CS 6530: Advanced Database Systems Fall 2022

Lecture 11
Logging & Recovery Protocols II

Prashant Pandey
prashant.pandey@utah.edu

Acknowledgement: Slides taken from Prof. Andy Pavlo, CMU

Some announcements…

- Project #2 due Tuesday October 25th
  - Any questions?
  - It will be a significant coding/debugging effort.
  - Recovery can be the hardest part!!
- Paper report #3 due on Thursday October 27th
- Final project
  - Start brainstorming about Final project in your group.
  - We will release details soon.
  - **Proposal deadline for final project November 1st**

CRASH RECOVERY

Recovery algorithms are techniques to ensure database consistency, transaction atomicity, and durability despite failures.

Recovery algorithms have two parts:

→ Actions during normal txn processing to ensure that the DBMS can recover from a failure.
→ Actions after a failure to recover the database to a state that ensures atomicity, consistency, and durability.

ARIES

Algorithms for **R**ecovery and **I**solation **E**xploiting **S**emantics

Developed at IBM Research in early 1990s for the DB2 DBMS.

Not all systems implement ARIES exactly as defined in this paper but they're close enough.

ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging

C. MOHAN
IBM Almaden Research Center
and
DON HADERLE
IBM Santa Teresa Laboratory
and
BRUCE LINDSAY, HAMID PIRAHESH and PETER SCHWARZ
IBM Almaden Research Center

In this paper we present a simple and efficient method, called ARIES (*Algorithm for Recovery and Isolation Exploiting Semantics*), which supports partial rollbacks of transactions, fine-granularity (e.g., record) locking and recovery using write-ahead logging (WAL). We introduce the paradigm of *repeating history* to redo all missing updates *before* performing the rollbacks of the loser transactions during restart after a system failure. ARIES uses a log sequence number in each page to correlate the state of a page with respect to logged updates of that page. All updates of a transaction are logged, including those performed during rollbacks. By appropriate chaining of the log records written during rollbacks to those written during forward progress, a bounded amount of logging is ensured during rollbacks even in the face of repeated failures during restart or of nested rollbacks We deal with a variety of features that are very important in building and operating an *industrial-strength* transaction processing system ARIES supports fuzzy checkpoints, selective and deferred restart, fuzzy image copies, media recovery, and high concurrency lock modes (e.g., increment/decrement) which exploit the semantics of the operations and require the ability to perform operation logging. ARIES is flexible with respect to the kinds of buffer management policies that can be implemented. It supports objects of varying length efficiently. By enabling parallelism during restart, page-oriented redo, and logical undo, it enhances concurrency and performance. We show why some of the System R paradigms for logging and recovery, which were based on the shadow page technique, need to be changed in the context of WAL. We compare ARIES to the WAL-based recovery methods of

Authors' addresses: C Mohan, Data Base Technology Institute, IBM Almaden Research Center, San Jose, CA 95120; D. Haderle, Data Base Technology Institute, IBM Santa Teresa Laboratory, San Jose, CA 95150; B. Lindsay, H. Pirahesh, and P. Schwarz, IBM Almaden Research Center, San Jose, CA 95120.
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1992 0362-5915/92/0300-0094 $1.50

ACM Transactions on Database Systems, Vol 17, No. 1, March 1992, Pages 94–162

ARIES – MAIN IDEAS

Write-Ahead Logging:

→ Any change is recorded in log on stable storage before the database change is written to disk.

→ Must use **STEAL** + **NO-FORCE** buffer pool policies.

Repeating History During Redo:

→ On restart, retrace actions and restore database to exact state before crash.

Logging Changes During Undo:

→ Record undo actions to log to ensure action is not repeated in the event of repeated failures.

TODAY'S AGENDA

Log Sequence Numbers

Recovery Algorithm

WAL RECORDS

We need to extend our log record format from last class to include additional info.

Every log record now includes a globally unique ***log sequence number*** (LSN).

Various components in the system keep track of ***LSNs*** that pertain to them…

LOG SEQUENCE NUMBERS

Name	Where	Definition
flushedLSN	Memory	Last LSN in log on disk
pageLSN	$page_x$	Newest update to $page_x$
recLSN	$page_x$	Oldest update to $page_x$ since it was last flushed
lastLSN	T_i	Latest record of txn T_i
MasterRecord	Disk	LSN of latest checkpoint

WRITING LOG RECORDS

Each data page contains a **pageLSN**.

→ The ***LSN*** of the most recent update to that page.

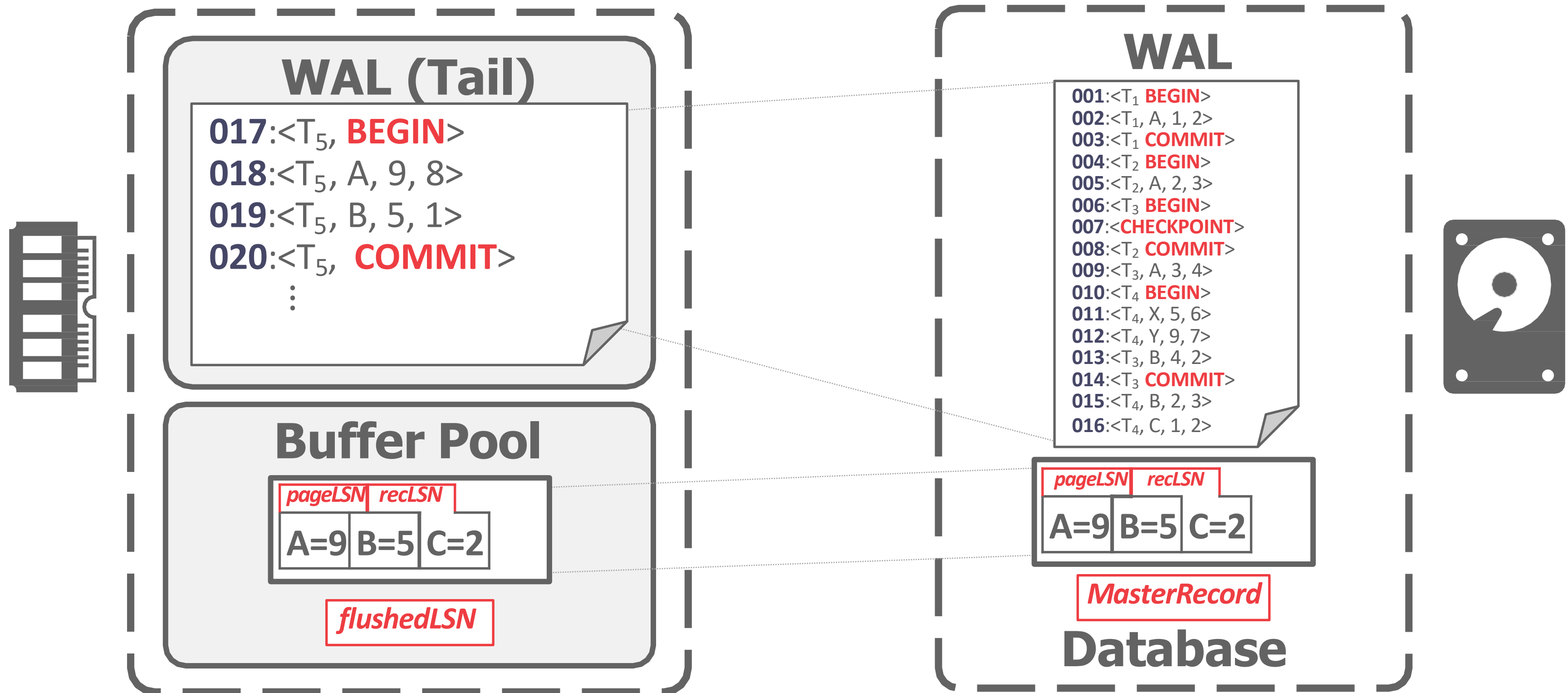
System keeps track of **flushedLSN**.

→ The max ***LSN*** flushed so far.

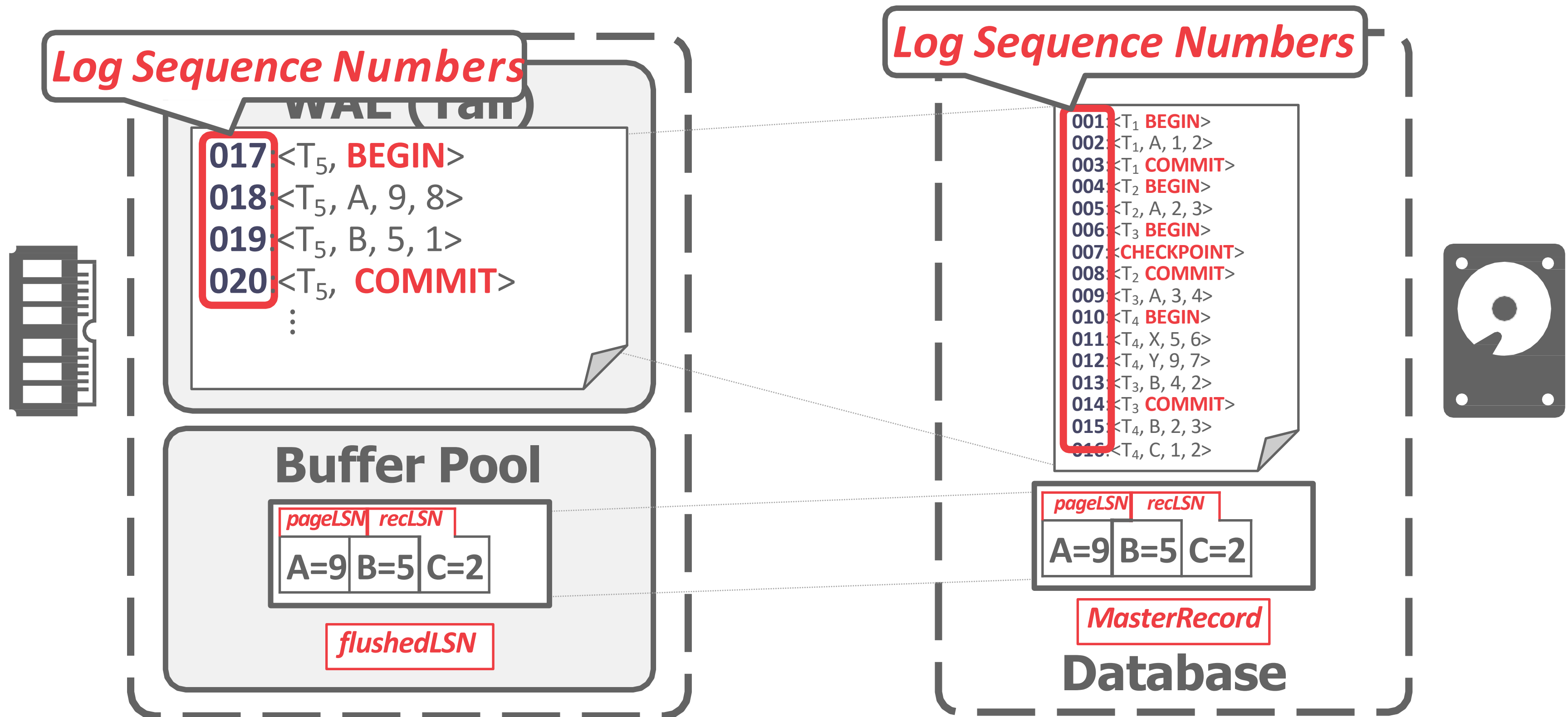
Before page **x** can be written to disk, we must flush log at least to the point where:

