

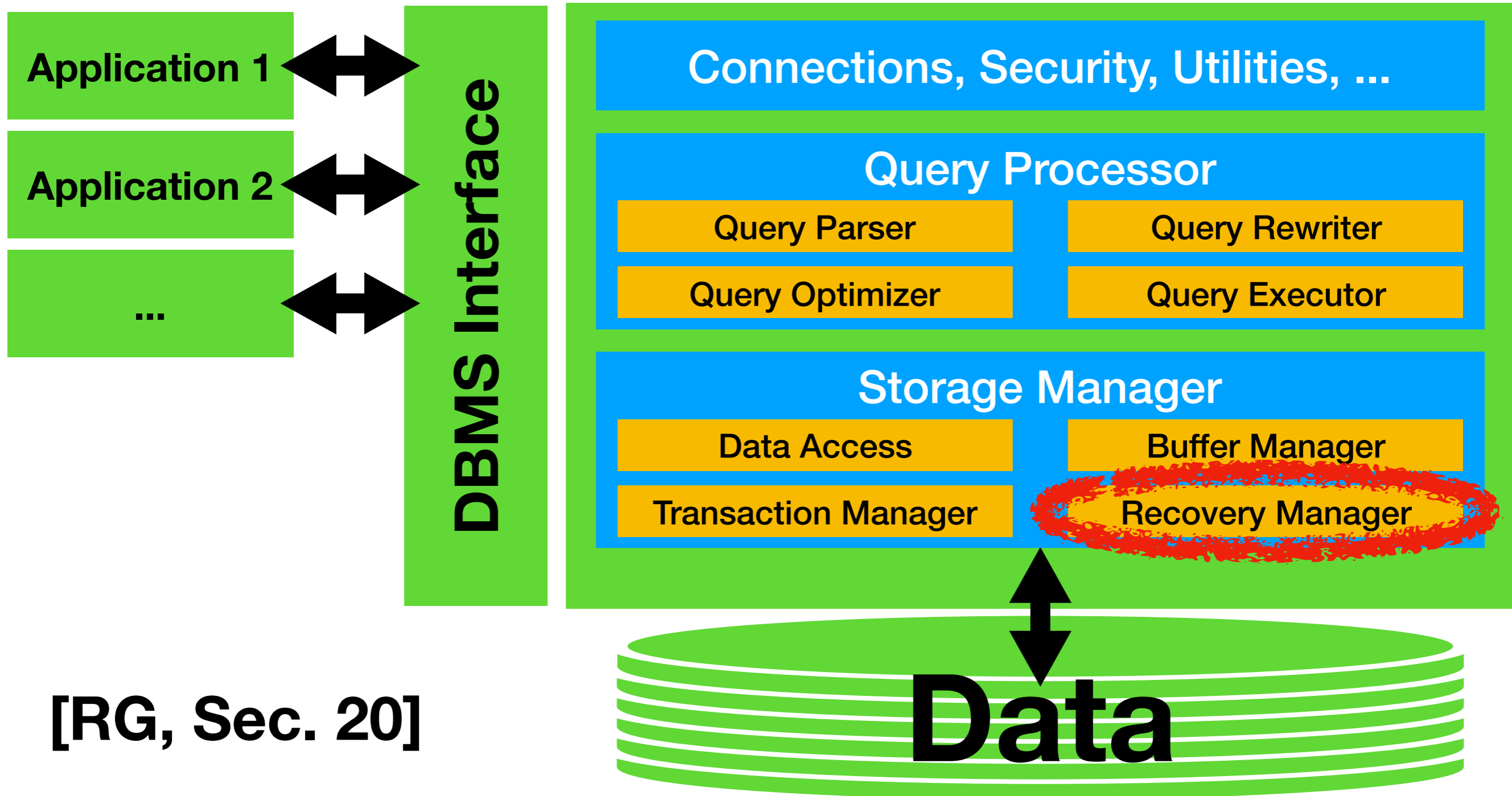
Recovery After System Crashes

Immanuel Trummer

itrummer@cornell.edu

www.itrummer.org

Database Management Systems (DBMS)

