handled safely







# **ARIES Recovery Algorithm**

**Database System Concepts**

©Silberschatz, Korth and Sudarshan

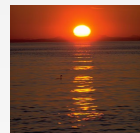
See [www.db-book.com](http://www.db-book.com) for conditions on re-use





# ARIES

- ARIES is a state of the art recovery method
  - Incorporates numerous optimizations to reduce overheads during normal processing and to speed up recovery
  - The “advanced recovery algorithm” we studied earlier is modeled after ARIES, but greatly simplified by removing optimizations
- Unlike the advanced recovery algorithm, ARIES
  1. Uses **log sequence number (LSN)** to identify log records
    - ▶ Stores LSNs in pages to identify what updates have already been applied to a database page
  2. Physiological redo
  3. Dirty page table to avoid unnecessary redos during recovery
  4. Fuzzy checkpointing that only records information about dirty pages, and does not require dirty pages to be written out at checkpoint time
    - ▶ More coming up on each of the above ...

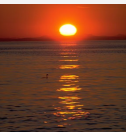




# ARIES Optimizations

## ■ Physiological redo

- Affected page is physically identified, action within page can be logical
  - ▶ Used to reduce logging overheads
    - e.g. when a record is deleted and all other records have to be moved to fill hole
      - » Physiological redo can log just the record deletion
      - » Physical redo would require logging of old and new values for much of the page
  - ▶ Requires page to be output to disk atomically
    - Easy to achieve with hardware RAID, also supported by some disk systems
    - Incomplete page output can be detected by checksum techniques,
      - » But extra actions are required for recovery
      - » Treated as a media failure





# ARIES Data Structures

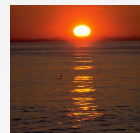
- ARIES uses several data structures
  - Log sequence number (LSN) identifies each log record
    - ▶ Must be sequentially increasing
    - ▶ Typically an offset from beginning of log file to allow fast access
      - Easily extended to handle multiple log files
  - Page LSN
  - Log records of several different types
  - Dirty page table





# ARIES Data Structures: Page LSN

- Each page contains a **PageLSN** which is the LSN of the last log record whose effects are reflected on the page
  - To update a page:
    - ▶ X-latch the page, and write the log record
    - ▶ Update the page
    - ▶ Record the LSN of the log record in PageLSN
    - ▶ Unlock page
  - To flush page to disk, must first S-latch page
    - ▶ Thus page state on disk is operation consistent
      - Required to support physiological redo
  - PageLSN is used during recovery to prevent repeated redo
    - ▶ Thus ensuring idempotence





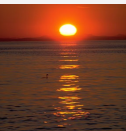
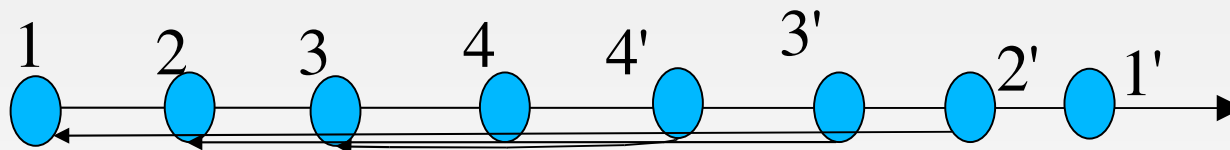
# ARIES Data Structures: Log Record

- Each log record contains LSN of previous log record of the same transaction

LSN	TransID	PrevLSN	RedoInfo	UndoInfo
-----	---------	---------	----------	----------

- LSN in log record may be implicit
- Special redo-only log record called **compensation log record (CLR)** used to log actions taken during recovery that never need to be undone
  - Serves the role of operation-abort log records used in advanced recovery algorithm
  - Has a field UndoNextLSN to note next (earlier) record to be undone
    - Records in between would have already been undone
    - Required to avoid repeated undo of already undone actions