→ **$pageLSN_x \leq flushedLSN$**

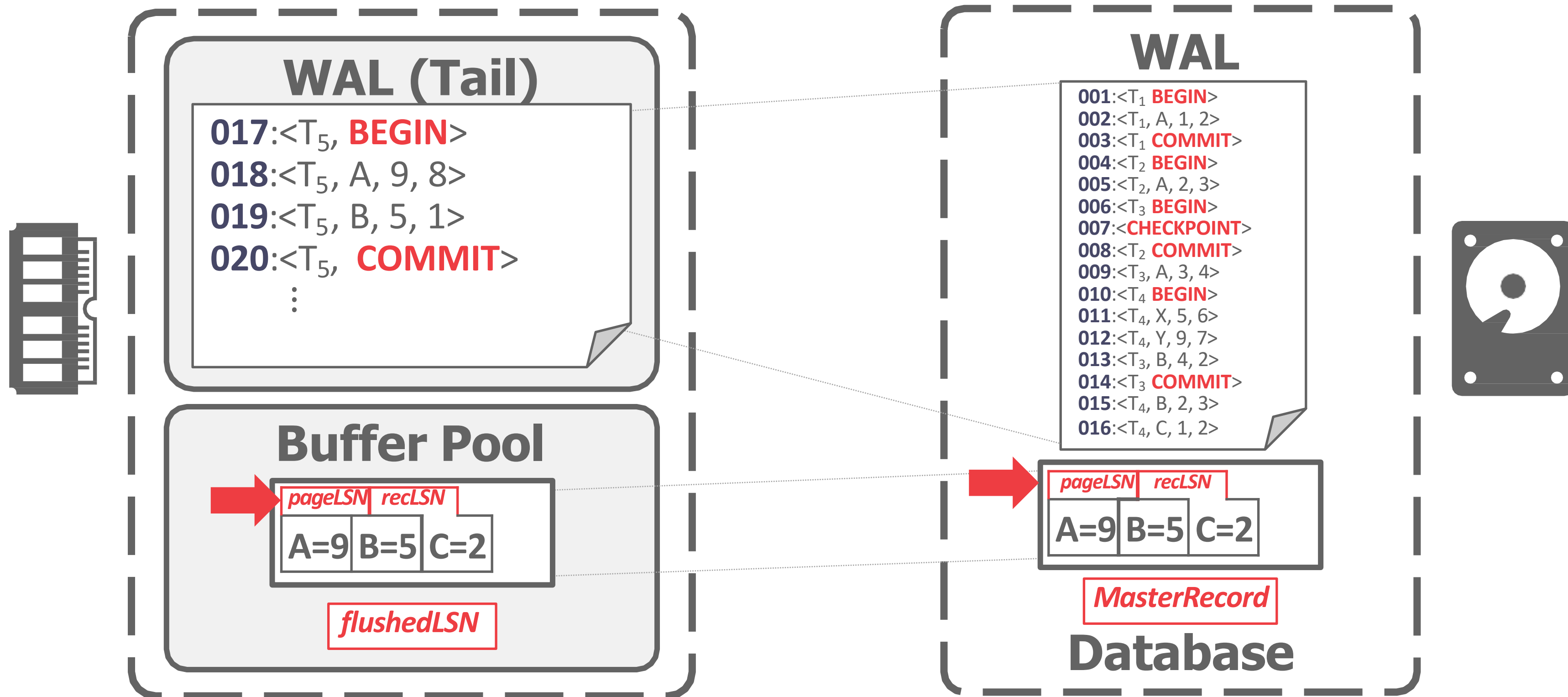
WRITING LOG RECORDS



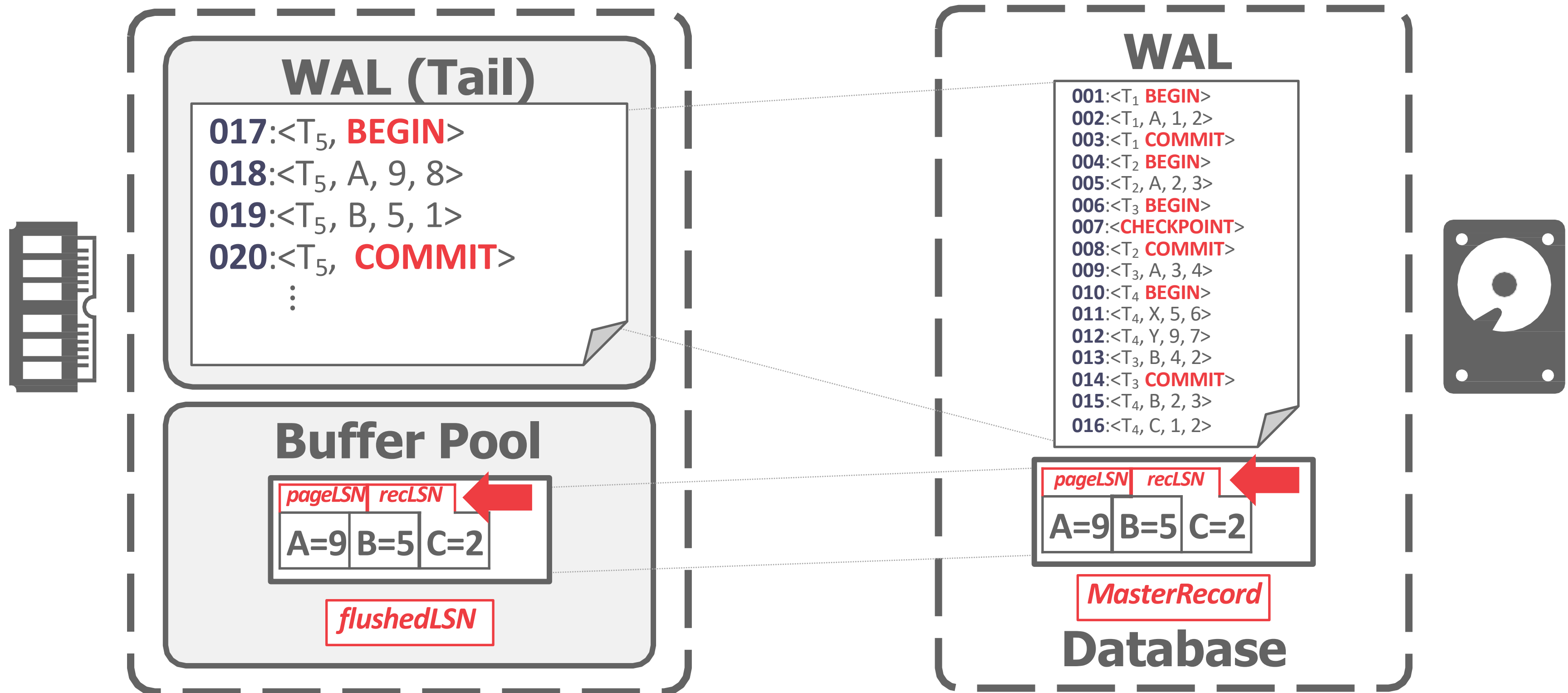
WRITING LOG RECORDS



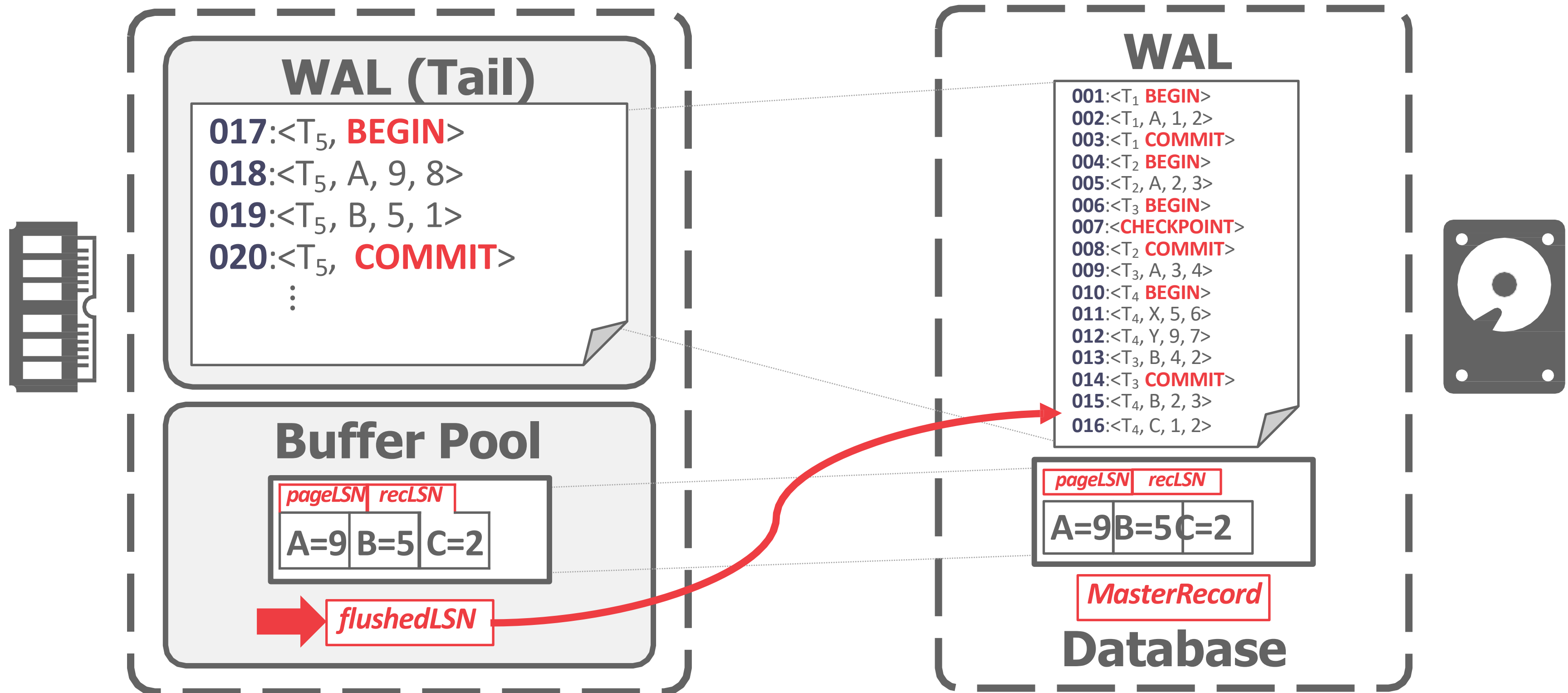
WRITING LOG RECORDS



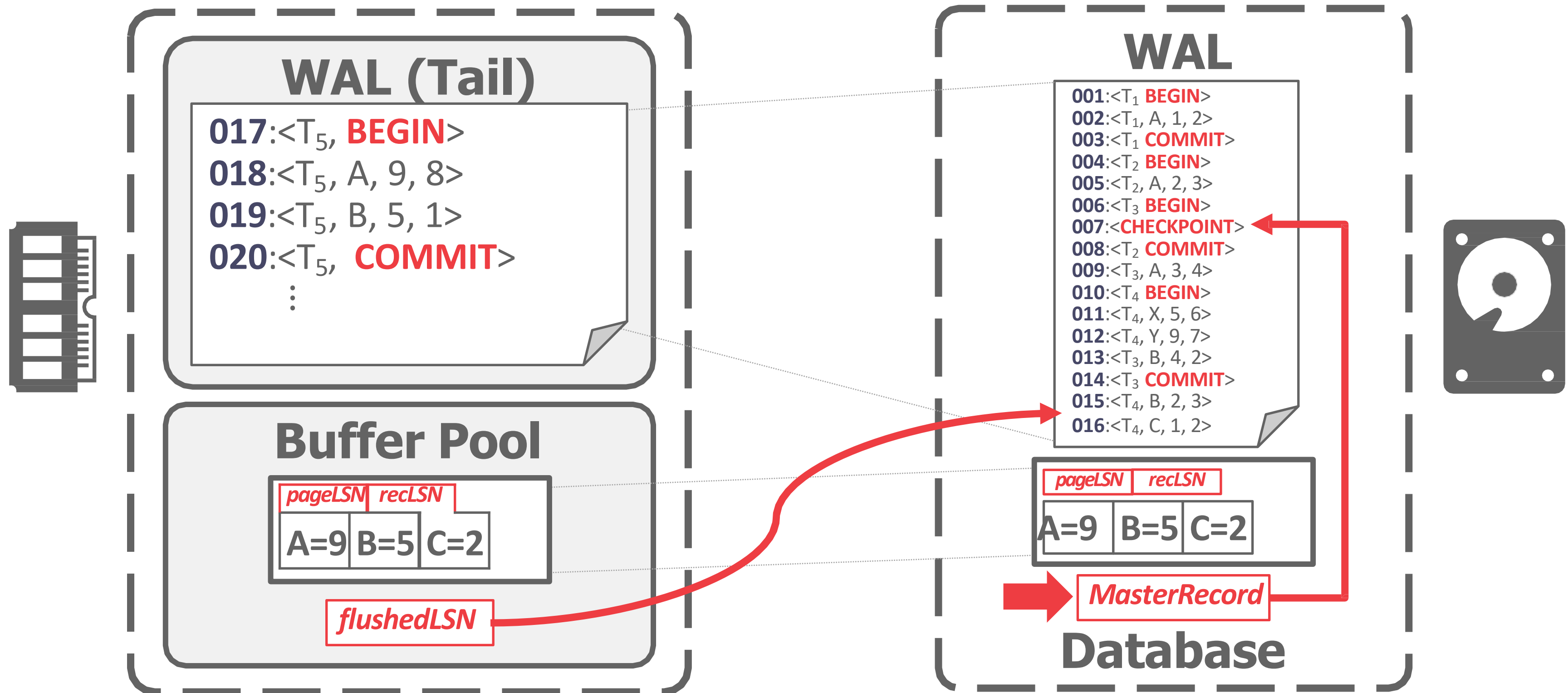
WRITING LOG RECORDS



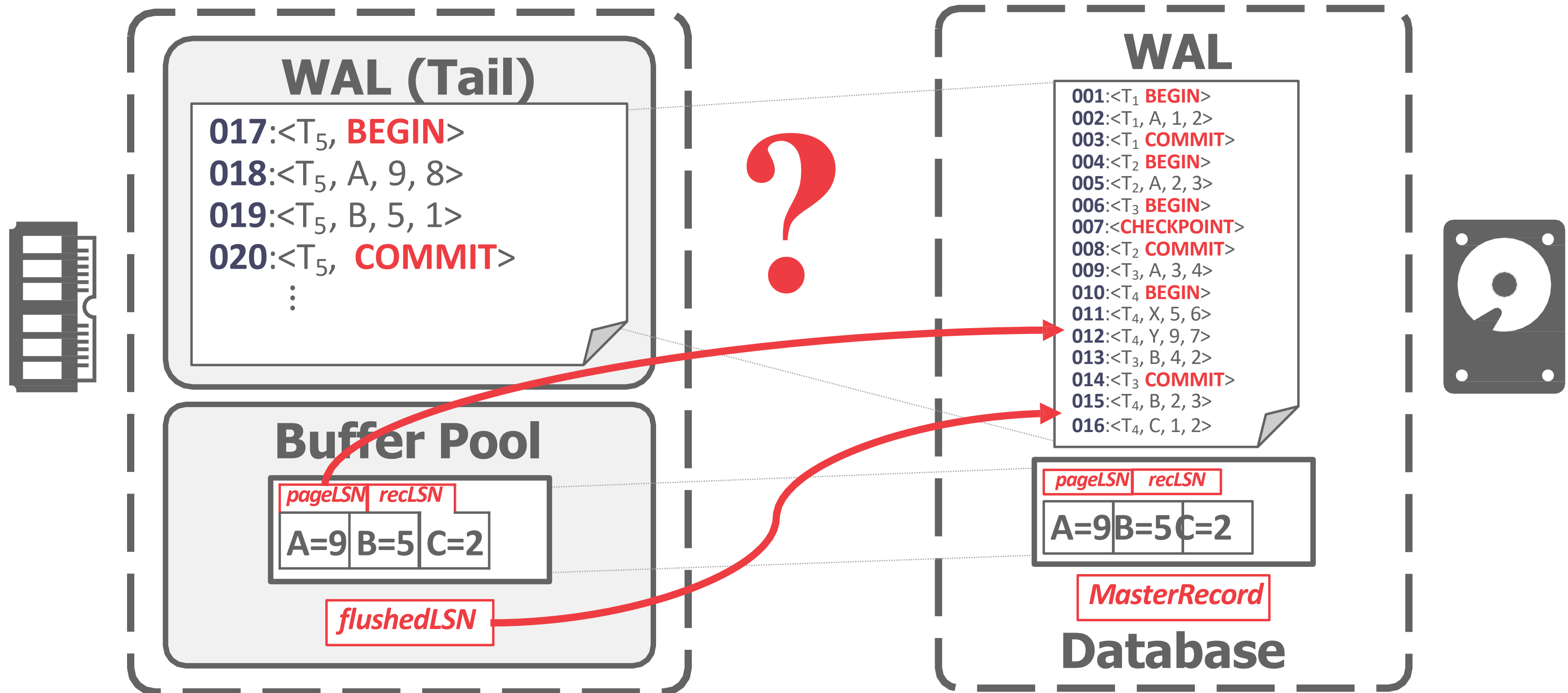
WRITING LOG RECORDS



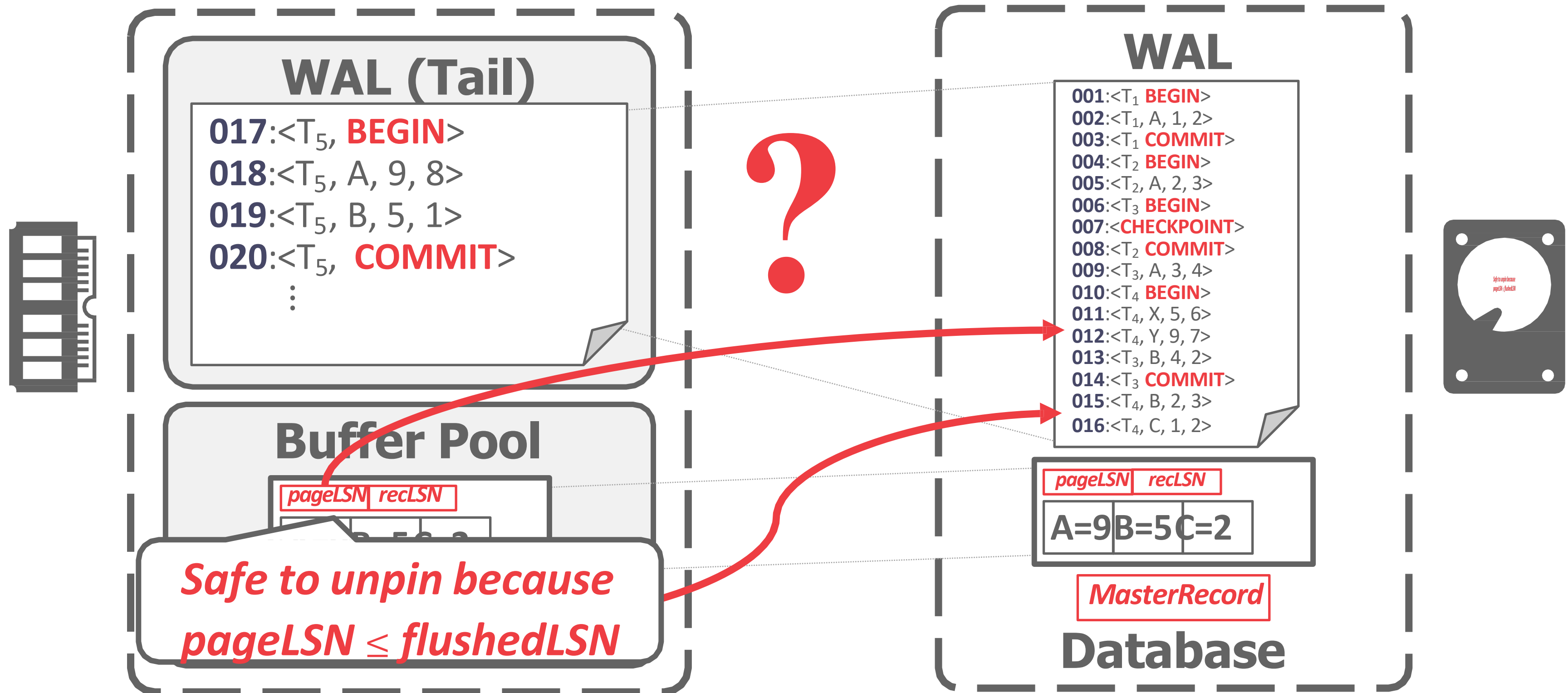
WRITING LOG RECORDS



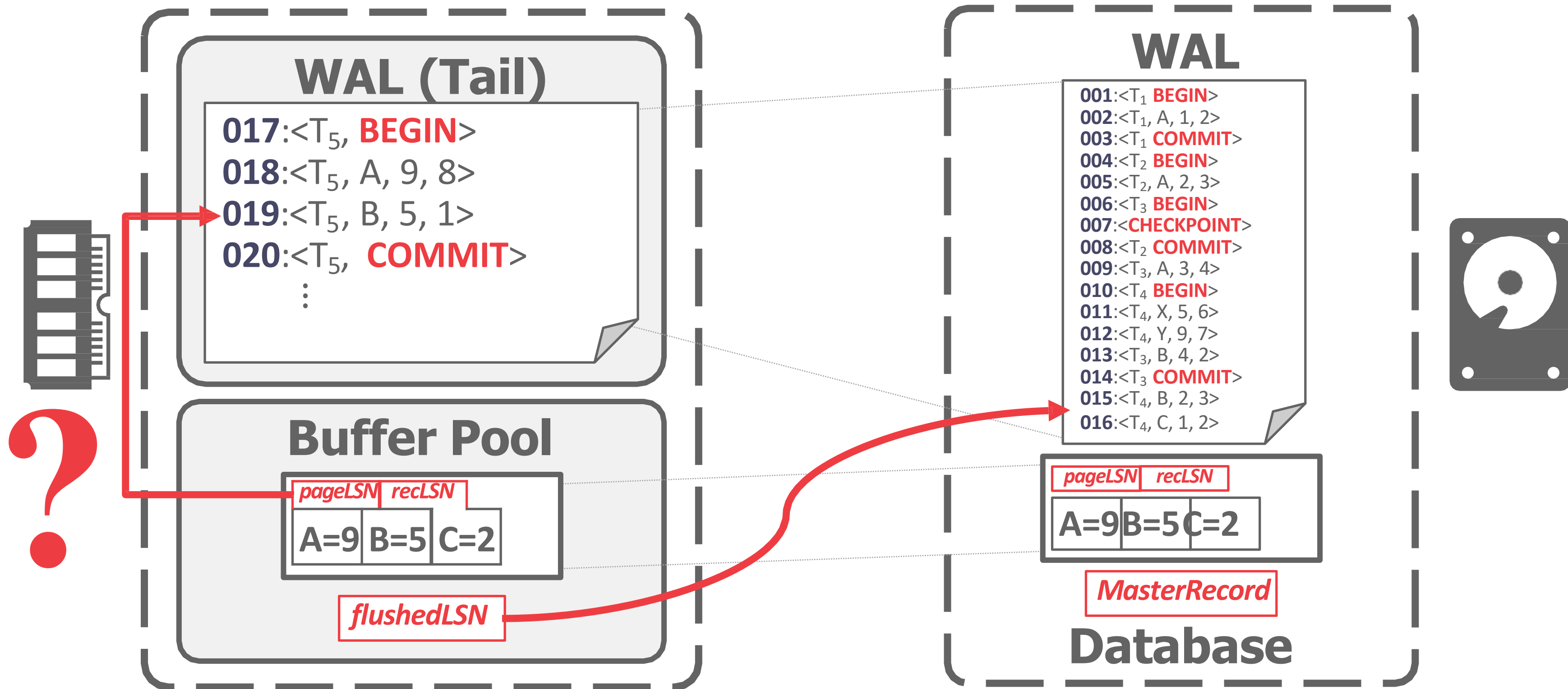
WRITING LOG RECORDS



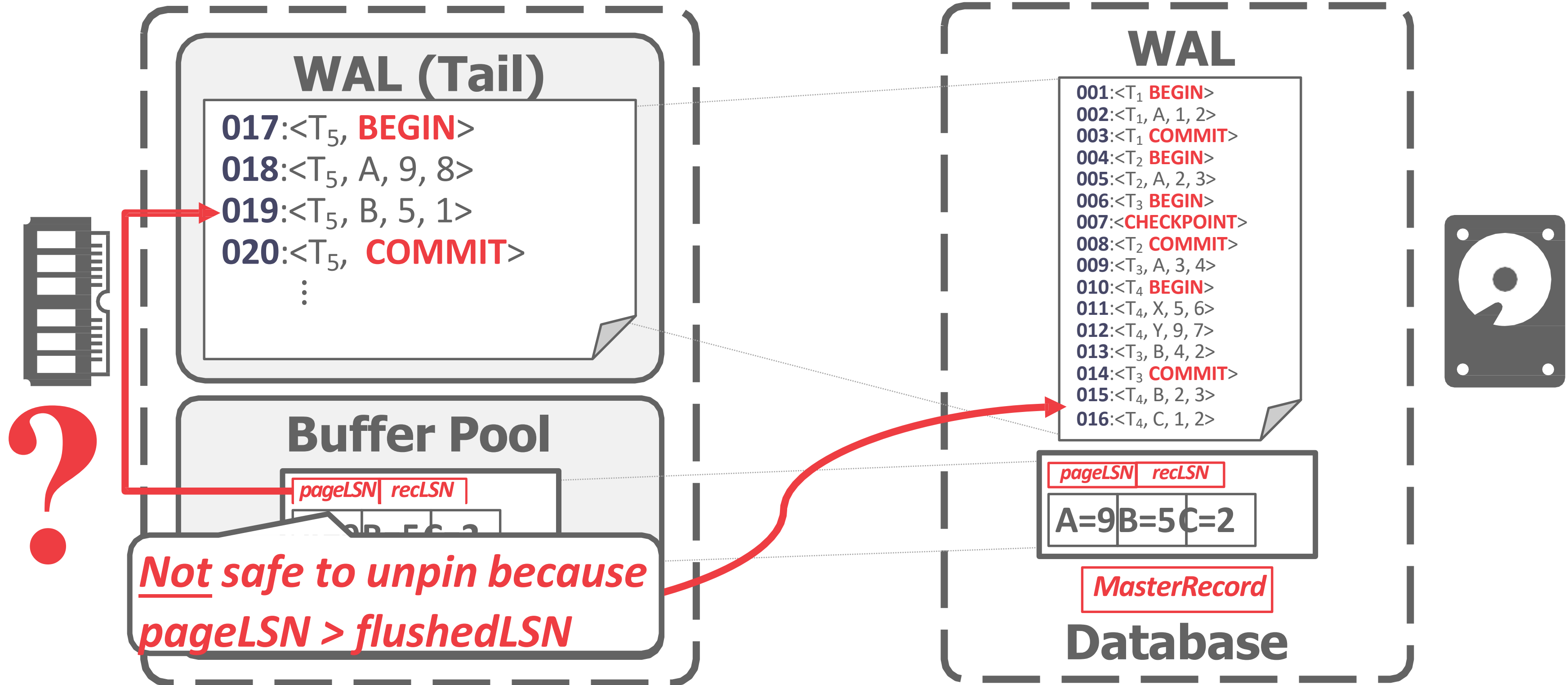
WRITING LOG RECORDS



WRITING LOG RECORDS



WRITING LOG RECORDS



WRITING LOG RECORDS

All log records have an ***LSN***.

Update the **pageLSN** every time a txn modifies a record in the page.

Update the **flushedLSN** in memory every time the DBMS writes out the WAL buffer to disk.

NORMAL EXECUTION

Each txn invokes a sequence of reads and writes, followed by commit or abort.

Assumptions in this lecture:

→ All log records fit within a single page.
→ Disk writes are atomic.
→ Single-versioned tuples with Strict 2PL.
→ **STEAL** + **NO-FORCE** buffer management with WAL.