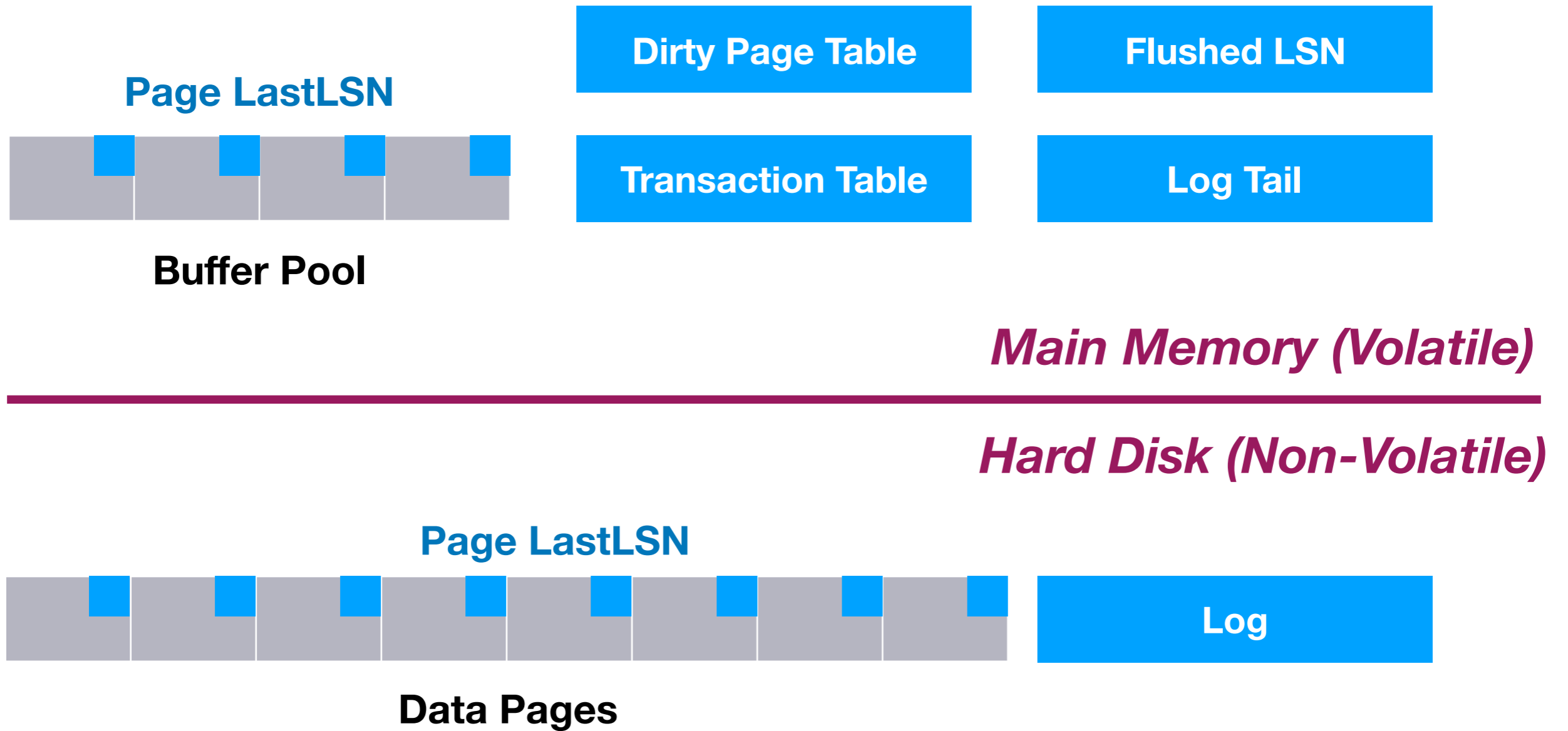


[RG, Sec. 20]

