CS460: Intro to Database Systems

Class 24: Crash Recovery

Instructor: Manos Athanassoulis

https://bu-disc.github.io/CS460/

Review: The ACID properties

Atomicity: All actions in the transaction happen, or none happen.

Consistency: If each transaction is consistent, and the DB starts consistent, it ends up consistent.

Isolation: Execution of one transaction is isolated from that of other transactions.

Durability: If a transaction commits, its effects persist.

Question: which ones does the Recovery Manager help with?

Atomicity & Durability (and also used for Consistency-related rollbacks)

Motivation

Atomicity:

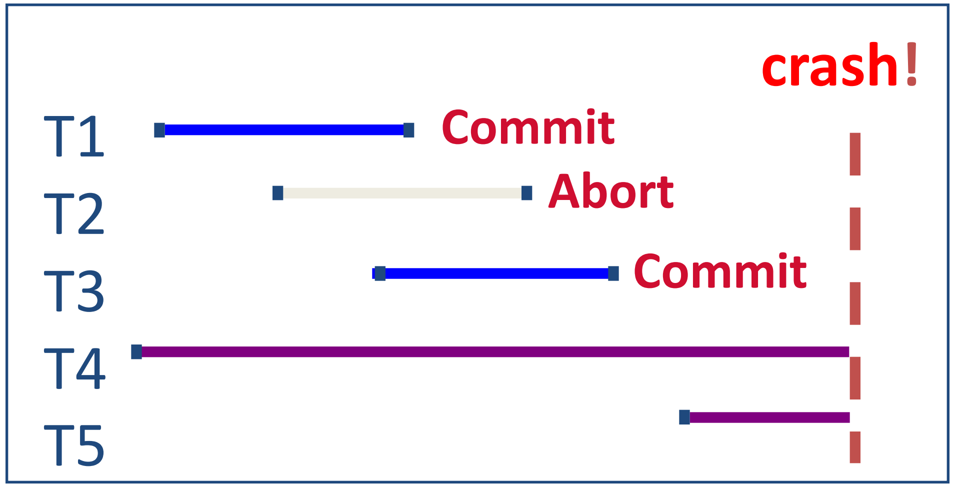
- Transactions may abort (“Rollback”).

Durability (& Atomicity):

- What if DBMS stops running? (Causes?)

Desired state after system restarts:

- T1 & T3 should be durable.
- T2, T4 & T5 should be aborted (effects should not be seen).



Assumptions

Concurrency control is in effect.

- Strict 2PL, in particular.

Updates are happening "in place".

- i.e., data is overwritten on (deleted from) the actual pages (not private copies)

What is simple scheme (without logging) to guarantee Atomicity & Durability?

- What happens during normal execution (what is the minimum lock granularity)?
- What happens when a transaction commits?
- What happens when a transaction aborts?

Buffer Management Plays a Key Role

- **Force policy – make sure that every update is on disk before commit.**
  - Provides durability without REDO logging.
  - But, can cause poor performance.

excessive I/Os:

if a highly used page is updated by 20 consecutive trxs, it will be over-written 20 times!!

- **No Steal policy – don't allow buffer-pool frames with <u>uncommited</u> updates to overwrite <u>committed</u> data on disk.**
  - Useful for ensuring atomicity without UNDO logging.
  - But can cause poor performance.

requires too much memory:

assumes all pages for all active transactions fit in the bufferpool!!

Buffer Management Plays a Key Role

- **Force policy – make sure that every update is on disk before commit.**
  - ~~Provides durability without~~ REDO logging.
  - But, can cause poor performance.