LSN	TransID	UndoNextLSN	RedoInfo
-----	---------	-------------	----------



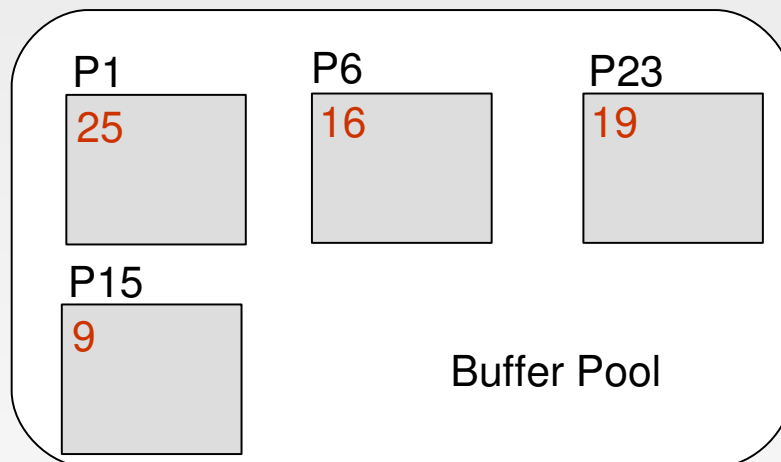
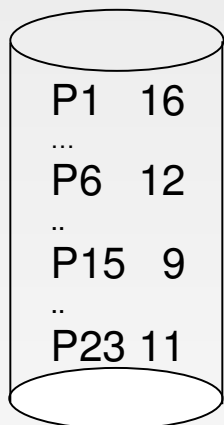


# ARIES Data Structures: DirtyPage Table

## ■ DirtyPageTable

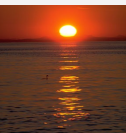
- List of pages in the buffer that have been updated
- Contains, for each such page
  - ▶ **PageLSN** of the page
  - ▶ **RecLSN** is an LSN such that log records before this LSN have already been applied to the page version on disk
    - Set to current end of log when a page is inserted into dirty page table (just before being updated)
    - Recorded in checkpoints, helps to minimize redo work

Page LSNs  
on disk



Page	PLSN	RLSN
P1	25	17
P6	16	15
P23	19	18

DirtyPage Table





# ARIES Data Structures: Checkpoint Log

## ■ Checkpoint log record

- Contains:

- ▶ DirtyPageTable and list of active transactions
- ▶ For each active transaction, LastLSN, the LSN of the last log record written by the transaction

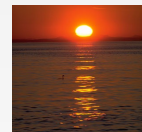
- Fixed position on disk notes LSN of last completed checkpoint log record

## ■ Dirty pages are not written out at checkpoint time

- ▶ Instead, they are flushed out continuously, in the background

## ■ Checkpoint is thus very low overhead

- can be done frequently







# ARIES Recovery Algorithm

ARIES recovery involves three passes

■ **Analysis pass:** Determines

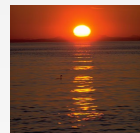
- Which transactions to undo
- Which pages were dirty (disk version not up to date) at time of crash
- **RedoLSN**: LSN from which redo should start

■ **Redo pass:**

- Repeats history, redoing all actions from RedoLSN
  - ▶ RecLSN and PageLSNs are used to avoid redoing actions already reflected on page

■ **Undo pass:**

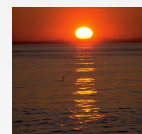
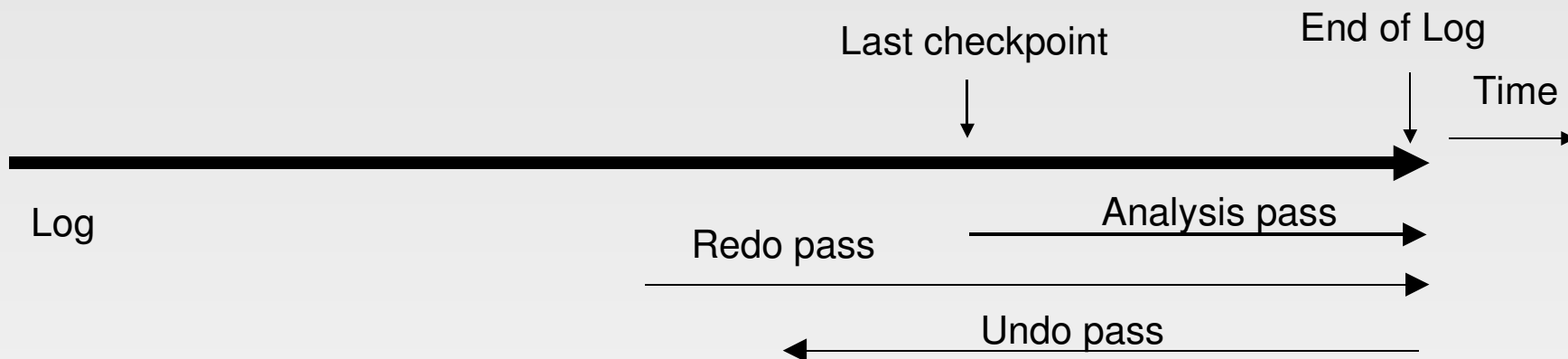
- Rolls back all incomplete transactions
  - ▶ Transactions whose abort was complete earlier are not undone
    - Key idea: no need to undo these transactions: earlier undo actions were logged, and are redone as required