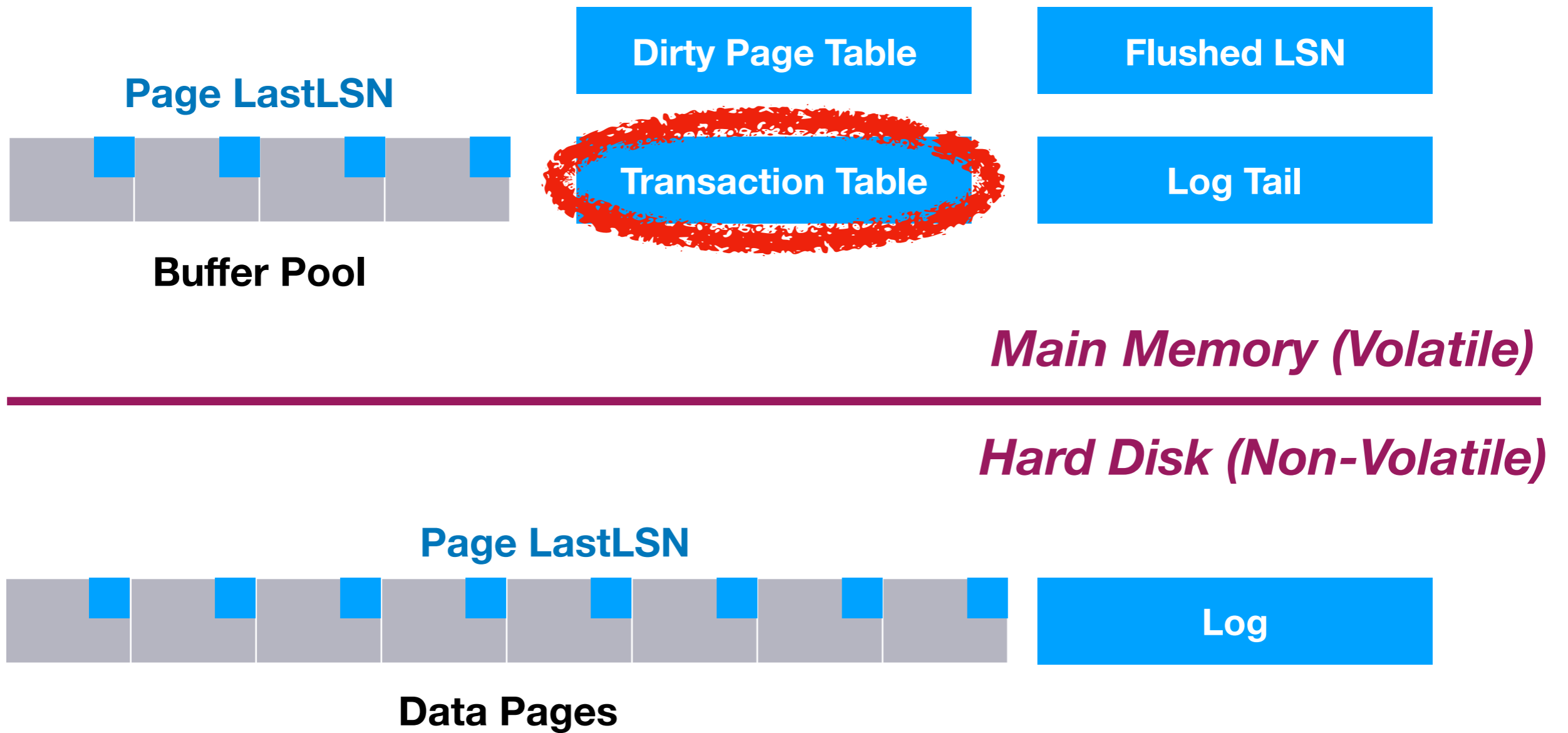
Outlook

- ARIES data structures
- **ARIES run time behavior**
- ARIES recovery algorithm

ARIES Data Structures



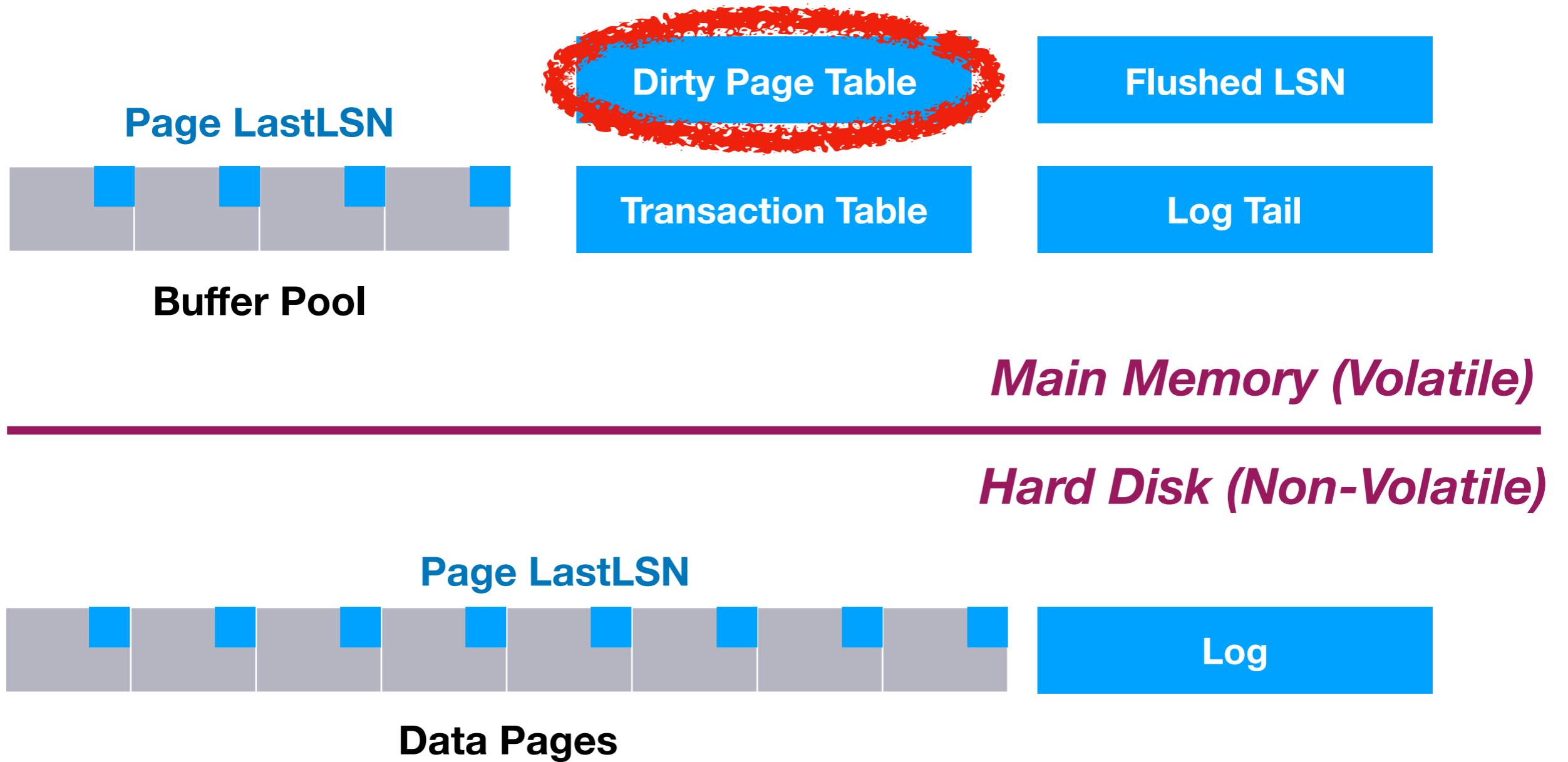
ARIES Data Structures



Updating Transaction Table

Scenario	Table Update
Transaction updates data	Update transaction lastLSN
Transaction commits	Update transaction status to committed
Transaction aborts	Update transaction status to aborted
Transaction ends	Remove from transaction table