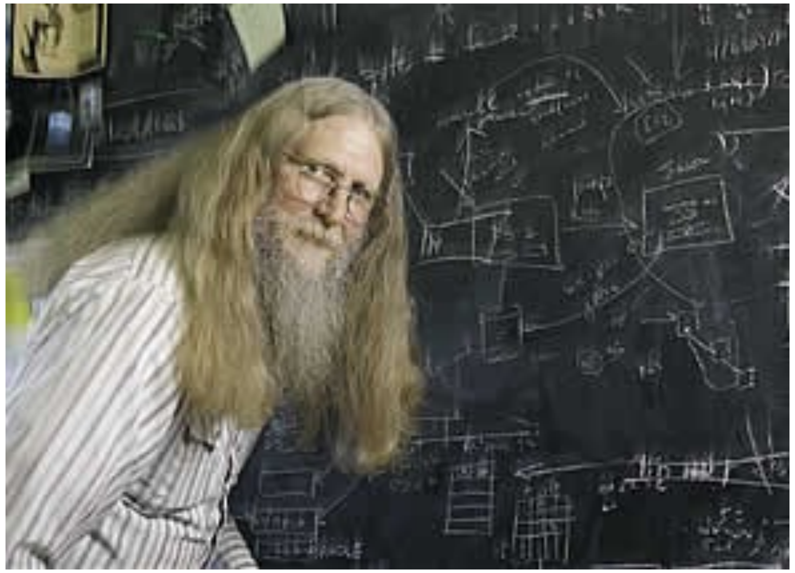
excessive I/Os:

if a highly used page is updated by 20 consecutive trxs, it will be over-written 20 times!!

- **No Steal policy – don't allow buffer-pool frames with uncommited updates to overwrite committed data on disk.**
  - ~~Useful for ensuring atomicity~~ without UNDO logging.
  - But can cause poor performance.

requires too much memory:

assumes all pages for all active transactions fit in the bufferpool!!



“three things are important in the database world: ***performance****,* ***performance****, and* ***performance****”*

Bruce Lindsay, IBM Research

ACM SIGMOD Edgar F. Codd Innovations award 2012

Preferred Policy: Steal/No-Force

More complicated but allows for highest performance

NO FORCE (allows updates of a committed transaction to NOT be on disk on commit time)

(complicates enforcing Durability)

- What if system crashes before a modified page written by a committed transaction makes it to disk?
- Write as **little** as possible, in a **convenient** place, at **commit** time, to support REDOing modifications.

STEAL (allows pages with uncommitted updates to overwrite committed data)

(complicates enforcing Atomicity)

- What if the transaction that performed updates aborts?
- What if system crashes before transaction is finished?
- Must remember the **old value** of P (to support UNDOing the write to page P).

Buffer Management summary

	No Steal	Steal
No Force		Fastest
Force	Slowest	

Performance Implications

	No Steal	Steal
No Force	No UNDO REDO	UNDO REDO
Force	No UNDO No REDO	UNDO No REDO

Logging/Recovery Implications

Basic Idea: Logging

Record REDO and UNDO information, for every update, in a *log*.

- Sequential writes to log (put it on a separate disk).
- Minimal info (diff) written to log, so multiple updates fit in a single log page.

Log: An ordered list of REDO/UNDO actions

- Log record contains:

  <XID, pageID, offset, length, old data, new data>

- and additional control info (which we'll see soon).

Write-Ahead Logging (WAL)

The Write-Ahead Logging Protocol:

1. Must force the log record for an update ***before*** the corresponding data page gets to disk.
2. Must force all log records for a Xact ***before commit***.
   (e.g., transaction is not committed until all its log records including its "commit" record are on the stable log.)

#1 (with UNDO info) helps guarantee Atomicity.

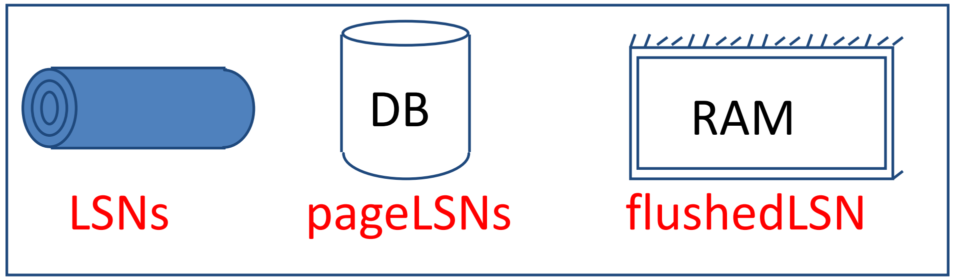
#2 (with REDO info) helps guarantee Durability.