TRANSACTION COMMIT

Write **COMMIT** record to log.

All log records up to txn's **COMMIT** record are flushed to disk.

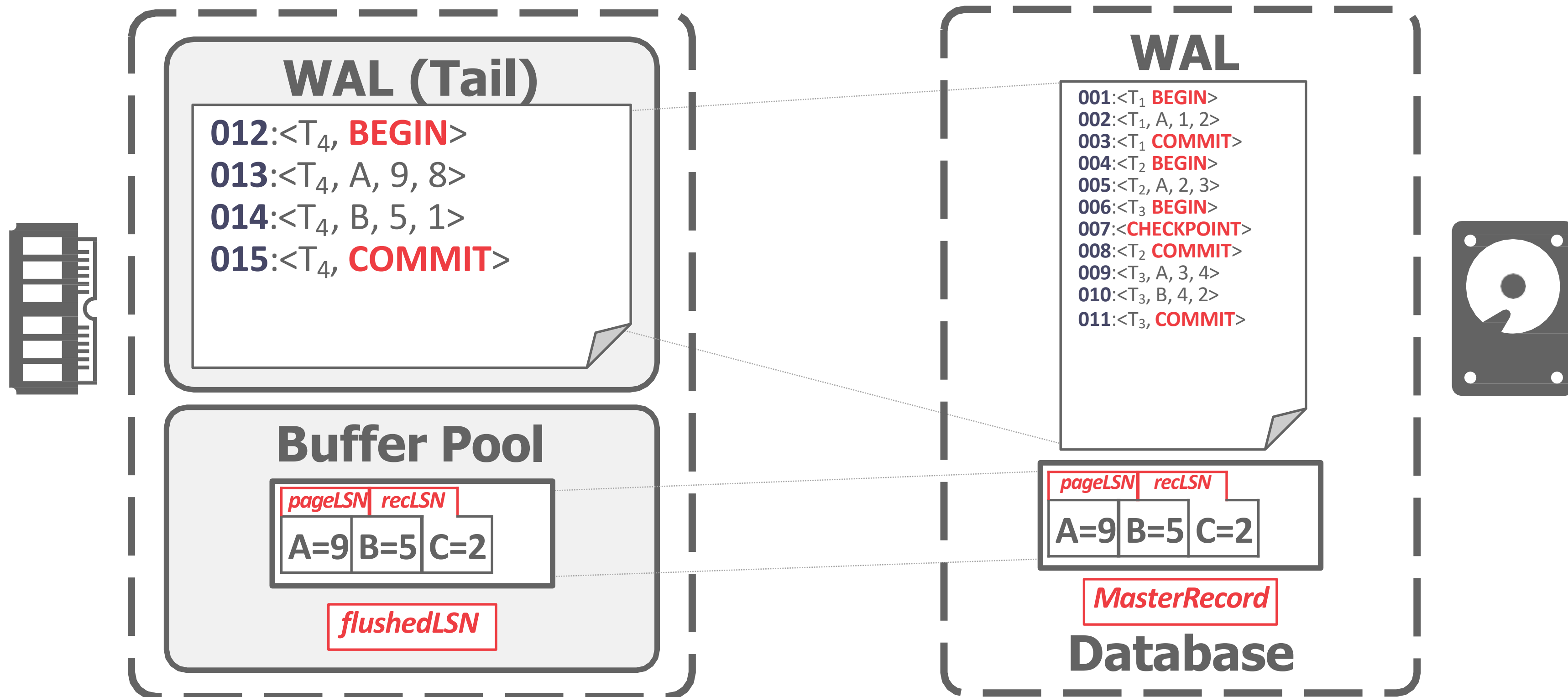
→ Log flushes are sequential, synchronous writes to disk.

→ Many log records per log page.

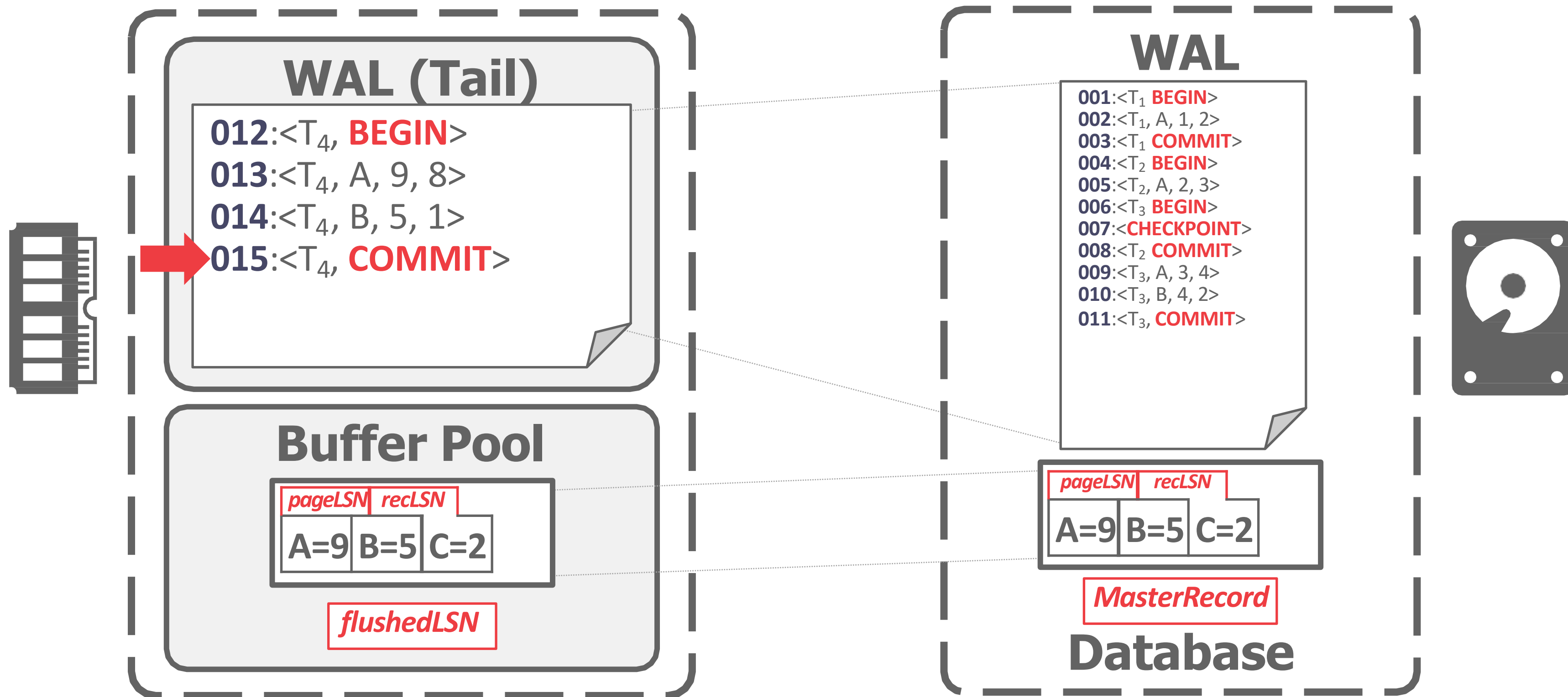
When the commit succeeds, write a special **TXN-END** record to log.

→ This does not need to be flushed immediately.

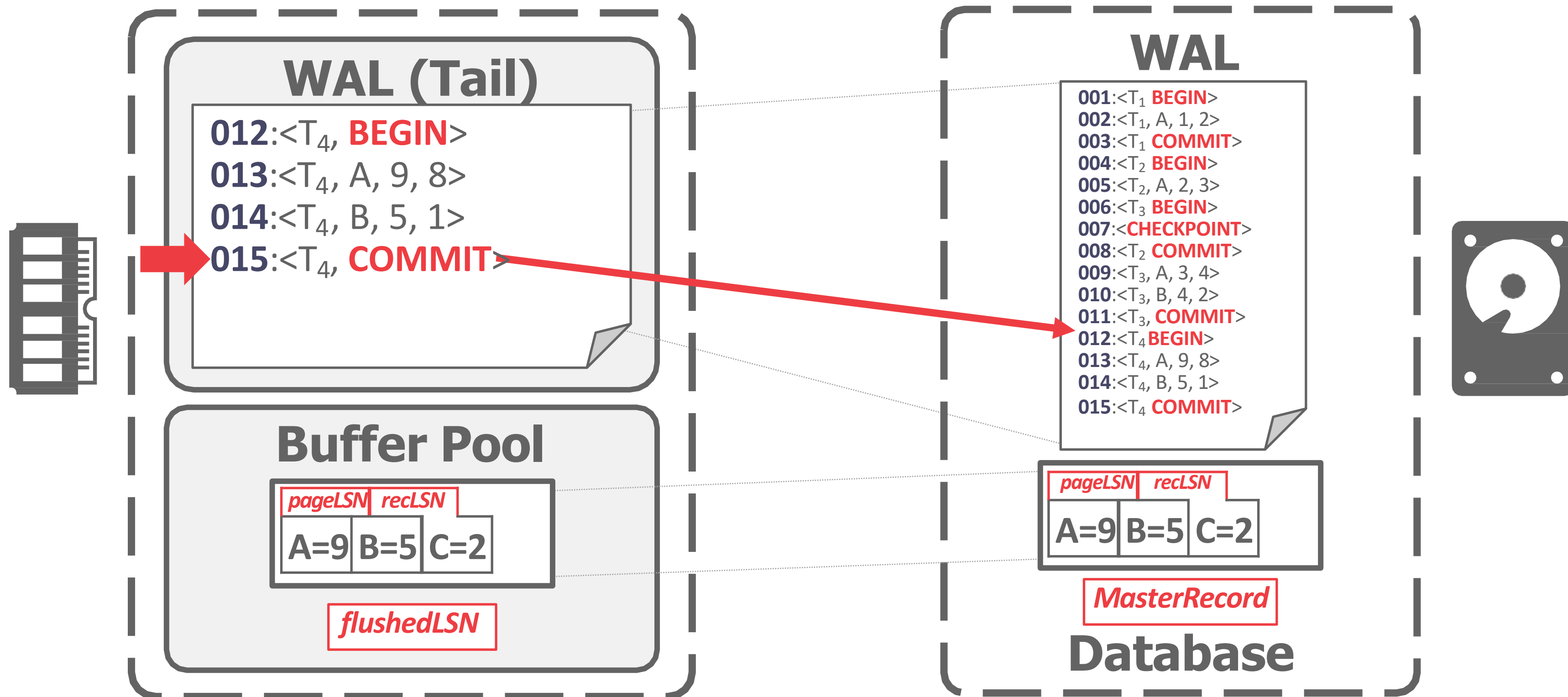
TRANSACTION COMMIT



TRANSACTION COMMIT



TRANSACTION COMMIT



TRANSACTION COMMIT

WAL (Tail)

012:<T_4, BEGIN>
013:<T_4, A, 9, 8>
014:<T_4, B, 5, 1>
015:<T_4, COMMIT>

Buffer Pool

pageLSN | recLSN

A=9 | B=5 | C=2

flushedLSN

flushedLSN = 015

WAL

001:<T_1 BEGIN>
002:<T_1, A, 1, 2>
003:<T_1 COMMIT>
004:<T_2 BEGIN>
005:<T_2, A, 2, 3>
006:<T_3 BEGIN>
007:<CHECKPOINT>
008:<T_2 COMMIT>
009:<T_3, A, 3, 4>
010:<T_3, B, 4, 2>
011:<T_3, COMMIT>
012:<T_4BEGIN>
013:<T_4, A, 9, 8>
014:<T_4, B, 5, 1>
015:<T_4 COMMIT>