# Aries Recovery: 3 Passes

- Analysis, redo and undo passes
- Analysis determines where redo should start
- Undo has to go back till start of earliest incomplete transaction

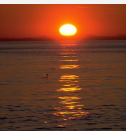




# ARIES Recovery: Analysis

## Analysis pass

- Starts from last complete checkpoint log record
  - Reads DirtyPageTable from log record
  - Sets RedoLSN = min of RecLSNs of all pages in DirtyPageTable
    - ▶ In case no pages are dirty, RedoLSN = checkpoint record's LSN
  - Sets undo-list = list of transactions in checkpoint log record
  - Reads LSN of last log record for each transaction in undo-list from checkpoint log record
- Scans forward from checkpoint
- .. Cont. on next page ...

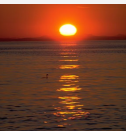




# ARIES Recovery: Analysis (Cont.)

## Analysis pass (cont.)

- Scans forward from checkpoint
  - If any log record found for transaction not in undo-list, adds transaction to undo-list
  - Whenever an update log record is found
    - ▶ If page is not in DirtyPageTable, it is added with RecLSN set to LSN of the update log record
  - If transaction end log record found, delete transaction from undo-list
  - Keeps track of last log record for each transaction in undo-list
    - ▶ May be needed for later undo
- At end of analysis pass:
  - RedoLSN determines where to start redo pass
  - RecLSN for each page in DirtyPageTable used to minimize redo work
  - All transactions in undo-list need to be rolled back



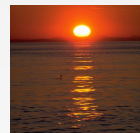


# ARIES Redo Pass

**Redo Pass:** Repeats history by replaying every action not already reflected in the page on disk, as follows:

- Scans forward from RedoLSN. Whenever an update log record is found:
  1. If the page is not in DirtyPageTable or the LSN of the log record is less than the RecLSN of the page in DirtyPageTable, then skip the log record
  2. Otherwise fetch the page from disk. If the PageLSN of the page fetched from disk is less than the LSN of the log record, redo the log record

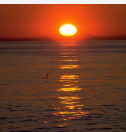
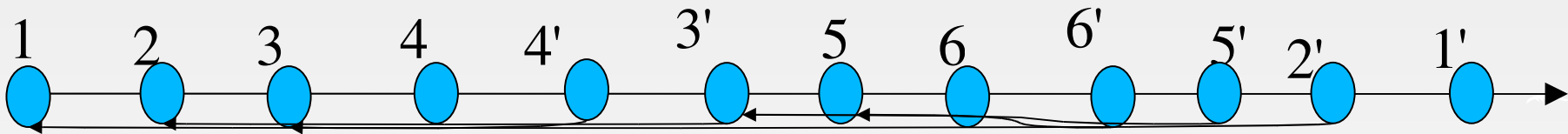
NOTE: if either test is negative the effects of the log record have already appeared on the page. First test avoids even fetching the page from disk!