This allows us to implement Steal/No-Force

Exactly how is logging (and recovery!) done?

- We'll look at the ARIES algorithm from IBM.

WAL & the Log



Each log record has a unique Log Sequence Number (LSN).

- LSNs are always increasing.

Each *data page* contains a pageLSN.

- The LSN of the most recent *log record* for an update to that page.

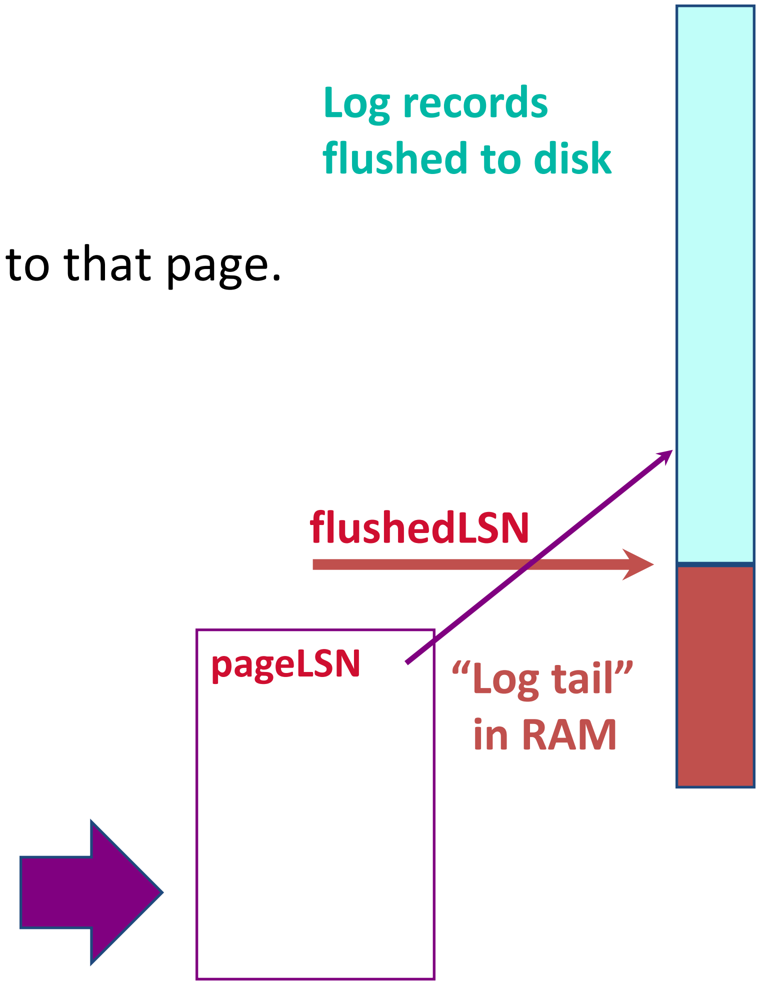
System keeps track of flushedLSN.

- The max LSN flushed so far.

WAL: For a page *i* to be written must flush log at least to the point where:

$pageLSN_i \le flushedLSN$

So that we can undo it!



Log Records

LogRecord fields:

LSN
prevLSN
XID
type

$$\text{update records only}\left\{\begin{array}{l}\text{pageID}\\\text{length}\\\text{offset}\\\text{before-image}\\\text{after-image}\end{array}\right.$$

prevLSN is the LSN of the previous log record written by *this* transaction
(so records of a transaction form a linked list backwards in time)

Possible log record types:

Update, Commit, Abort

Checkpoint (for log maintenance)

Compensation Log Records (CLRs)

– for UNDO actions

End (end of commit or abort)

Other Log-Related State

In-memory metadata:

Transaction Table

- One entry per currently active transactions (removed when trx commits or aborts)
- Contains XID, status (running/committing/aborting), and lastLSN (most recent LSN written by transaction). why? All active trxs at crash time have to be aborted!