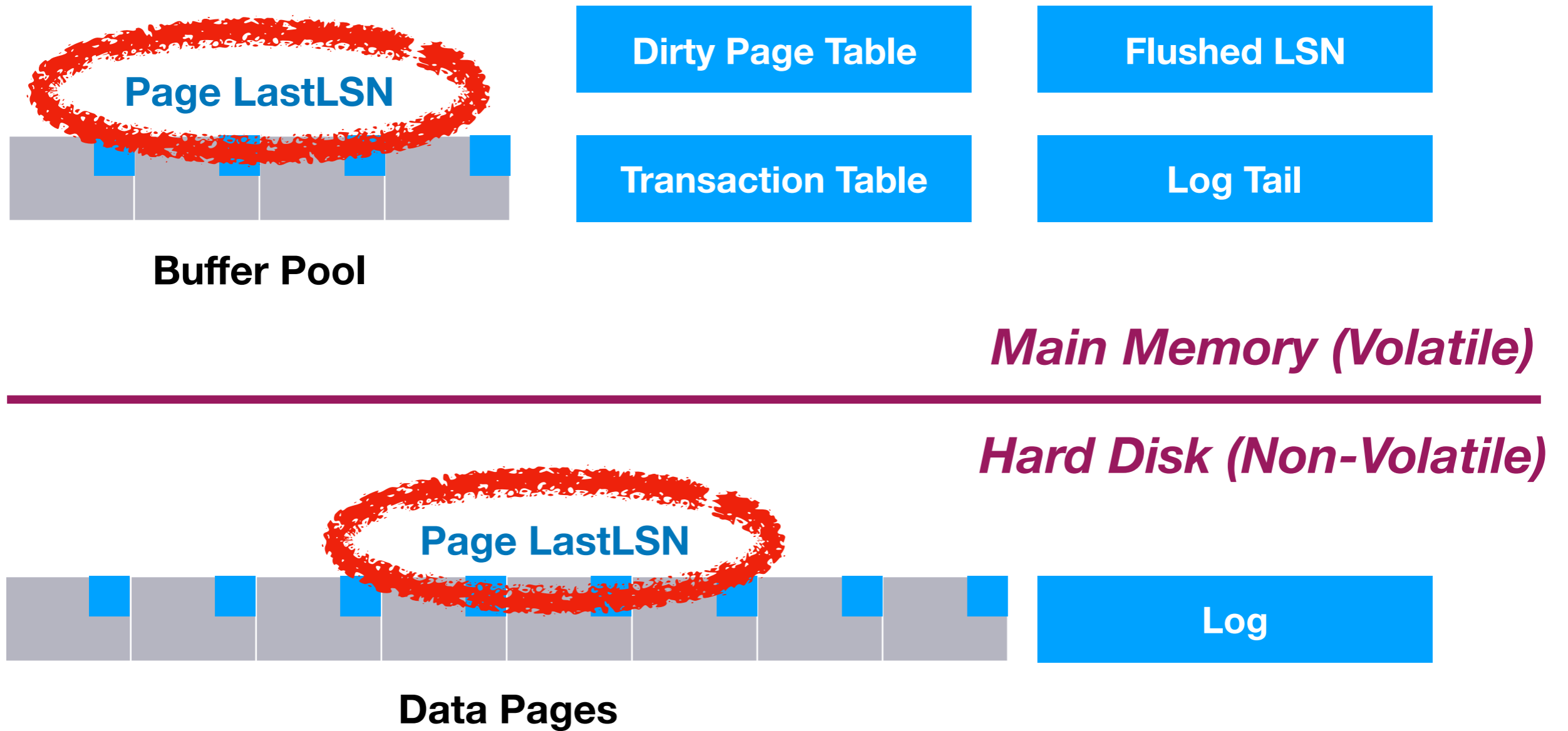
ARIES Data Structures



Updating Dirty Page Table

Scenario	Table Update
Page Changed	If page not in table: add page with current LSN as recLSN
Page Written to Disk	Remove page if in table

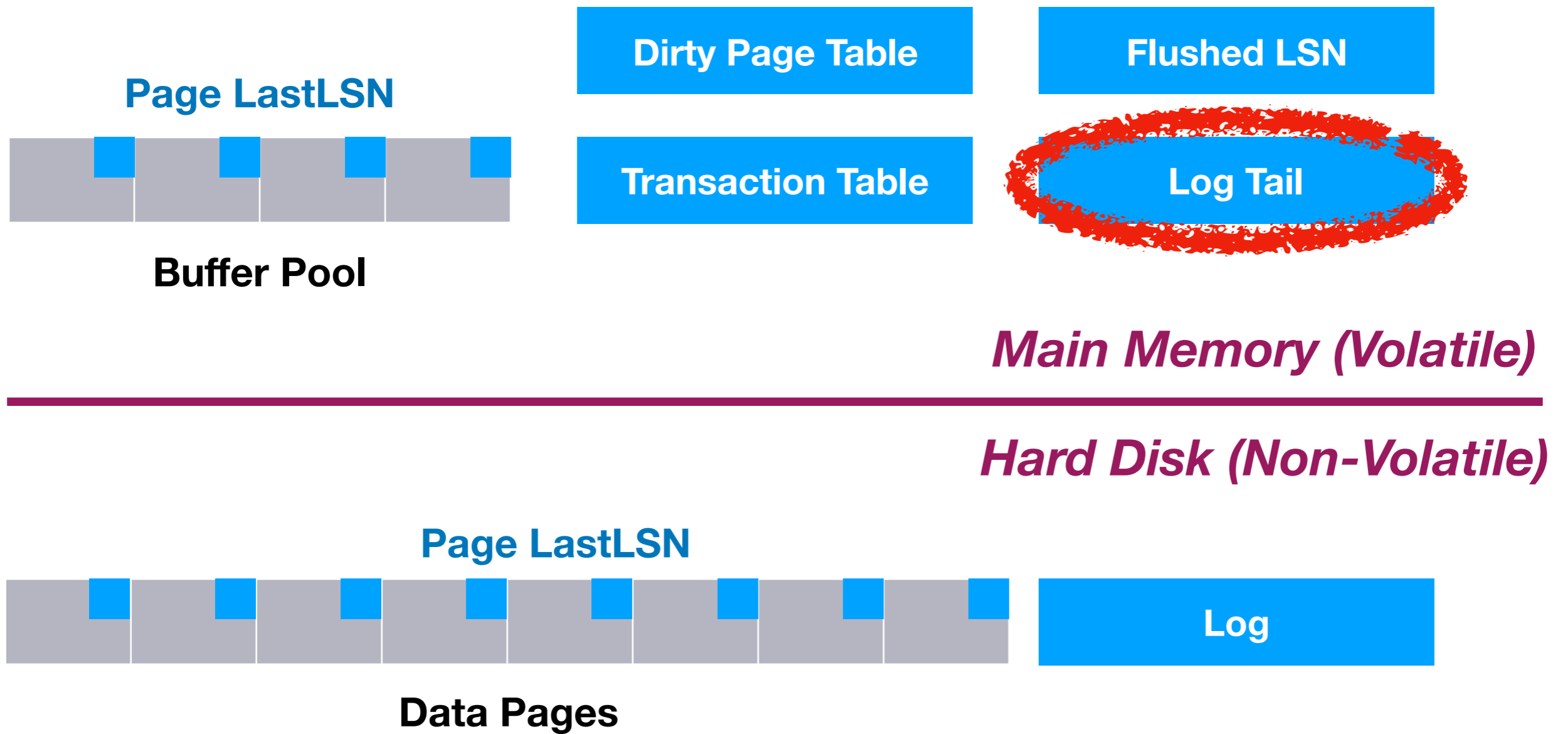
ARIES Data Structures



Updating PageLSN

Scenario	Table Update
Data update	Update PageLSN in memory
Page Written to Disk	PageLSN is copied to disk