# ARIES Undo Actions

- When an undo is performed for an update log record
  - Generate a CLR containing the undo action performed (actions performed during undo are logged physically or physiologically).
    - ▶ CLR for record  $n$  noted as  $n'$  in figure below
  - Set UndoNextLSN of the CLR to the PrevLSN value of the update log record
    - ▶ Arrows indicate UndoNextLSN value
- ARIES supports partial rollback
  - Used e.g. to handle deadlocks by rolling back just enough to release reqd. locks
  - Figure indicates forward actions after partial rollbacks
    - ▶ records 3 and 4 initially, later 5 and 6, then full rollback

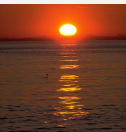




# ARIES: Undo Pass

## Undo pass:

- Performs backward scan on log undoing all transaction in undo-list
  - Backward scan optimized by skipping unneeded log records as follows:
    - ▶ Next LSN to be undone for each transaction set to LSN of last log record for transaction found by analysis pass.
    - ▶ At each step pick largest of these LSNs to undo, skip back to it and undo it
    - ▶ After undoing a log record
      - For ordinary log records, set next LSN to be undone for transaction to PrevLSN noted in the log record
      - For compensation log records (CLRs) set next LSN to be undo to UndoNextLSN noted in the log record
        - » All intervening records are skipped since they would have been undone already
- Undos performed as described earlier





# Other ARIES Features

- Recovery Independence
  - Pages can be recovered independently of others
    - ▶ E.g. if some disk pages fail they can be recovered from a backup while other pages are being used
- Savepoints:
  - Transactions can record savepoints and roll back to a savepoint
    - ▶ Useful for complex transactions
    - ▶ Also used to rollback just enough to release locks on deadlock

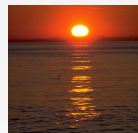






# Other ARIES Features (Cont.)

- Fine-grained locking:
  - Index concurrency algorithms that permit tuple level locking on indices can be used
    - ▶ These require logical undo, rather than physical undo, as in advanced recovery algorithm
- Recovery optimizations: For example:
  - Dirty page table can be used to prefetch pages during redo
  - Out of order redo is possible:
    - ▶ redo can be postponed on a page being fetched from disk, and performed when page is fetched.
    - ▶ Meanwhile other log records can continue to be processed





# Remote Backup Systems

**Database System Concepts**

©Silberschatz, Korth and Sudarshan

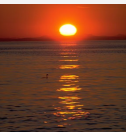
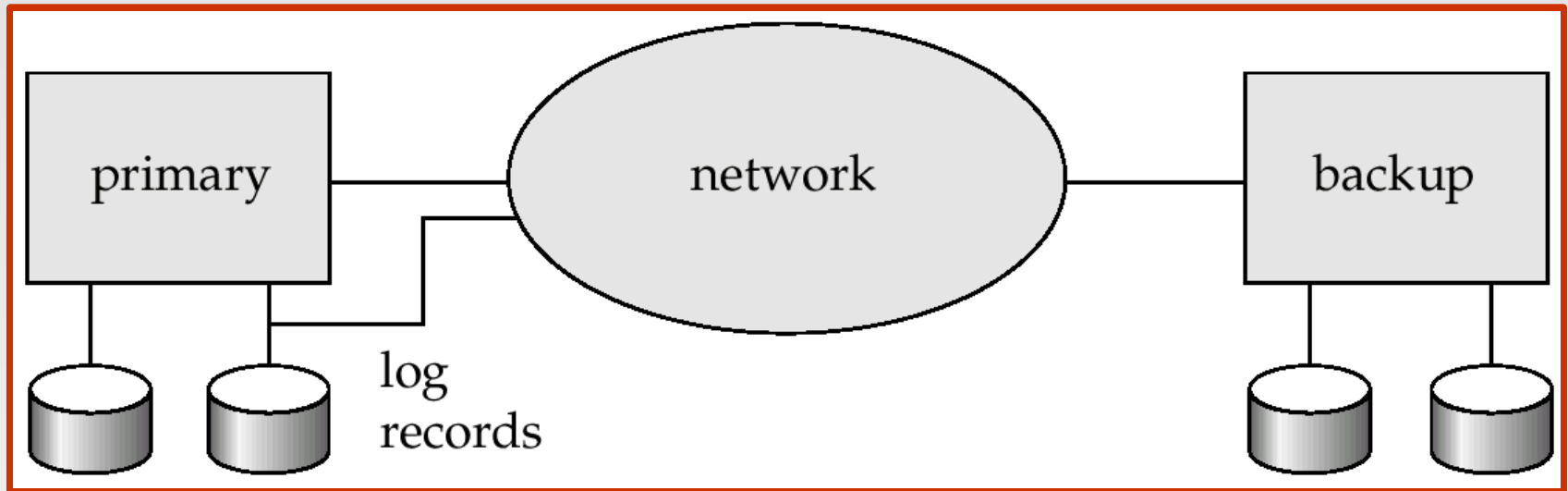
See [www.db-book.com](http://www.db-book.com) for conditions on re-use