Dirty Page Table

- One entry per dirty page in bufferpool (removed when the page is flushed to disk)
- Contains recLSN (recovery LSN) – the LSN of the log record which first caused the page to be dirty

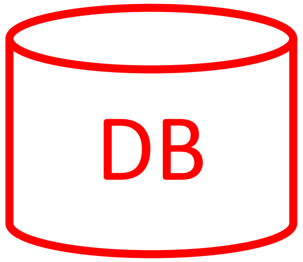
why?

This is the first record which may have to be redone!

The Big Picture: What's Stored Where

LogRecords

- prevLSN
- XID
- type
  - update
  - commit
  - abort
  - checkpoint
  - CLR
  - end
- pageID
- length
- offset
- before-image
- after-image

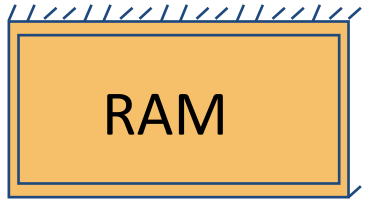


Data pages

each with a pageLSN

master record

LSN of most recent checkpoint



Xact Table

- lastLSN
- status

Dirty Page Table

- recLSN

flushedLSN

EXECUTING TRANSACTIONS WITH WAL

Normal Execution of a transaction

Series of reads & writes, followed by commit or abort.

- We will assume that disk write is atomic.
  - In practice, additional details to deal with non-atomic writes.

Strict 2PL.

STEAL, NO-FORCE buffer management, with Write-Ahead Logging.

Transaction Commit

Write commit record to log.

All log records up to transaction's commit record are flushed to disk.

- Guarantees that flushedLSN $\geq$ lastLSN.
- Note that log flushes are sequential, synchronous writes to disk.
- Many log records per log page.

Commit() returns.

Write end record to log.

Simple Transaction Abort

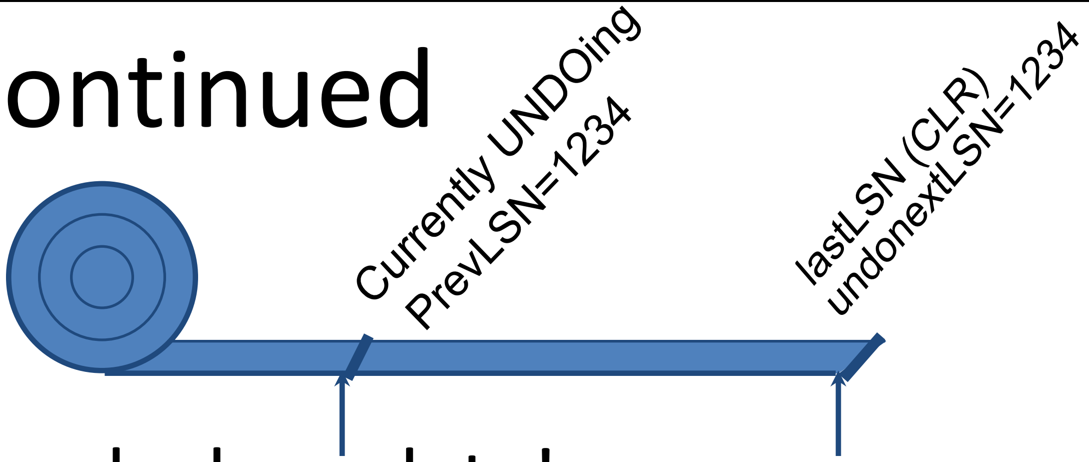
For now, consider an explicit abort of a Xact.

- No crash involved.

We want to "play back" the log in reverse order, UNDOing updates.

- Get lastLSN of Xact from Xact table.
- Can follow chain of log records backward via the prevLSN field.
- Write a "CLR" (compensation log record) for each undone operation.
- Write an *Abort* log record before starting to rollback operations.