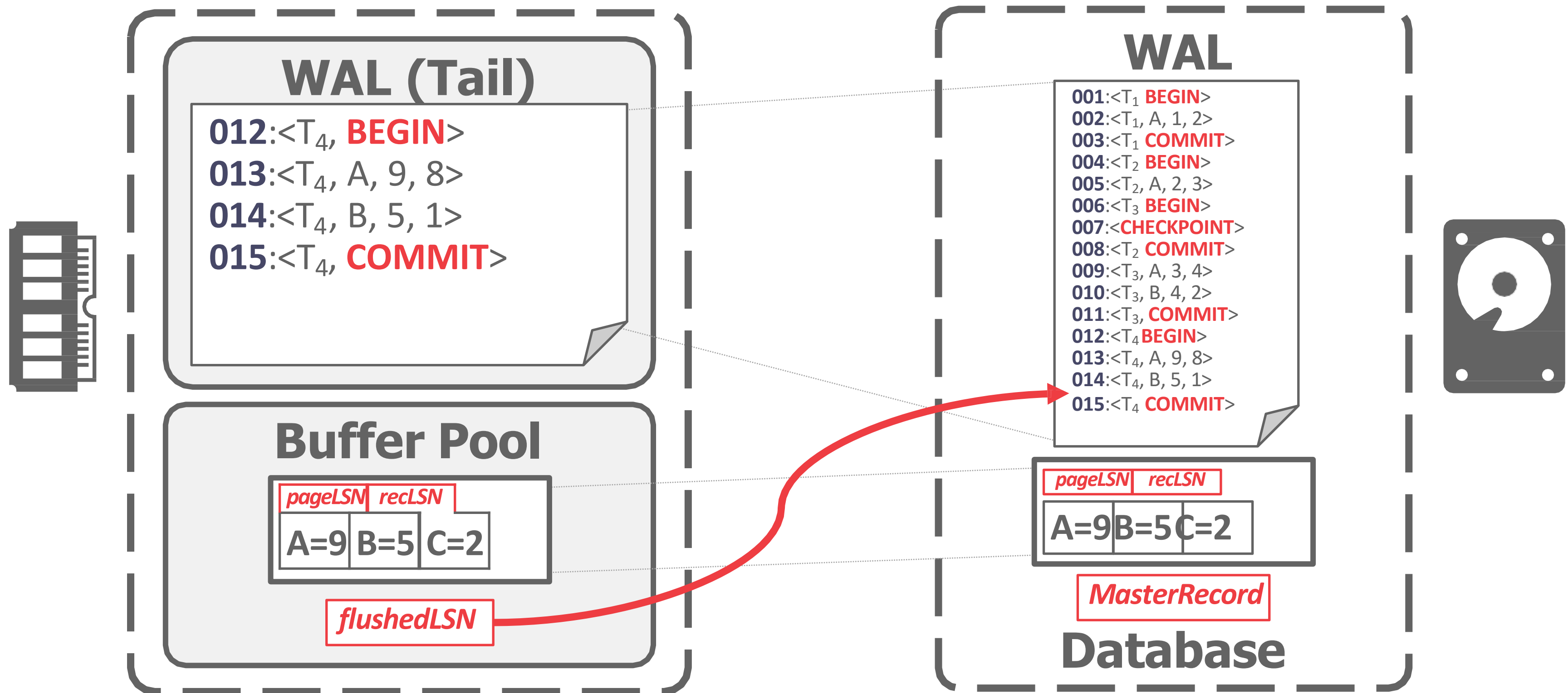
pageLSN | recLSN

A=9 | B=5 | C=2

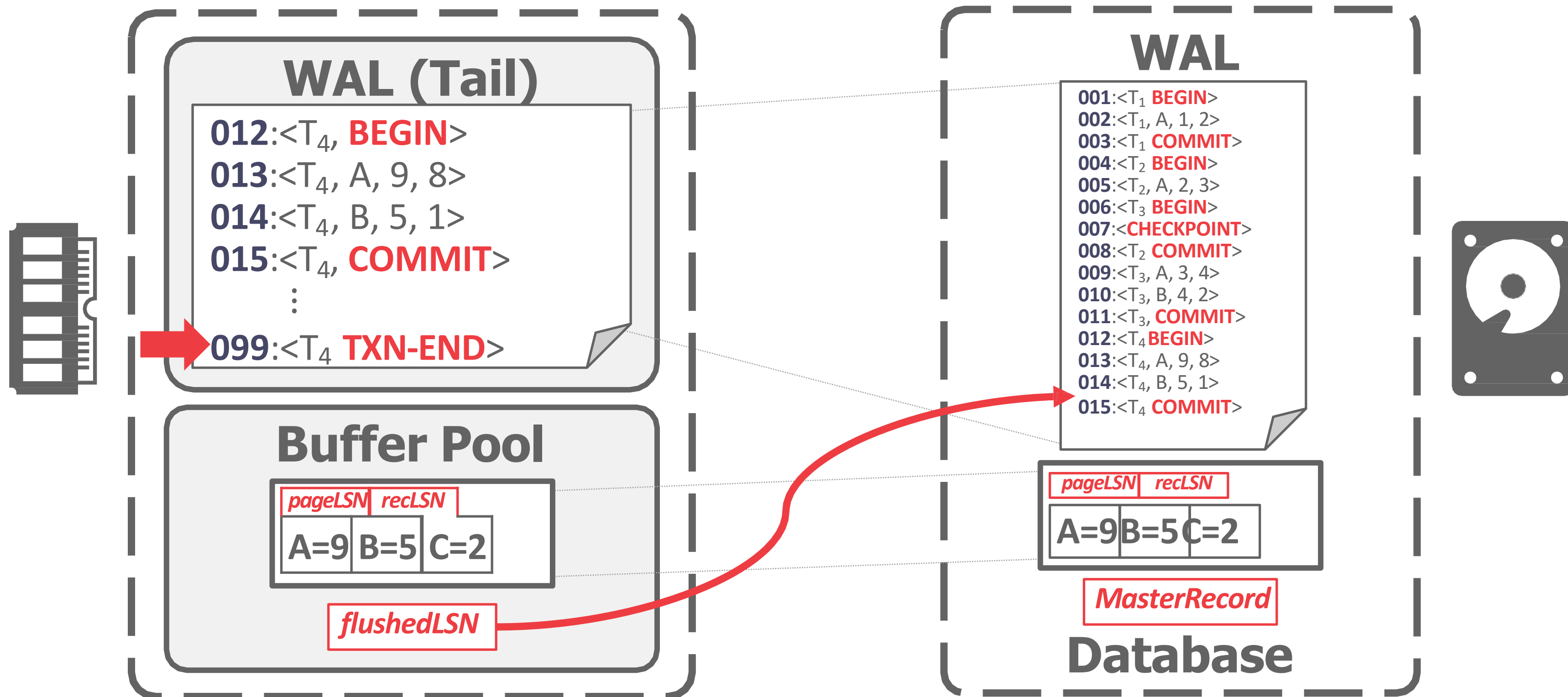
MasterRecord

Database

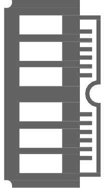
TRANSACTION COMMIT



TRANSACTION COMMIT



TRANSACTION COMMIT



We can trim the in-memory log up to flushedLSN

012:<T_4, BEGI
013:<T_4, A, 9, 8>
014:<T_4, B, 5, 1>
015:<T_4, COMMIT>

099:<T_4 TXN-END>

Buffer Pool

pageLSN	recLSN	
A=9	B=5	C=2

flushedLSN

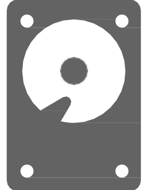
WAL

001:<T_1 BEGIN>
002:<T_1, A, 1, 2>
003:<T_1 COMMIT>
004:<T_2 BEGIN>
005:<T_2, A, 2, 3>
006:<T_3 BEGIN>
007:<CHECKPOINT>
008:<T_2 COMMIT>
009:<T_3, A, 3, 4>
010:<T_3, B, 4, 2>
011:<T_3, COMMIT>
012:<T_4BEGIN>
013:<T_4, A, 9, 8>
014:<T_4, B, 5, 1>
015:<T_4 COMMIT>

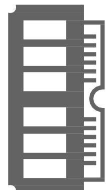
pageLSN	recLSN	
A=9	B=5	C=2

MasterRecord

Database



TRANSACTION COMMIT



We can trim the in-memory log up to flushedLSN

⋮

099:<T_4 TXN-END>

Buffer Pool

pageLSN	recLSN	
A=9	B=5	C=2

flushedLSN

WAL

001:<T_1 BEGIN>
002:<T_1, A, 1, 2>
003:<T_1 COMMIT>
004:<T_2 BEGIN>
005:<T_2, A, 2, 3>
006:<T_3 BEGIN>
007:<CHECKPOINT>
008:<T_2 COMMIT>
009:<T_3, A, 3, 4>
010:<T_3, B, 4, 2>
011:<T_3, COMMIT>
012:<T_4BEGIN>
013:<T_4, A, 9, 8>
014:<T_4, B, 5, 1>
015:<T_4 COMMIT>

pageLSN	recLSN	
A=9	B=5	C=2

MasterRecord

Database

TODAY'S AGENDA

~~Log Sequence Numbers~~

Recovery Algorithm

ARIES – RECOVERY PHASES

Phase #1 – Analysis

→ Read WAL from last **MasterRecord** to identify dirty pages in the buffer pool and active txns at the time of the crash.

Phase #2 – Redo

→ Repeat all actions starting from an appropriate point in the log (even txns that will abort).

Phase #3 – Undo

→ Reverse the actions of txns that did not commit before the crash.

ACTIVE TRANSACTION TABLE

One entry per currently active txn.

→ **txnId**: Unique txn identifier.

→ **status**: The current "mode" of the txn.

→ **lastLSN**: Most recent ***LSN*** created by txn.

Entry removed after the TXN-END message.

Txn Status Codes:

→ **R**→ Running

→ **C**→ Committing

→ **U**→ Candidate for Undo

DIRTY PAGE TABLE

Keep track of which pages in the buffer pool contain changes from transactions that have not been flushed to disk.

One entry per dirty page in the buffer pool:

→ **recLSN**: The ***LSN*** of the log record that first caused the page to be dirty.

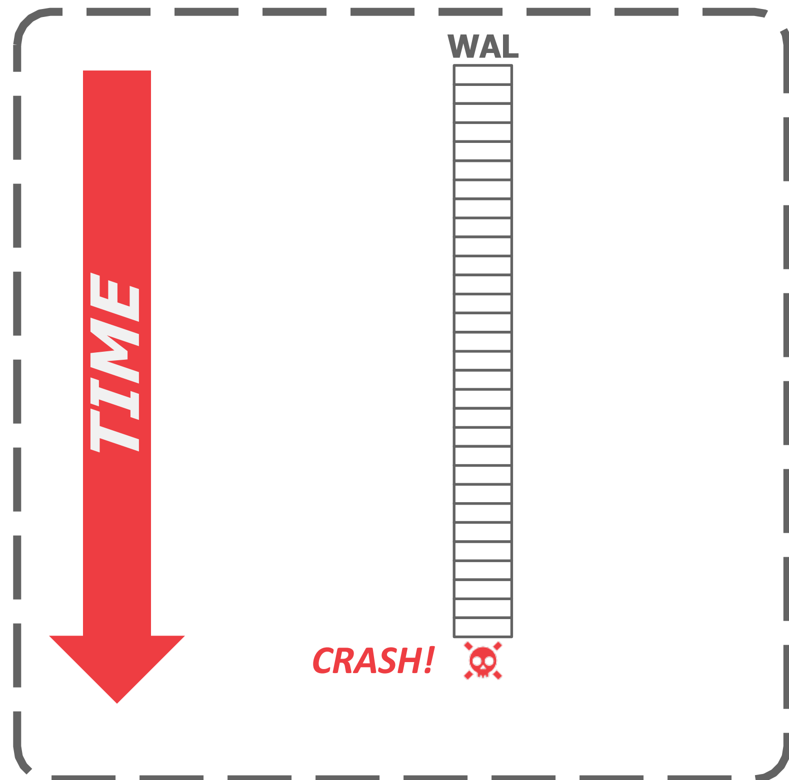
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



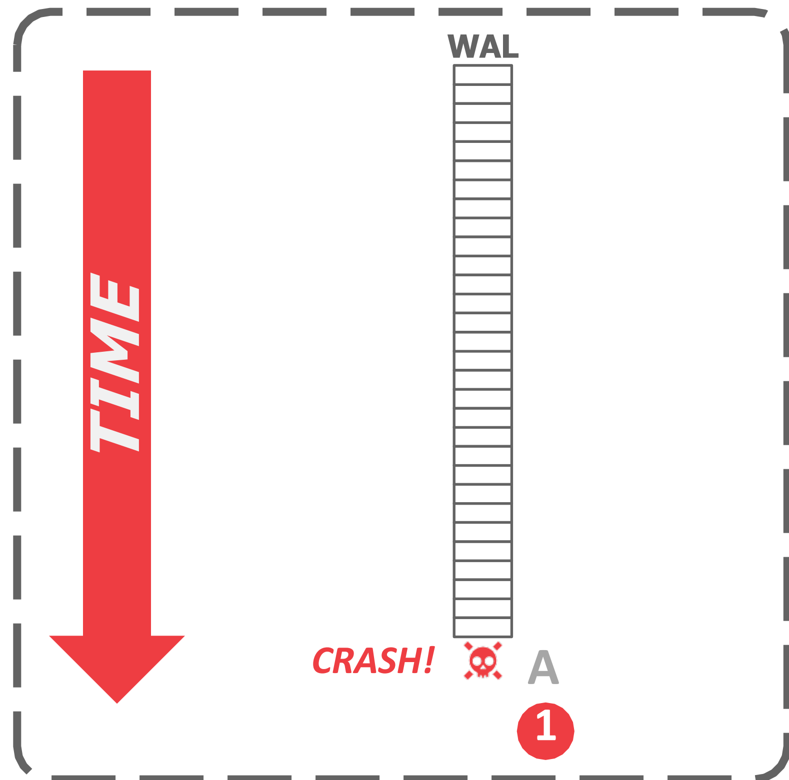
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