# Remote Backup Systems

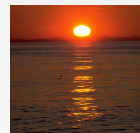
- Remote backup systems provide high availability by allowing transaction processing to continue even if the primary site is destroyed.





# Remote Backup Systems (Cont.)

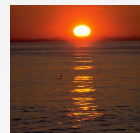
- **Detection of failure:** Backup site must detect when primary site has failed
  - to distinguish primary site failure from link failure maintain several communication links between the primary and the remote backup.
  - Heart-beat messages
- **Transfer of control:**
  - To take over control backup site first perform recovery using its copy of the database and all the long records it has received from the primary.
    - ▶ Thus, completed transactions are redone and incomplete transactions are rolled back.
  - When the backup site takes over processing it becomes the new primary
  - To transfer control back to old primary when it recovers, old primary must receive redo logs from the old backup and apply all updates locally.





# Remote Backup Systems (Cont.)

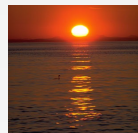
- **Time to recover:** To reduce delay in takeover, backup site periodically processes the redo log records (in effect, performing recovery from previous database state), performs a checkpoint, and can then delete earlier parts of the log.
- **Hot-Spare** configuration permits very fast takeover:
  - Backup continually processes redo log record as they arrive, applying the updates locally.
  - When failure of the primary is detected the backup rolls back incomplete transactions, and is ready to process new transactions.
- Alternative to remote backup: distributed database with replicated data
  - Remote backup is faster and cheaper, but less tolerant to failure
    - ▶ more on this in Chapter 19





# Remote Backup Systems (Cont.)

- Ensure durability of updates by delaying transaction commit until update is logged at backup; avoid this delay by permitting lower degrees of durability.
- **One-safe:** commit as soon as transaction's commit log record is written at primary
  - Problem: updates may not arrive at backup before it takes over.
- **Two-very-safe:** commit when transaction's commit log record is written at primary and backup
  - Reduces availability since transactions cannot commit if either site fails.
- **Two-safe:** proceed as in two-very-safe if both primary and backup are active. If only the primary is active, the transaction commits as soon as its commit log record is written at the primary.
  - Better availability than two-very-safe; avoids problem of lost transactions in one-safe.





# End of Chapter

**Database System Concepts**

©Silberschatz, Korth and Sudarshan

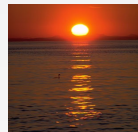
See [www.db-book.com](http://www.db-book.com) for conditions on re-use