Abort, continued



- To perform UNDO, must have a lock on data!
  - No problem (we’re doing Strict 2PL)!
- Before restoring old value of a page, write a CLR:
  - You continue logging while you UNDO!!
  - CLR has one extra field: undonextLSN
    - Points to the next LSN to undo (i.e., the prevLSN of the record we’re currently undoing).
  - CLRs *never* Undone (but they might be Redone when repeating history: guarantees Atomicity!)
- At end of UNDO, write an “end” log record.

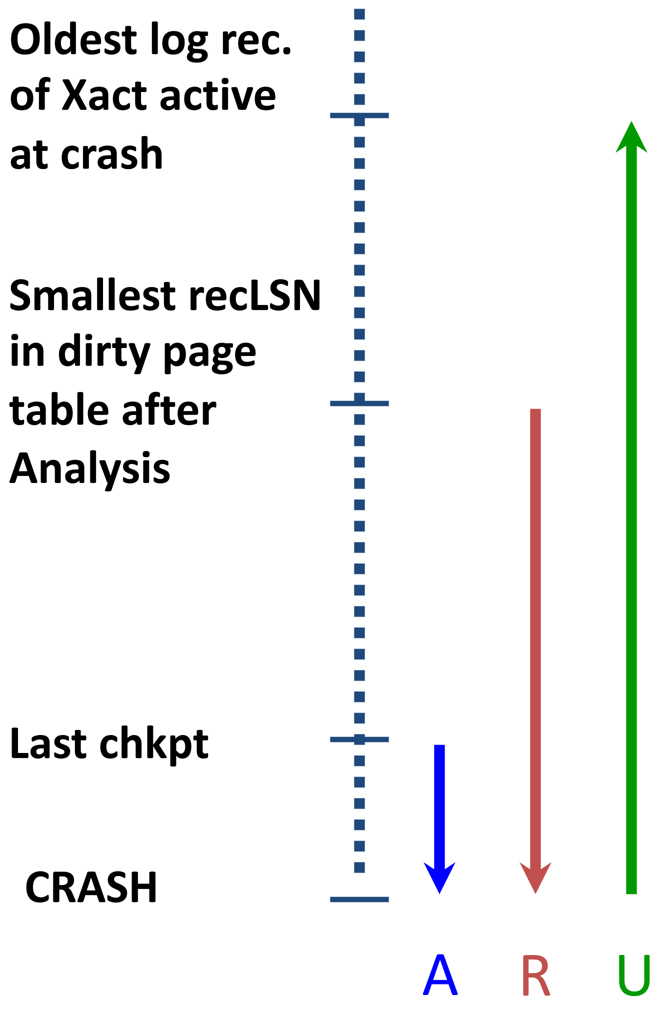
Checkpointing

Conceptually, keep log around for all time.
Obviously, this has performance/implementation problems…

Periodically, the DBMS creates a checkpoint, in order to minimize the time taken to recover in the event of a system crash. Write to log:

- begin_checkpoint record: Indicates when checkpoint began.
- end_checkpoint record: Contains current *transaction table* and *dirty page table*. This is a 'fuzzy checkpoint':
  - Other Xacts continue to run; so, these tables are accurate only as of the time of the begin_checkpoint record.
  - No attempt to force dirty pages to disk; effectiveness of checkpoint limited by oldest unwritten change to a dirty page.
- Store LSN of most recent checkpoint record in a safe place (*master* record).

Crash Recovery: Big Picture


Oldest log rec. of Xact active at crash
Smallest recLSN in dirty page table after Analysis
Last chkpt
CRASH
A
R
U

- Start from a checkpoint (found via master record).
- Three phases. Need to do:
  - Analysis - Figure out which transactions committed since checkpoint, which failed.
  - REDO ***all*** actions.
    (repeat history)
  - UNDO effects of failed transactions.

Recovery: The Analysis Phase

Re-establish knowledge of state at checkpoint.

- via transaction table and dirty page table stored in the checkpoint

Scan log forward from checkpoint.

- End record: Remove Xact from Xact table.
- All Other records: Add Xact to Xact table, set lastLSN=LSN, change Xact status on commit.
- also, for Update records: If page P not in Dirty Page Table, Add P to DPT, set its recLSN=LSN.