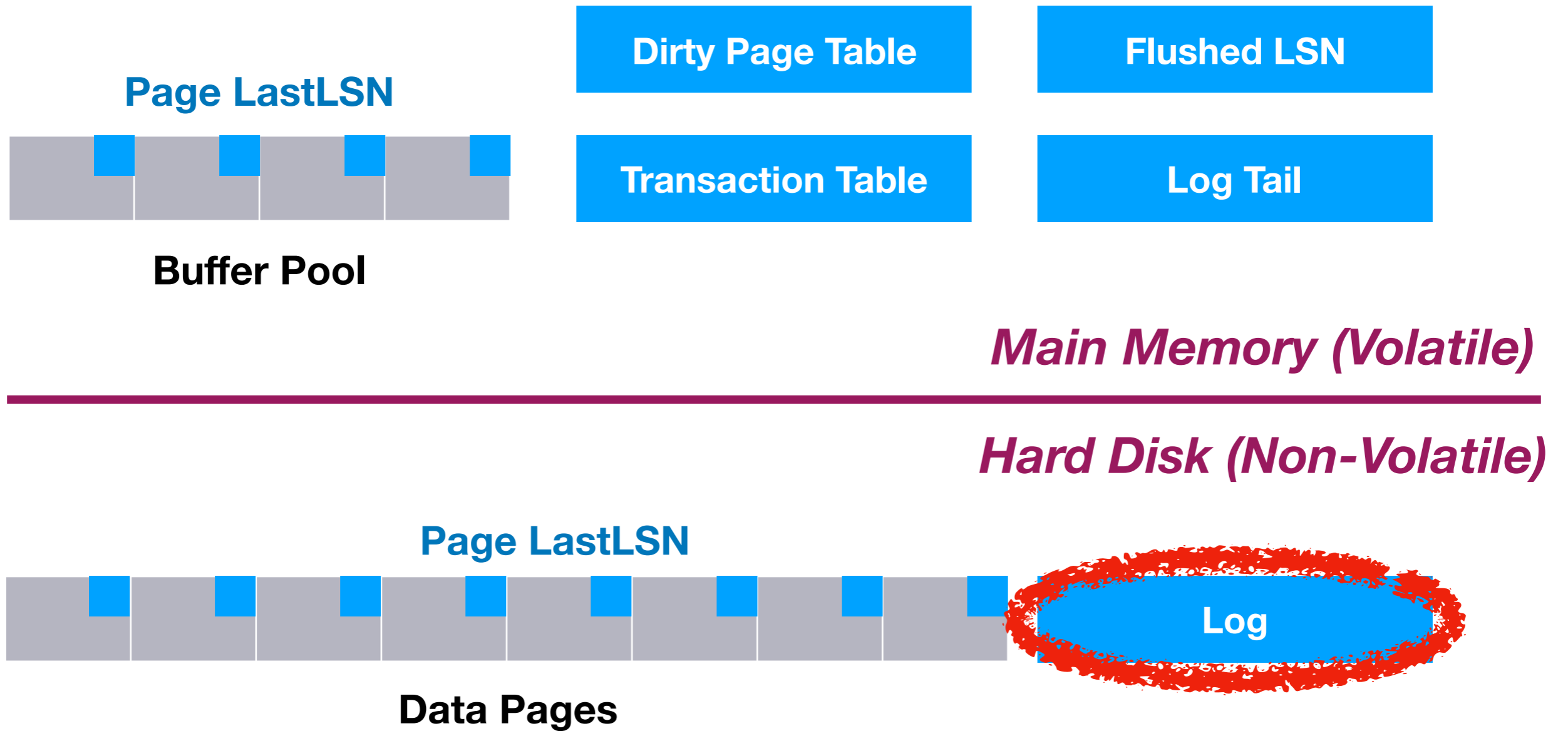
ARIES Data Structures



Updating Log Tail Buffer

Scenario	Log Update
Transaction updates data	Write update log entry
Transaction commits	Write commit log entry
Transaction aborts	Write abort log entry
Undo transaction update	Write compensation log record
Finished transaction cleanup	Write end log record