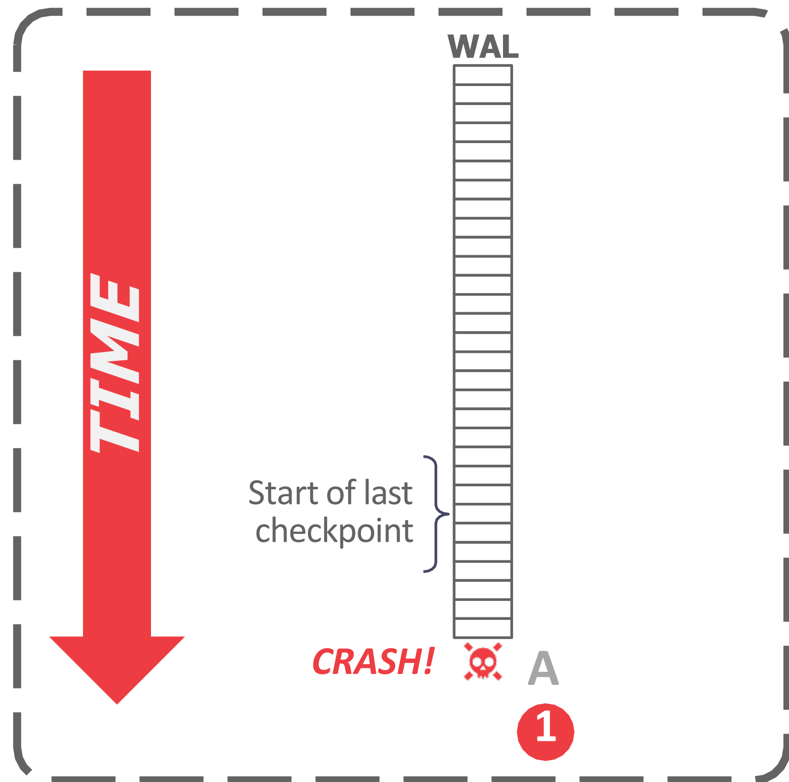
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



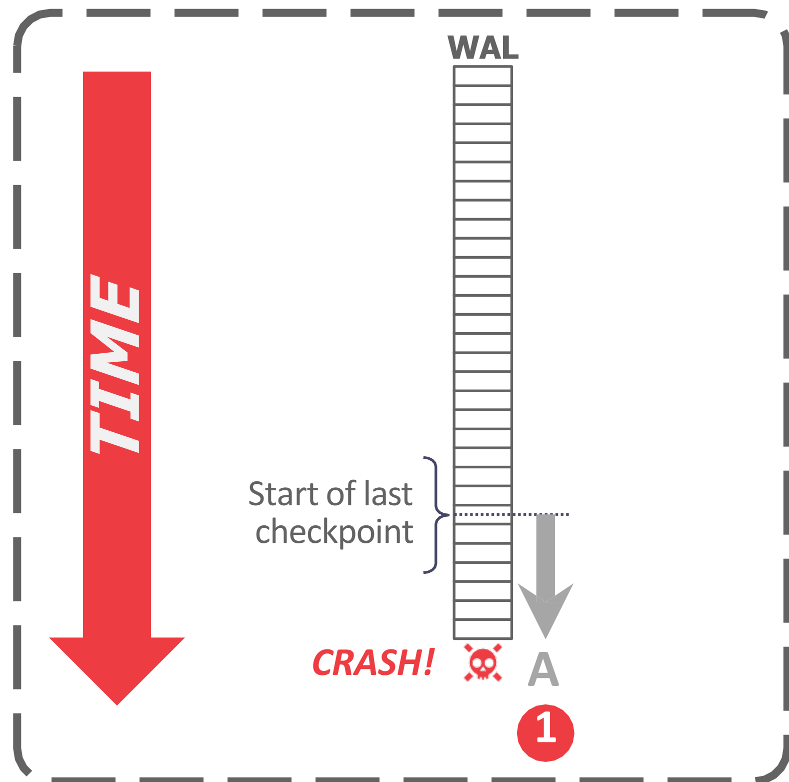
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



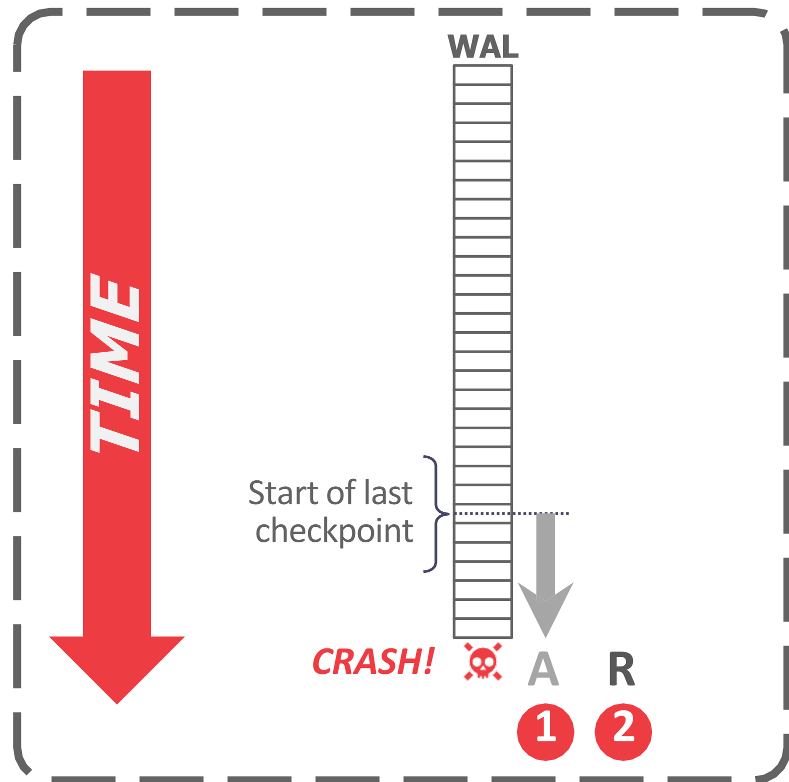
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



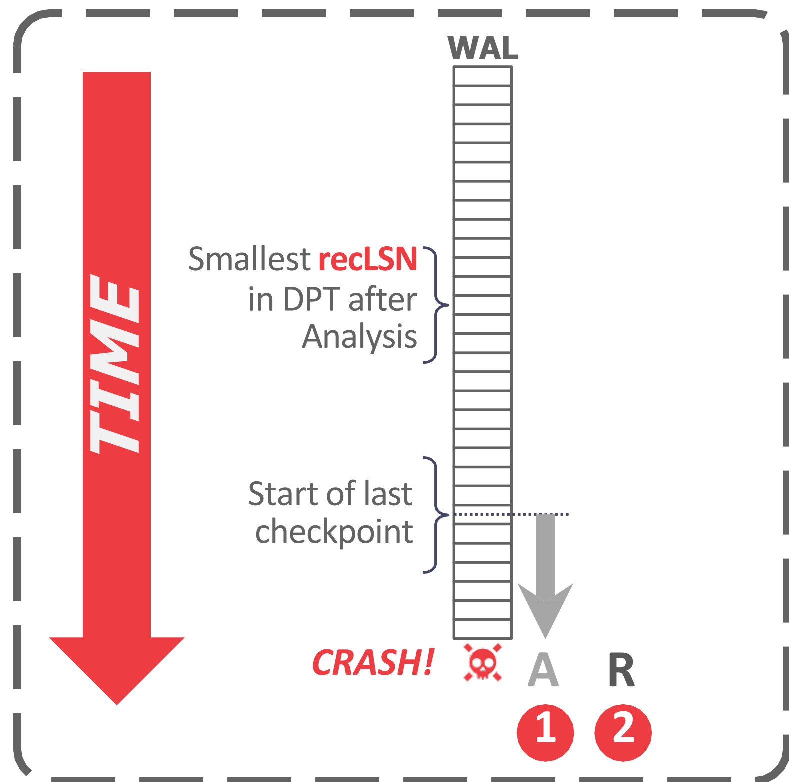
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



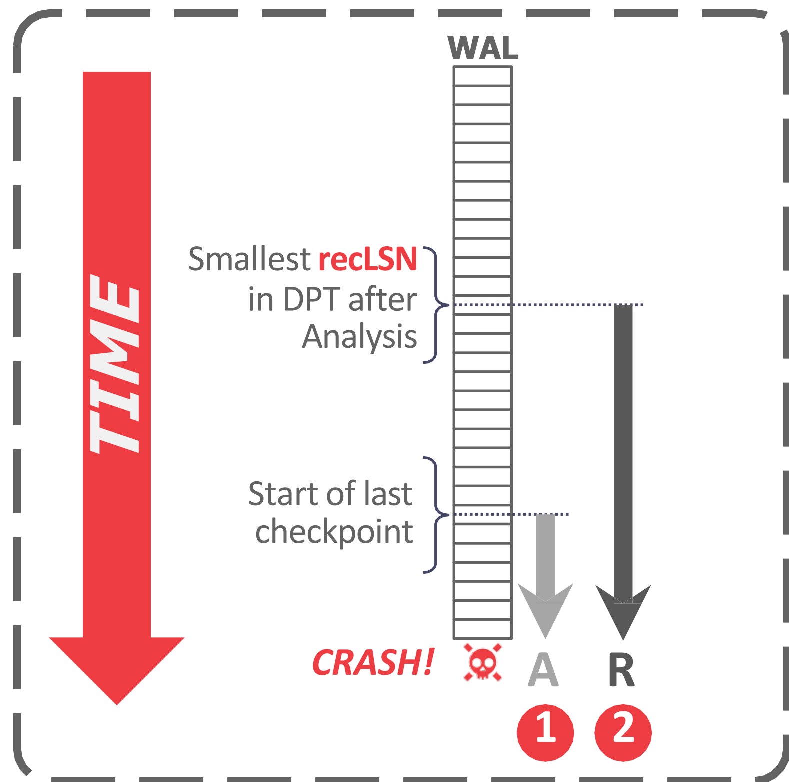
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



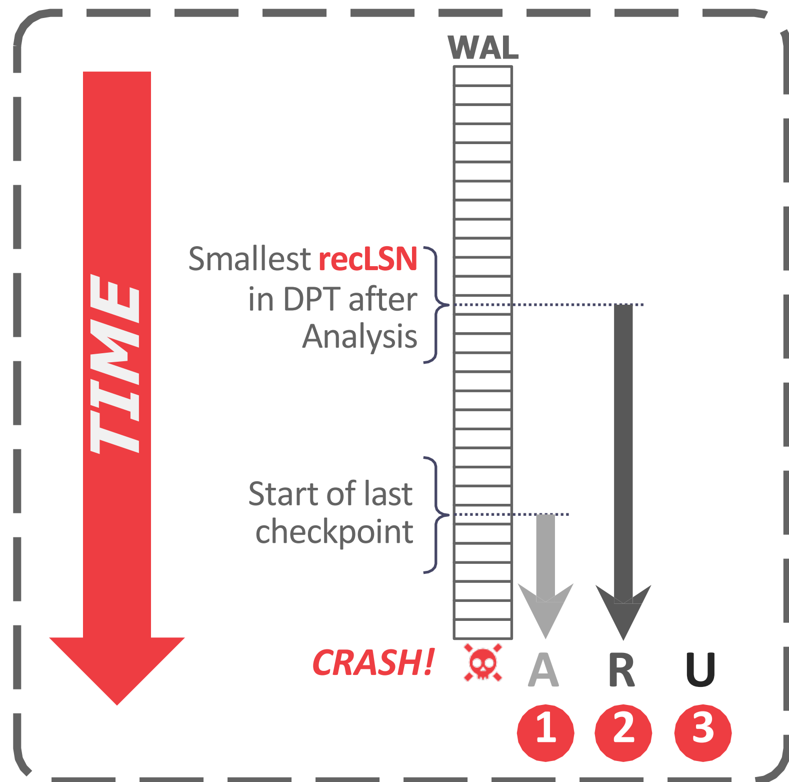
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



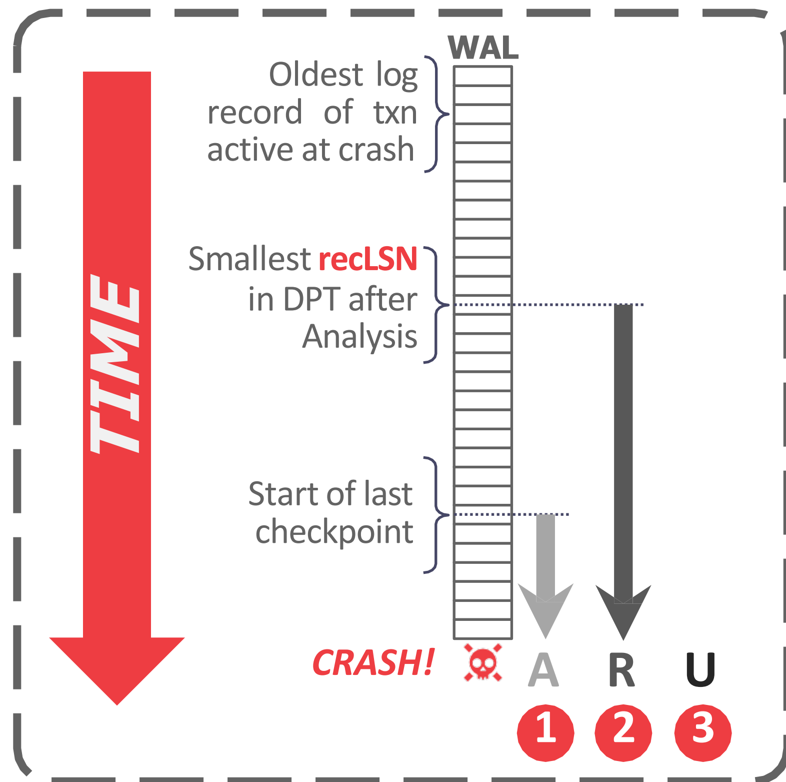
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT** found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