At end of Analysis…

- transaction table says which xacts were active at time of crash.
- DPT says which dirty pages *might not* have made it to disk

Phase 2: The REDO Phase

We *Repeat History* to reconstruct state at crash:

- Reapply *all* updates (even of aborted transactions!), redo CLRs.

Scan forward from log rec containing smallest recLSN in DPT.

Q: why start here? the first update that dirtied the page (update that might not made it to the disk)

For each update log record or CLR, REDO the action unless one of the following holds:

- Affected page is not in the Dirty Page Table (all changes to this page made it to disk)
- Affected page is in D.P.T., but has recLSN > LSN (the specific update made it to disk)
- pageLSN (in DB) ≥ LSN. (this last case requires I/O) (ensure update is on disk)

To REDO an action:

- Reapply logged action.
- Set pageLSN to LSN. No additional logging, no forcing!

Phase 3: The UNDO Phase

ToUndo={lastLSNs of all Xacts in the Xact Table}

Repeat:

- Choose (and remove) largest LSN among ToUndo.
- If this LSN is a CLR and undonextLSN==NULL

  Write an End record for this transation.
- If this LSN is a CLR, and undonextLSN != NULL

  Add undonextLSN to ToUndo
- Else this LSN is an update. Undo the update, write a CLR, add prevLSN to ToUndo.

Until ToUndo is empty.

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

RAM

Active Transaction Table

Dirty Page Table

Master Record:

last checkpoint at LSN 00

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

RAM

Active Transaction Table

T1, running, 10

Dirty Page Table

P5, 10

Master Record:

last checkpoint at LSN 00

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

RAM

Active Transaction Table

T1, running, 10
T2, running, 20

Dirty Page Table

P5, 10
P3, 20

Master Record:

last checkpoint at LSN 00

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

RAM

Active Transaction Table

T1, ~~running, 10~~ aborting, 30

T2, running, 20

Dirty Page Table

P5, 10

P3, 20

Master Record:

last checkpoint at LSN 00

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

RAM

Active Transaction Table

T1, aborting, ~~30~~ 40

T2, running, 20

Dirty Page Table

P5, 10

P3, 20

Master Record:

last checkpoint at LSN 00

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

RAM

Active Transaction Table

~~T1, aborting, 40~~

T2, running, 20

Dirty Page Table

P5, 10

P3, 20

Master Record:

last checkpoint at LSN 00

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

RAM

Active Transaction Table

~~T1, aborting, 40~~
T2, running, 20
T3, running, 50

Dirty Page Table

P5, 10
P3, 20
P1, 50

Master Record:

last checkpoint at LSN 00

Example: Analysis Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH

Master Record:

last checkpoint at LSN 00

RAM

Active Transaction Table

~~T1, aborting, 40~~

T2, running, ~~20~~ 60

T3, running, 50

Dirty Page Table

P5, 10

P3, 20

P1, 50

P5 already dirty!

ToUndo

50

60

Analysis phase done!

→ need to REDO from 10

→ need to UNDO T2 and T3, ToUndo={50,60}

Example: Redo Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART

Master Record:

last checkpoint at LSN 00

RAM

Active Transaction Table

T2, running, 60

T3, running, 50

Dirty Page Table

P5, 10

P3, 20

P1, 50

ToUndo

50

60

Redo everything from 10

No logging – no forcing!

Example: Undo Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART
70	***CLR***: Undo T2 LSN 60, undoNextLSN=20

Master Record:

last checkpoint at LSN 00

RAM

Active Transaction Table

T2, ~~running, 60~~ aborting, 70

T3, running, 50

Dirty Page Table

P5, 10

P3, 20

P1, 50

ToUndo

50

~~60~~

20

Example: Undo Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART
70	***CLR***: Undo T2 LSN 60, undoNextLSN=20
80	***CLR***: Undo T3 LSN 50, undoNextLSN=NULL