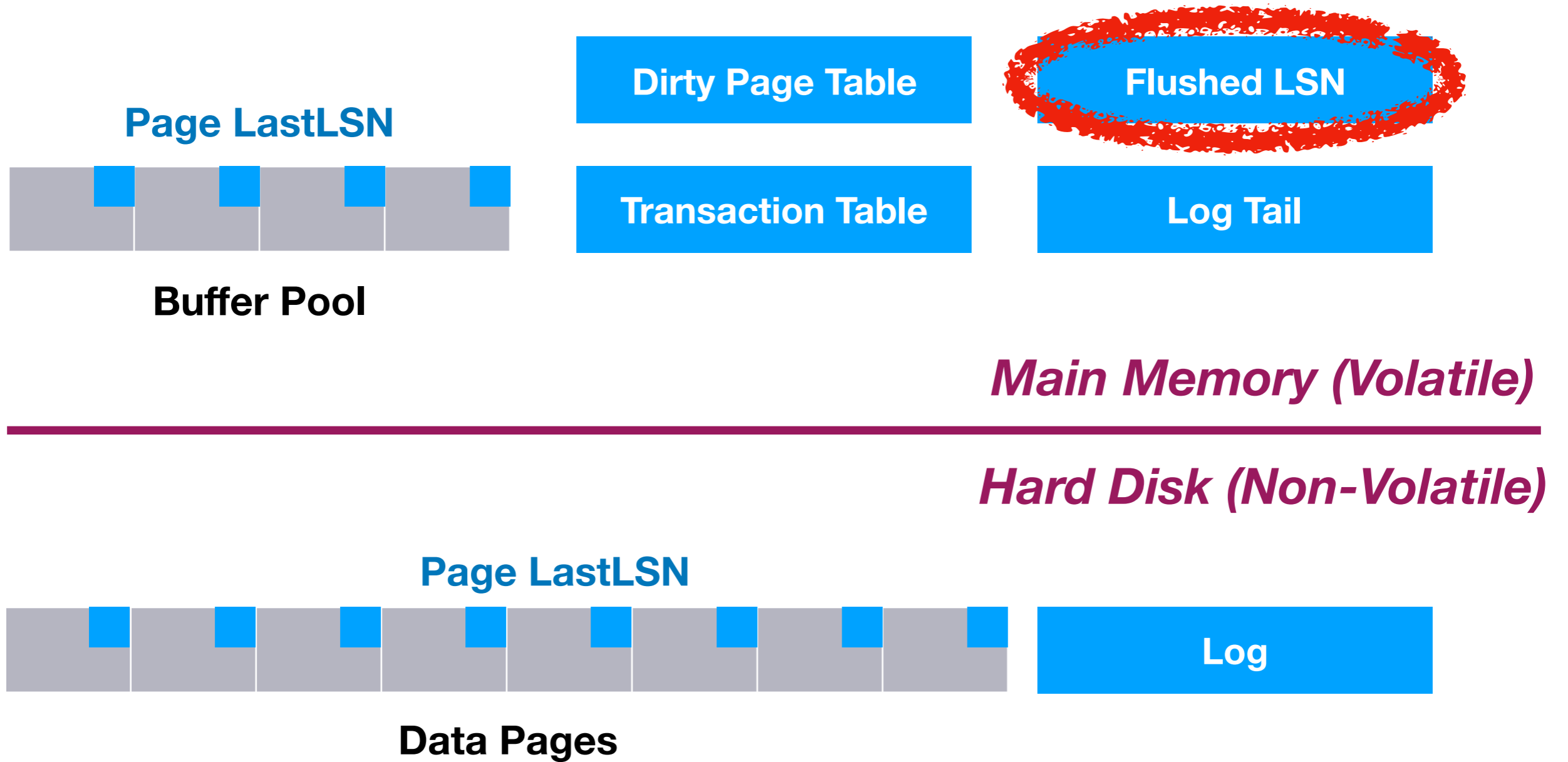
ARIES Data Structures



Updating Log on Disk

Scenario	Disk Log Update
Transaction commits	Before commit: Flush log entries up to last transaction entry
Page written to disk	Before writing: Flush log entries until last entry affecting page (pageLSN)