# Shadow Paging

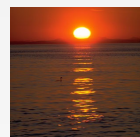
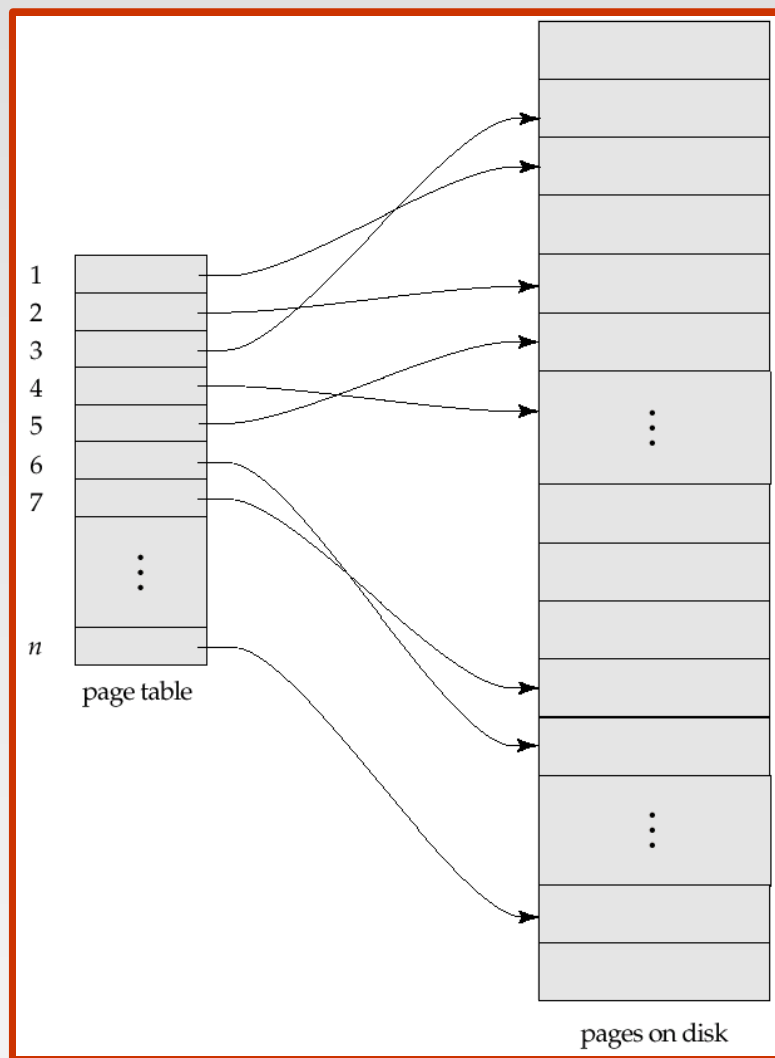
- **Shadow paging** is an alternative to log-based recovery; this scheme is useful if transactions execute serially
- Idea: maintain *two* page tables during the lifetime of a transaction –the **current page table**, and the **shadow page table**
- Store the shadow page table in nonvolatile storage, such that state of the database prior to transaction execution may be recovered.
  - Shadow page table is never modified during execution
- To start with, both the page tables are identical. Only current page table is used for data item accesses during execution of the transaction.
- Whenever any page is about to be written for the first time
  - A copy of this page is made onto an unused page.
  - The current page table is then made to point to the copy
  - The update is performed on the copy







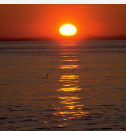
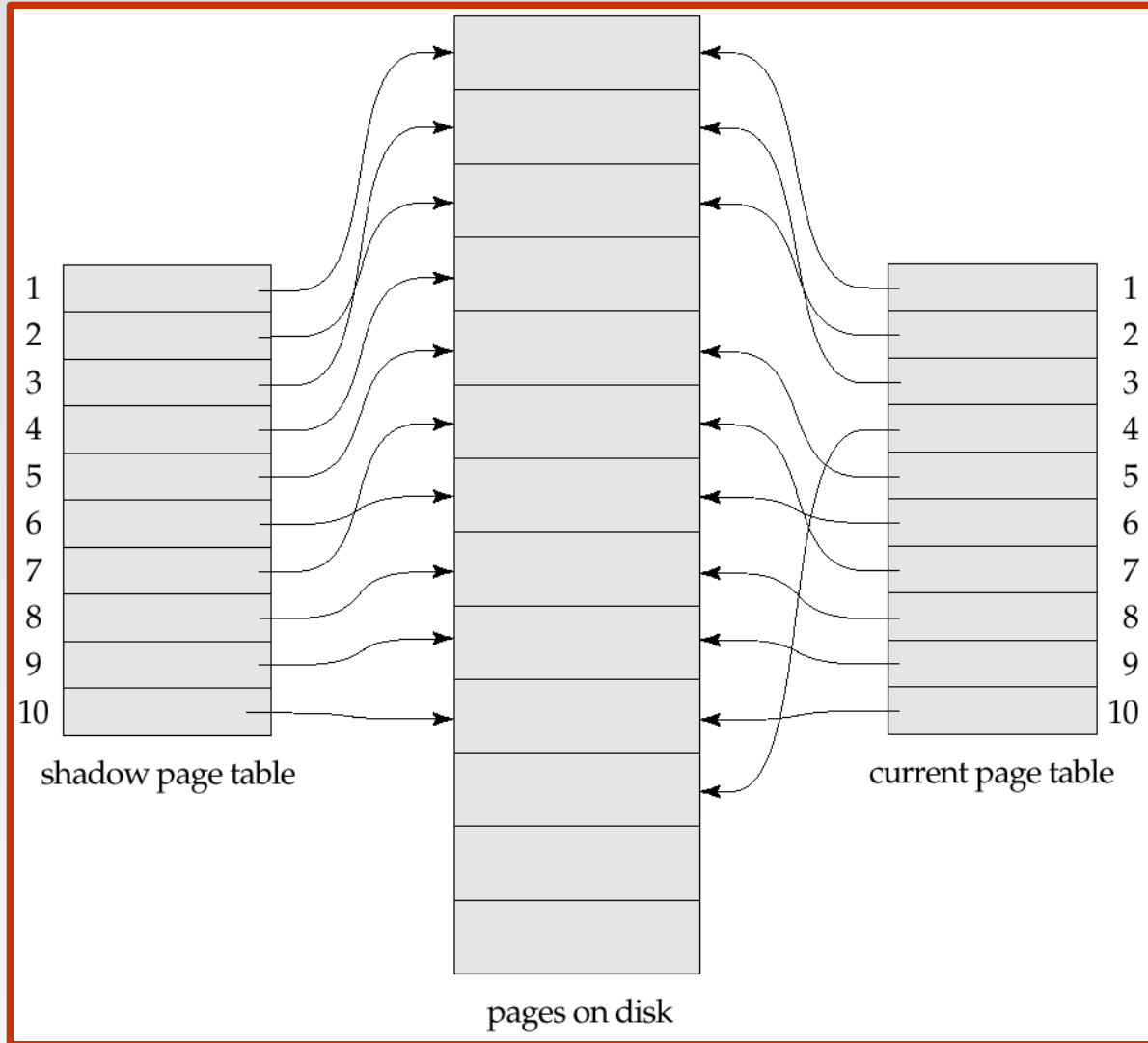
# Sample Page Table