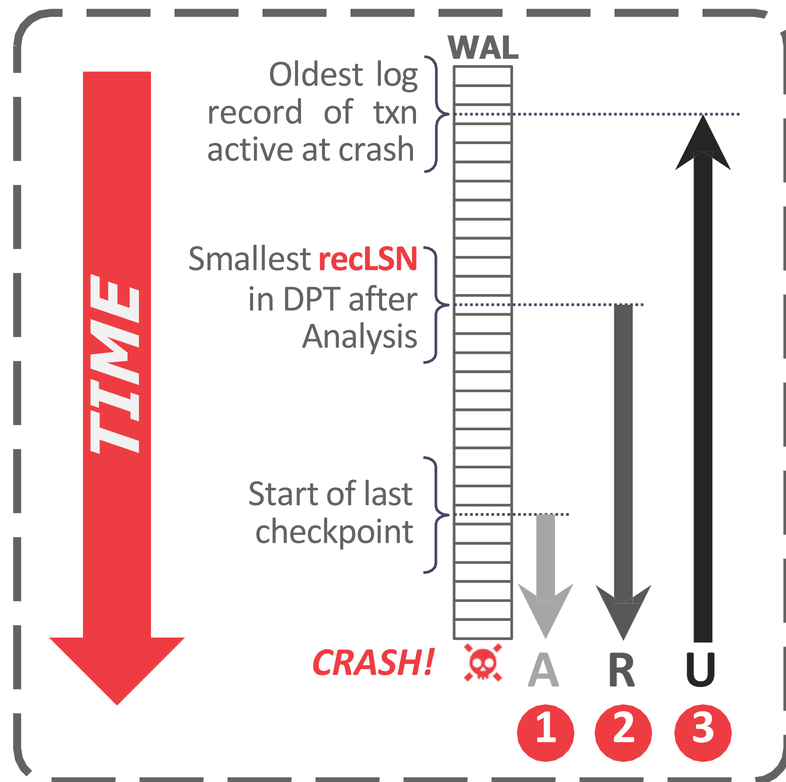
ARIES – OVERVIEW

Start from last **BEGIN-CHECKPOINT**
found via **MasterRecord**.

Analysis: Figure out which txns committed or failed since checkpoint.

Redo: Repeat all actions.

Undo: Reverse effects of failed txns.



ANALYSIS PHASE

Scan log forward from last successful checkpoint.

If you find a **TXN-END** record, remove its corresponding txn from **ATT**.

All other records:

→ Add txn to **ATT** with status **UNDO**.

→ On commit, change txn status to **COMMIT**.

For **UPDATE** records:

→ If page **P** not in **DPT**, add **P** to **DPT**, set its **recLSN**=**LSN**.

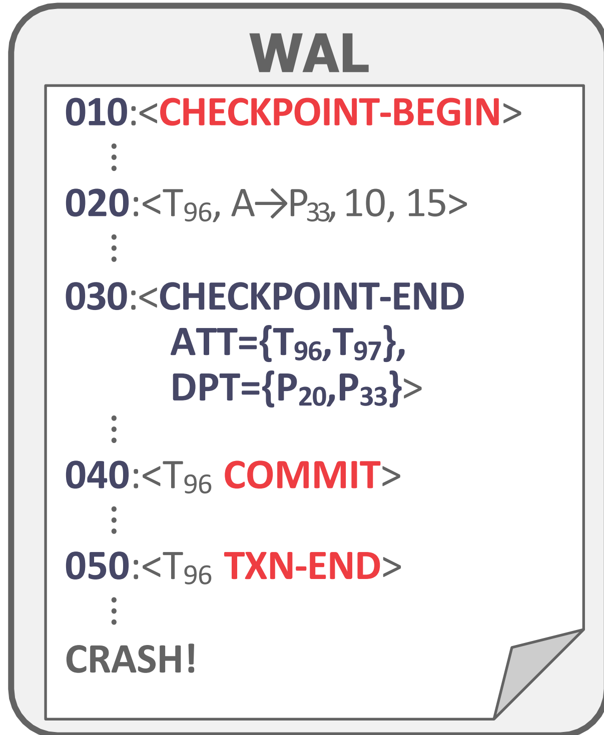
ANALYSIS PHASE

At end of the Analysis Phase:

→ **ATT** identifies which txns were active at time of crash.

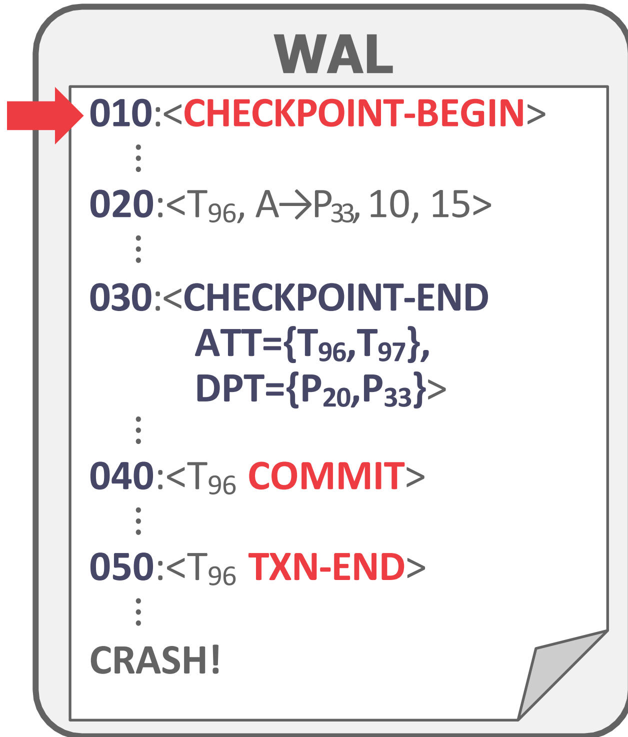
→ **DPT** identifies which dirty pages might not have made it to disk.