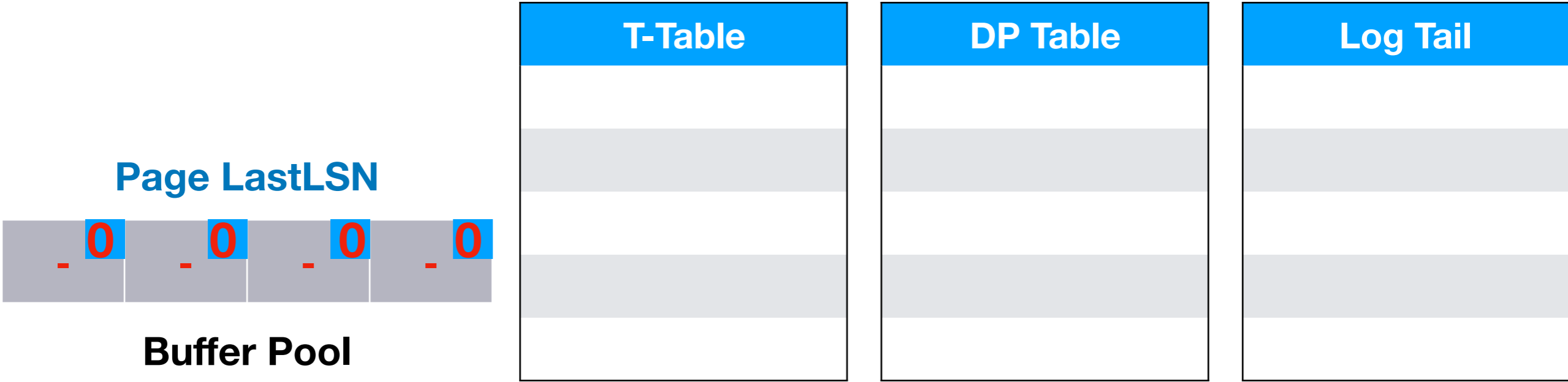
ARIES Data Structures



Updating FlushedLSN

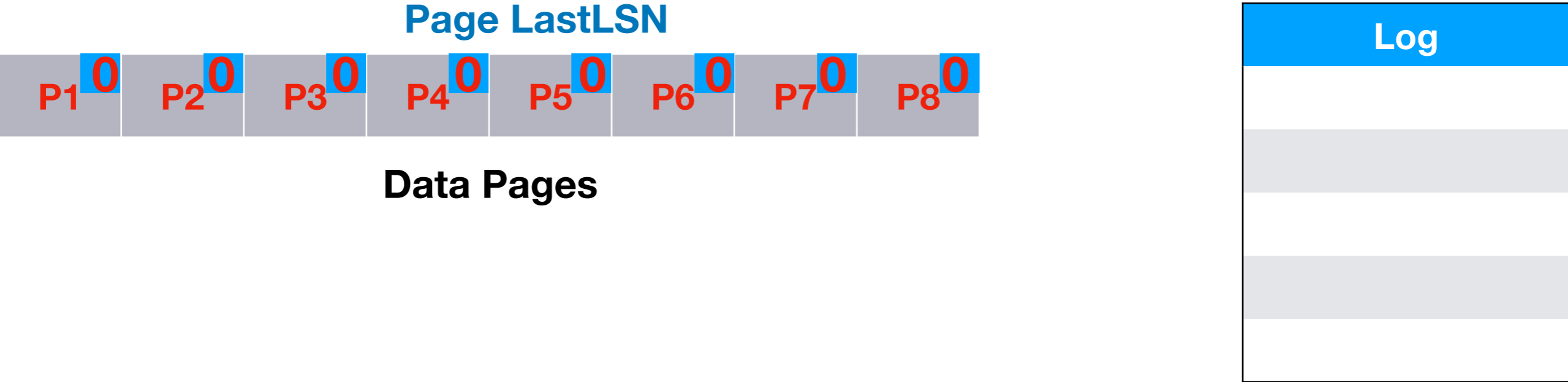
Scenario	Update to FlushedLSN
Log written to hard disk until LSN X	Update FlushedLSN to X

ARIES Example (Run Time)

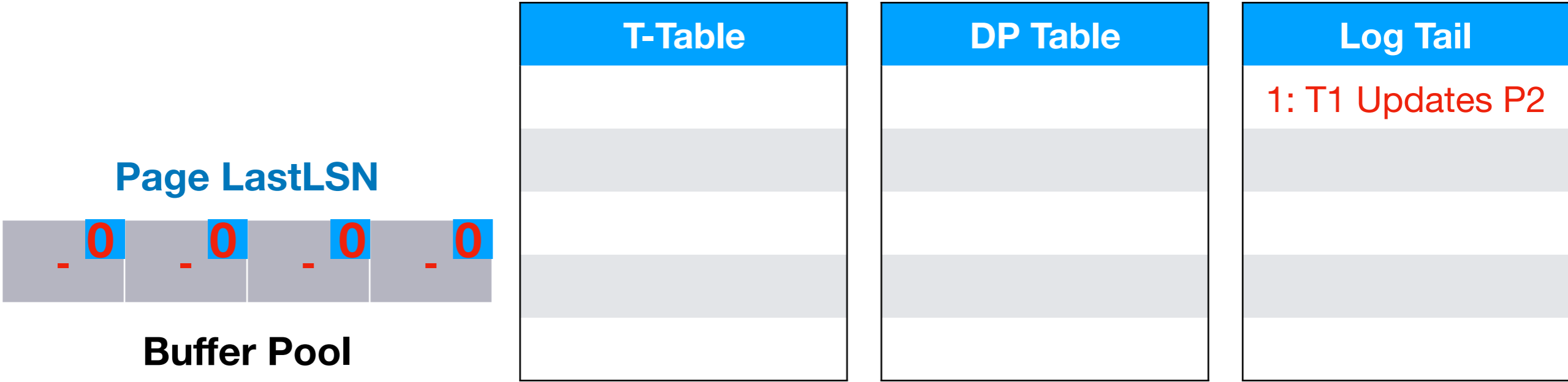


Main Memory (Volatile)

Hard Disk (Non-Volatile)

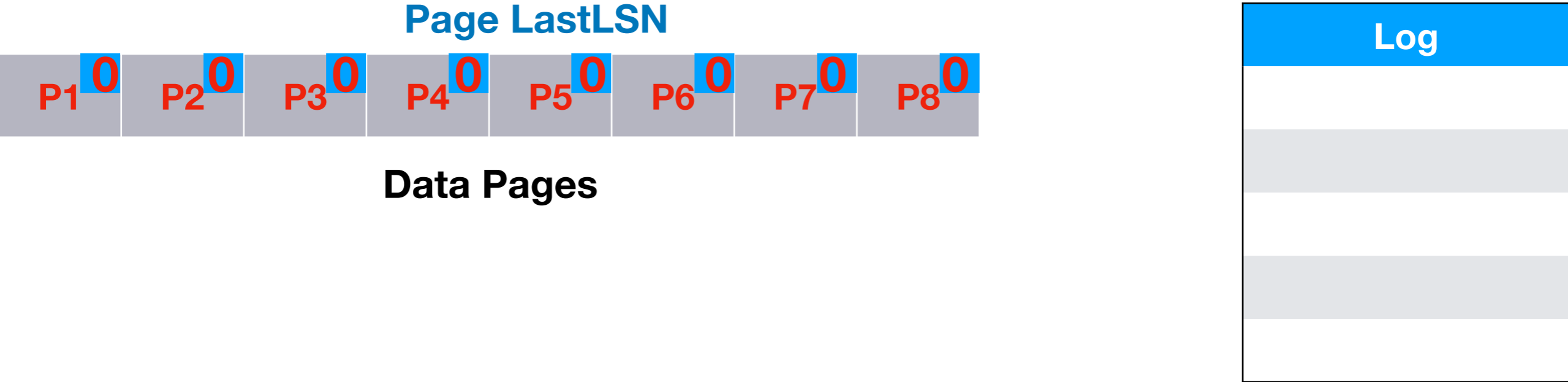


ARIES Example (Run Time)