# Example of Shadow Paging

Shadow and current page tables after write to page 4





# Shadow Paging (Cont.)

- To commit a transaction :
  1. Flush all modified pages in main memory to disk
  2. Output current page table to disk
  3. Make the current page table the new shadow page table, as follows:
    - keep a pointer to the shadow page table at a fixed (known) location on disk.
    - to make the current page table the new shadow page table, simply update the pointer to point to current page table on disk
- Once pointer to shadow page table has been written, transaction is committed.
- No recovery is needed after a crash — new transactions can start right away, using the shadow page table.
- Pages not pointed to from current/shadow page table should be freed (garbage collected).





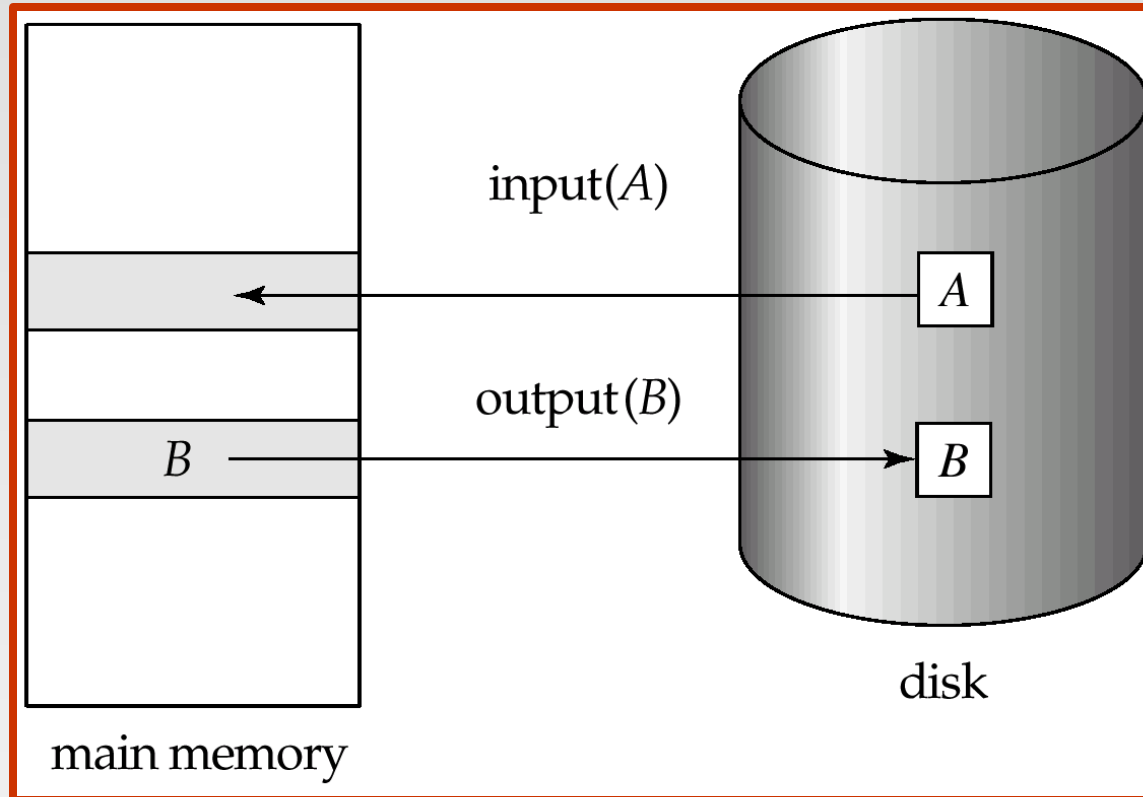
# Show Paging (Cont.)