ANALYSIS PHASE EXAMPLE



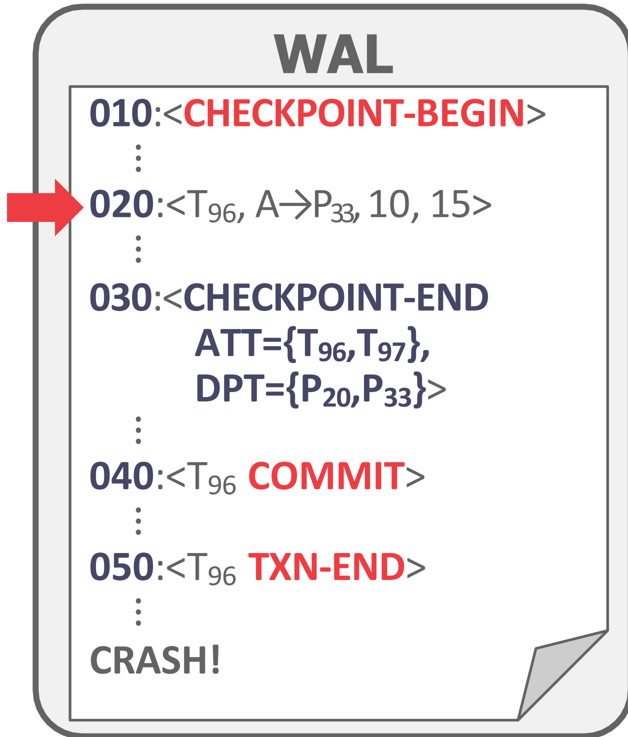
LSN	ATT	DPT
010		
020		
030		
040		
050		

ANALYSIS PHASE EXAMPLE



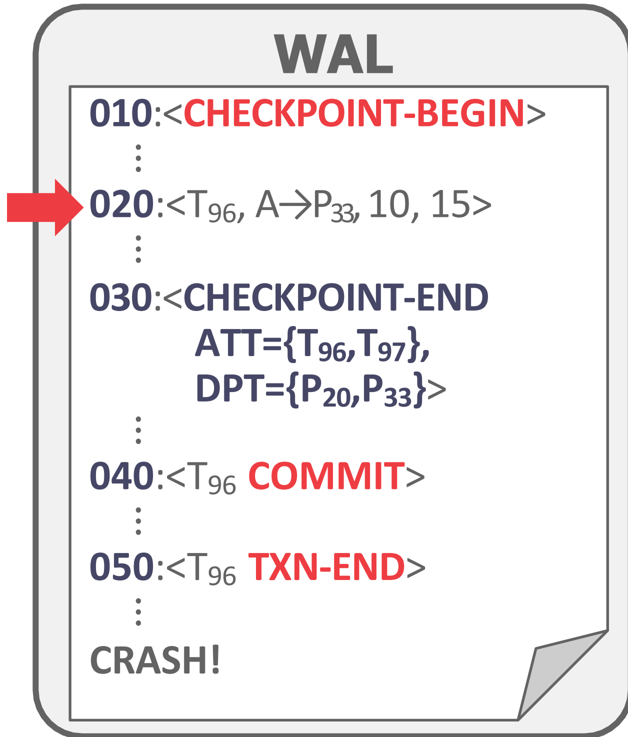
LSN	ATT	DPT
010		
020		
030		
040		
050		

ANALYSIS PHASE EXAMPLE



LSN	ATT	DPT
010		
020		
030		
040		
050		

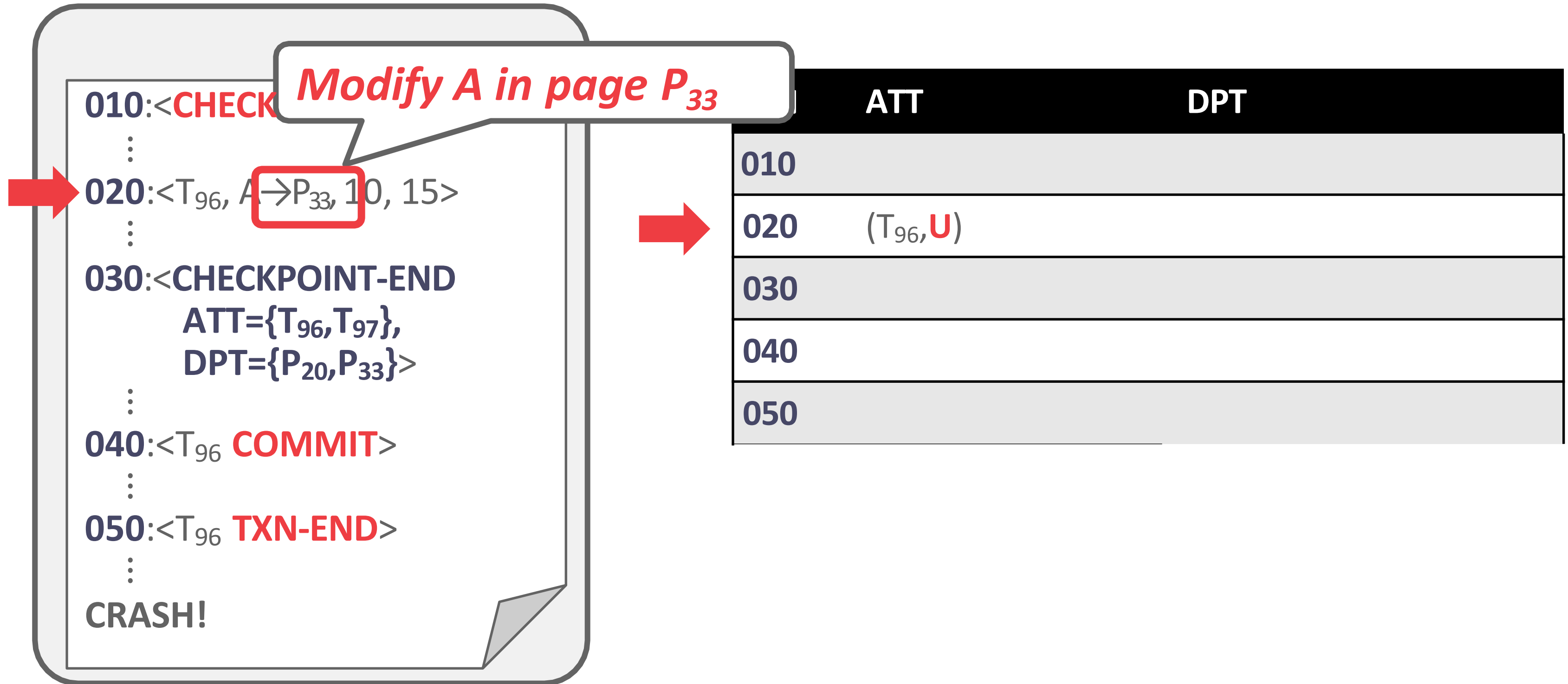
ANALYSIS PHASE EXAMPLE



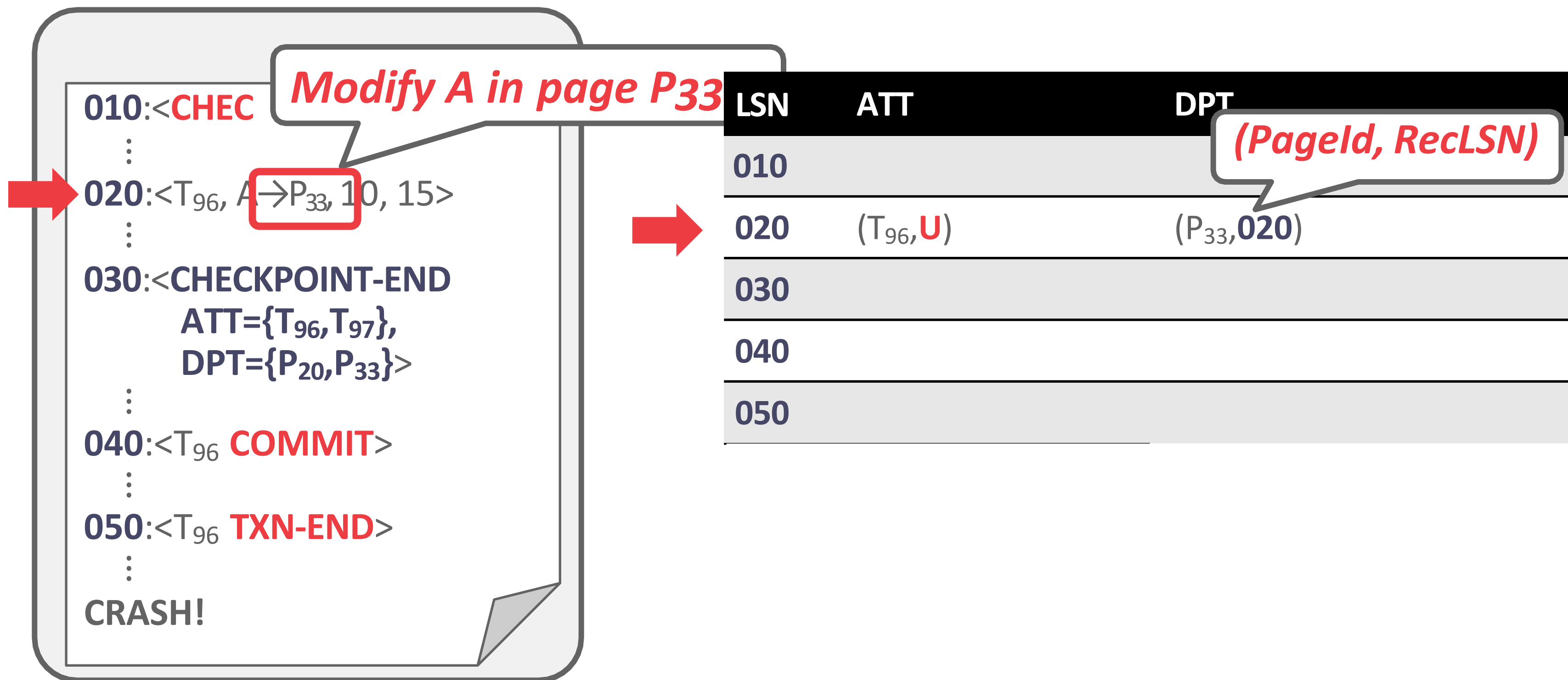
(TxnId, Status)

LSN	ATT	DPT
010		
020	$(T_{96}, \mathbf{U})$	
030		
040		
050		

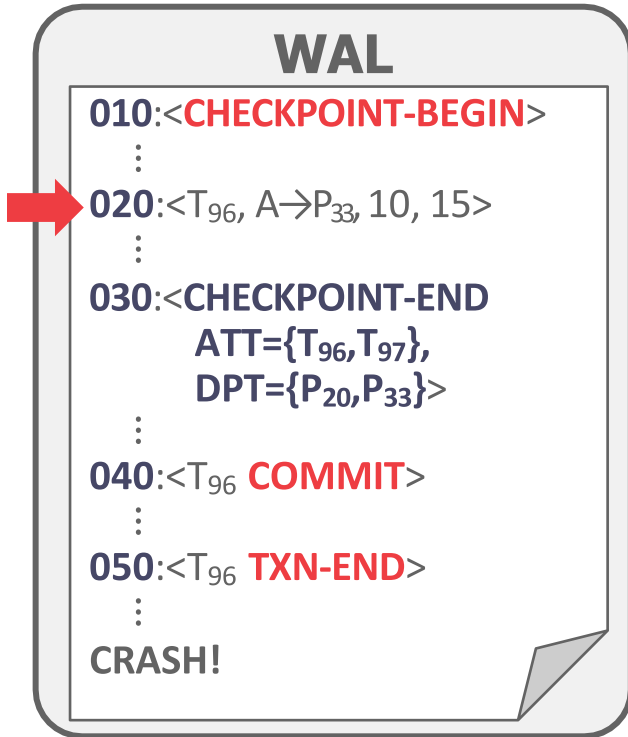
ANALYSIS PHASE EXAMPLE



ANALYSIS PHASE EXAMPLE

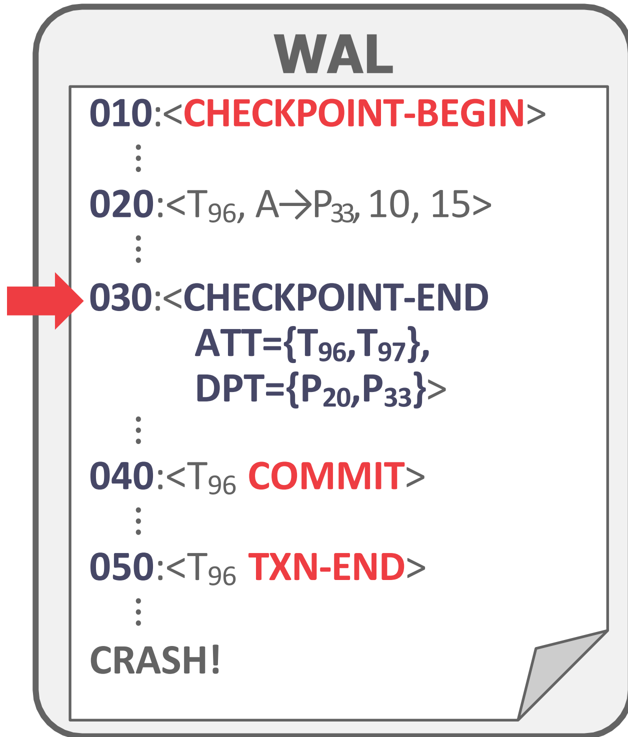


ANALYSIS PHASE EXAMPLE



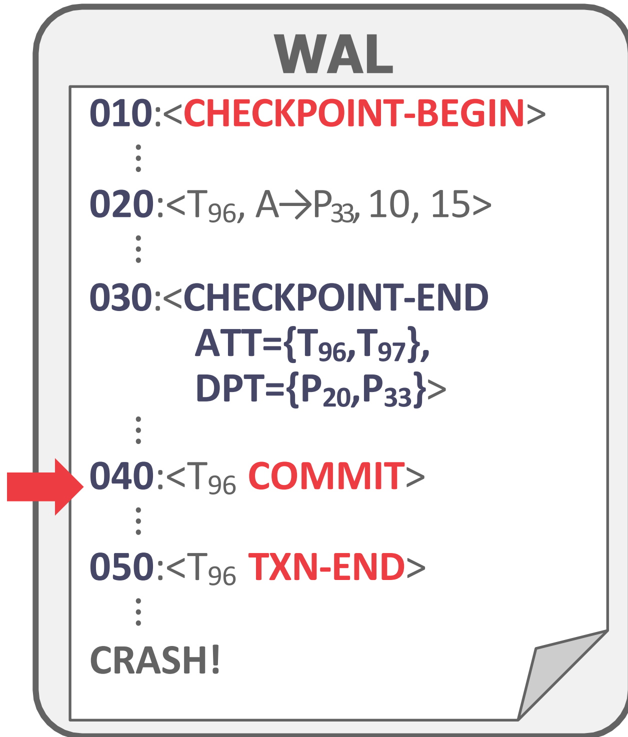
LSN	ATT	DPT
010		
020	(T_{96},U)	(P_{33},020)
030		
040		
050		

ANALYSIS PHASE EXAMPLE



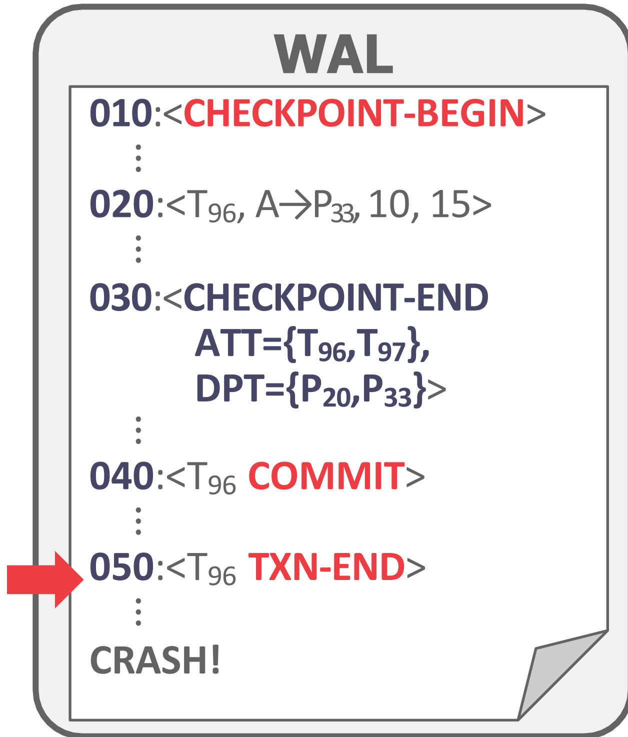
LSN	ATT	DPT
010		
020	(T_{96},**U**)	(P_{33},**020**)
030	(T_{96},**U**), (T_{97},**U**)	(P_{33},**020**), (P_{20},**008**)
040		
050		

ANALYSIS PHASE EXAMPLE



LSN	ATT	DPT
010		
020	(T_{96},U)	(P_{33},020)
030	(T_{96},U), (T_{97},U)	(P_{33},020), (P_{20},008)
040	(T_{96},C), (T_{97},U)	(P_{33},020), (P_{20},008)
050		

ANALYSIS PHASE EXAMPLE



LSN	ATT	DPT
010		
020	(T_{96},**U**)	(P_{33},**020**)
030	(T_{96},**U**), (T_{97},**U**)	(P_{33},**020**), (P_{20},**008**)
040	(T_{96},**C**), (T_{97},**U**)	(P_{33},**020**), (P_{20},**008**)
050	(T_{97},**U**)	(P_{33},**020**), (P_{20},**008**)

REDO PHASE

The goal is to repeat history to reconstruct state at the moment of the crash:

→ Reapply all updates (even aborted txns!) and redo **CLRs**.

There are techniques that allow the DBMS to avoid unnecessary reads/writes, but we will ignore that in this lecture...

REDO PHASE

Scan forward from the log record containing smallest **recLSN** in **DPT**.

For each update log record or ***CLR*** with a given ***LSN***, redo the action unless:

→ Affected page is not in **DPT**, <u>or</u>

→ Affected page is in **DPT** but that record's ***LSN*** is less than the page's **recLSN**.

REDO PHASE

To redo an action:

→ Reapply logged action.

→ Set **pageLSN** to log record's ***LSN***.

→ No additional logging, no forced flushes!

At the end of Redo Phase, write **TXN-END** log records for all txns with status **C** and remove them from the **ATT**.

UNDO PHASE

Undo all txns that were active at the time of crash and therefore will never commit.

→ These are all the txns with **U** status in the **ATT** after the Analysis Phase.

Process them in reverse ***LSN*** order using the **lastLSN** to speed up traversal.

Write a **CLR** for every modification.

ADDITIONAL CRASH ISSUES (1)

What does the DBMS do if it crashes during recovery in the Analysis Phase?

What does the DBMS do if it crashes during recovery in the Redo Phase?

ADDITIONAL CRASH ISSUES (1)

What does the DBMS do if it crashes during recovery in the Analysis Phase?

→ Nothing. Just run recovery again.

What does the DBMS do if it crashes during recovery in the Redo Phase?

ADDITIONAL CRASH ISSUES (1)

What does the DBMS do if it crashes during recovery in the Analysis Phase?

→ Nothing. Just run recovery again.

What does the DBMS do if it crashes during recovery in the Redo Phase?

→ Again nothing. Redo everything again.

ADDITIONAL CRASH ISSUES (2)

How can the DBMS improve performance during recovery in the Redo Phase?

How can the DBMS improve performance during recovery in the Undo Phase?

ADDITIONAL CRASH ISSUES (2)

How can the DBMS improve performance during recovery in the Redo Phase?

→ Assume that it is not going to crash again and flush all changes to disk asynchronously in the background.

How can the DBMS improve performance during recovery in the Undo Phase?

ADDITIONAL CRASH ISSUES (2)

How can the DBMS improve performance during recovery in the Redo Phase?

→ Assume that it is not going to crash again and flush all changes to disk asynchronously in the background.

How can the DBMS improve performance during recovery in the Undo Phase?

→ Lazily rollback changes before new txns access pages.

→ Rewrite the application to avoid long-running txns.

CONCLUSION

Mains ideas of ARIES:

→ WAL with **STEAL**/**NO-FORCE**

→ Redo everything since the earliest dirty page

→ Undo txns that never commit

→ Write **CLRs** when undoing, to survive failures during restarts

Log Sequence Numbers:

→ ***LSNs*** identify log records; linked into backwards chains per transaction via **prevLSN**.

→ **pageLSN** allows comparison of data page and log records.