Master Record:
last checkpoint at LSN 00

RAM

Active Transaction Table
T2, aborting, 70
T3, ~~running, 50~~ aborting, 80

Dirty Page Table

P5	10
P3	20
P1	50

ToUndo
~~50~~
20

Example: Undo Phase

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART
70	***CLR***: Undo T2 LSN 60, undoNextLSN=20
80	***CLR***: Undo T3 LSN 50, undoNextLSN=NULL
85	***End***, T3, prevLSN=80

Master Record:
last checkpoint at LSN 00

RAM

Active Transaction Table
T2, aborting, 70
~~T3, aborting, 80~~

Dirty Page Table
P5, 10
P3, 20
P1, 50

ToUndo
20

Example: Second Crash

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART
70	***CLR***: Undo T2 LSN 60, undoNextLSN=20
80	***CLR***: Undo T3 LSN 50, undoNextLSN=NULL
85	***End***, T3, prevLSN=80
	CRASH

Master Record:
last checkpoint at LSN 00

RAM

Active Transaction Table

Dirty Page Table

ToUndo

We lost all metadata!

We perform analysis and we reach exactly at the same point.

Example: Analysis & Redo

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART
70	***CLR***: Undo T2 LSN 60, undoNextLSN=20
80	***CLR***: Undo T3 LSN 50, undoNextLSN=NULL
85	***End***, T3, prevLSN=80
	CRASH, RESTART

Master Record:

last checkpoint at LSN 00

RAM

Active Transaction Table

T2, aborting, 70

Dirty Page Table

P5, 10

P3, 20

P1, 50

ToUndo

20

Example: Undo Phase (2nd)

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART
70	***CLR***: Undo T2 LSN 60, undoNextLSN=20
80	***CLR***: Undo T3 LSN 50, undoNextLSN=NULL
85	***End***, T3, prevLSN=80
	CRASH, RESTART
90	***CLR***: Undo T2 LSN 20, undoNextLSN=NULL

Master Record:

last checkpoint at LSN 00

RAM

Active Transaction Table

T2, aborting, ~~70~~ 90

Dirty Page Table

P5, 10

P3, 20

P1, 50

ToUndo

~~20~~

Example: Undo Phase (2nd)

LSN	LOG
00	Begin Checkpoint
05	End Checkpoint
10	***Update***, T1, P5, prevLSN=NULL
20	***Update***, T2, P3, prevLSN=NULL
30	***Abort***, T1, prevLSN=10
40	***CLR***: Undo T1 LSN 10, undoNextLSN=NULL
45	***End***, T1, prevLSN=30
50	***Update***, T3, P1, prevLSN=NULL
60	***Update***, T2, P5, prevLSN=20
	CRASH, RESTART
70	***CLR***: Undo T2 LSN 60, undoNextLSN=20
80	***CLR***: Undo T3 LSN 50, undoNextLSN=NULL
85	***End***, T3, prevLSN=80
	CRASH, RESTART
90	***CLR***: Undo T2 LSN 20, undoNextLSN=NULL
95	***End***, T2, prevLSN=90

Master Record:

last checkpoint at LSN 00

RAM

Active Transaction Table

~~T2, aborting, 90~~

Dirty Page Table

P5, 10

P3, 20

P1, 50

ToUndo

Recovery completed!

Normal execution can resume!

Additional Crash Issues

What happens if system crashes during Analysis? During REDO?

How do you limit the amount of work in REDO?

- Flush asynchronously in the background.

How do you limit the amount of work in UNDO?

- Avoid long-running transactions.

Summary of Logging/Recovery

Recovery Manager guarantees Atomicity & Durability.

Use WAL to allow STEAL/NO-FORCE without sacrificing correctness.

LSNs identify log records; linked into backwards chains per transaction (via prevLSN).

pageLSN allows comparison of data page and log records.

Summary, continued

Checkpointing: A quick way to limit the amount of log to scan on recovery.

Recovery works in 3 phases:

- Analysis: Forward from checkpoint.
- Redo: Forward from oldest recLSN.
- Undo: Backward from end to first LSN of oldest Xact alive at crash.

Upon Undo, write CLRs.

Redo “repeats history”: Simplifies the logic!