- Advantages of shadow-paging over log-based schemes
  - no overhead of writing log records
  - recovery is trivial
- Disadvantages :
  - Copying the entire page table is very expensive
    - ▶ Can be reduced by using a page table structured like a B<sup>+</sup>-tree
      - No need to copy entire tree, only need to copy paths in the tree that lead to updated leaf nodes
  - Commit overhead is high even with above extension
    - ▶ Need to flush every updated page, and page table
  - Data gets fragmented (related pages get separated on disk)
  - After every transaction completion, the database pages containing old versions of modified data need to be garbage collected
  - Hard to extend algorithm to allow transactions to run concurrently
    - ▶ Easier to extend log based schemes





# Block Storage Operations





# Portion of the Database Log Corresponding to $T_0$ and $T_1$

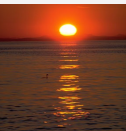
$\langle T_0 \text{ start} \rangle$   
 $\langle T_0, A, 950 \rangle$   
 $\langle T_0, B, 2050 \rangle$   
 $\langle T_0 \text{ commit} \rangle$   
 $\langle T_1 \text{ start} \rangle$   
 $\langle T_1, C, 600 \rangle$   
 $\langle T_1 \text{ commit} \rangle$





# State of the Log and Database Corresponding to $T_0$ and $T_1$

Log	Database
$\langle T_0 \text{ start} \rangle$	
$\langle T_0, A, 950 \rangle$	
$\langle T_0, B, 2050 \rangle$	
$\langle T_0 \text{ commit} \rangle$	$A = 950$ $B = 2050$
$\langle T_1 \text{ start} \rangle$	
$\langle T_1, C, 600 \rangle$	
$\langle T_1 \text{ commit} \rangle$	$C = 600$





# Portion of the System Log Corresponding to $T_0$ and $T_1$

$\langle T_0 \text{ start} \rangle$

$\langle T_0, A, 1000, 950 \rangle$

$\langle T_0, B, 2000, 2050 \rangle$

$\langle T_0 \text{ commit} \rangle$

$\langle T_1 \text{ start} \rangle$

$\langle T_1, C, 700, 600 \rangle$

$\langle T_1 \text{ commit} \rangle$







# State of System Log and Database Corresponding to $T_0$ and $T_1$

Log	Database
$\langle T_0 \text{ start} \rangle$	
$\langle T_0, A, 1000, 950 \rangle$	
$\langle T_0, B, 2000, 2050 \rangle$	
	$A = 950$
	$B = 2050$
$\langle T_0 \text{ commit} \rangle$	
$\langle T_1 \text{ start} \rangle$	
$\langle T_1, C, 700, 600 \rangle$	
	$C = 600$
$\langle T_1 \text{ commit} \rangle$	