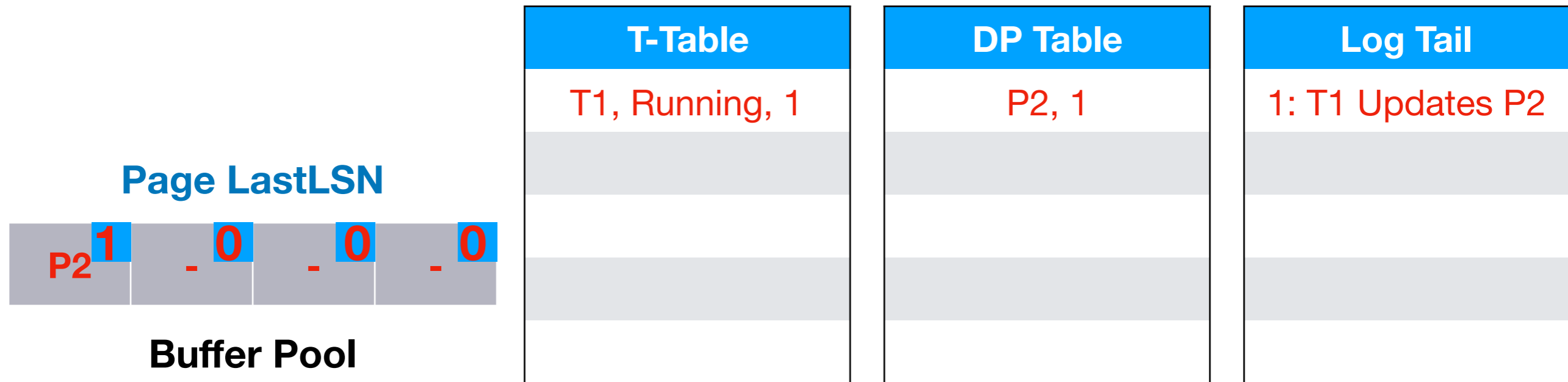


Main Memory (Volatile)

Hard Disk (Non-Volatile)

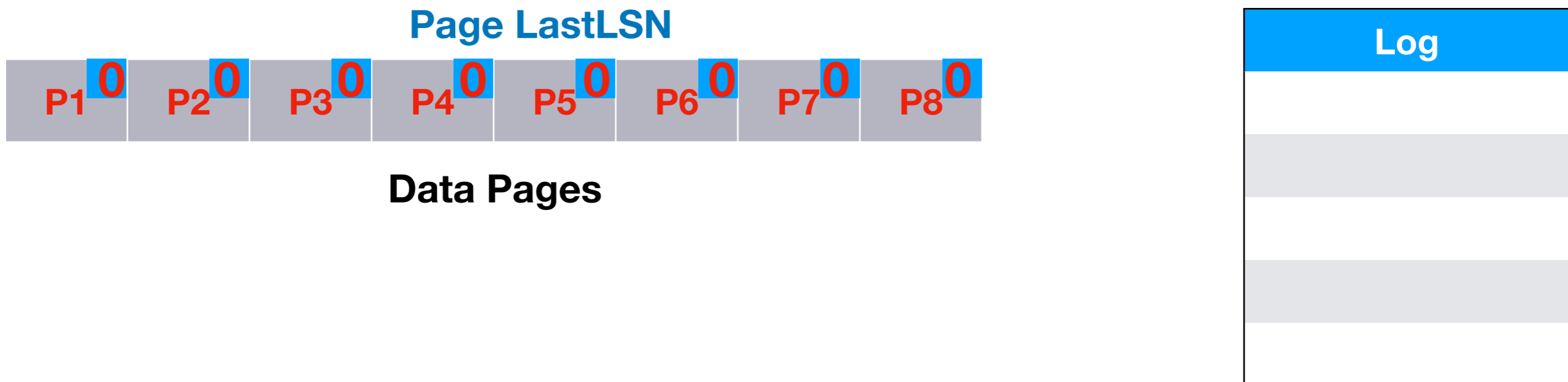


ARIES Example (Run Time)

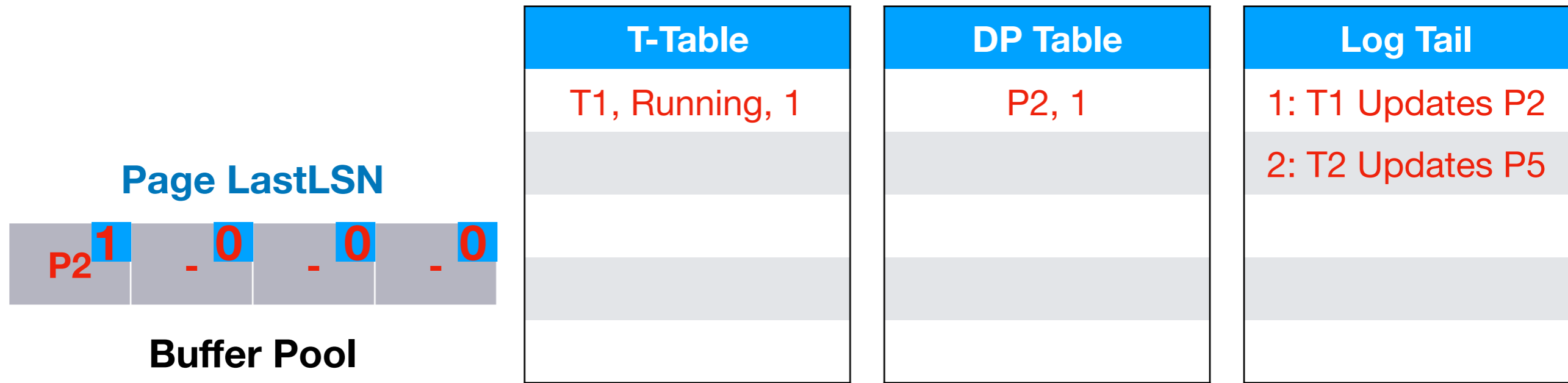


Main Memory (Volatile)

Hard Disk (Non-Volatile)



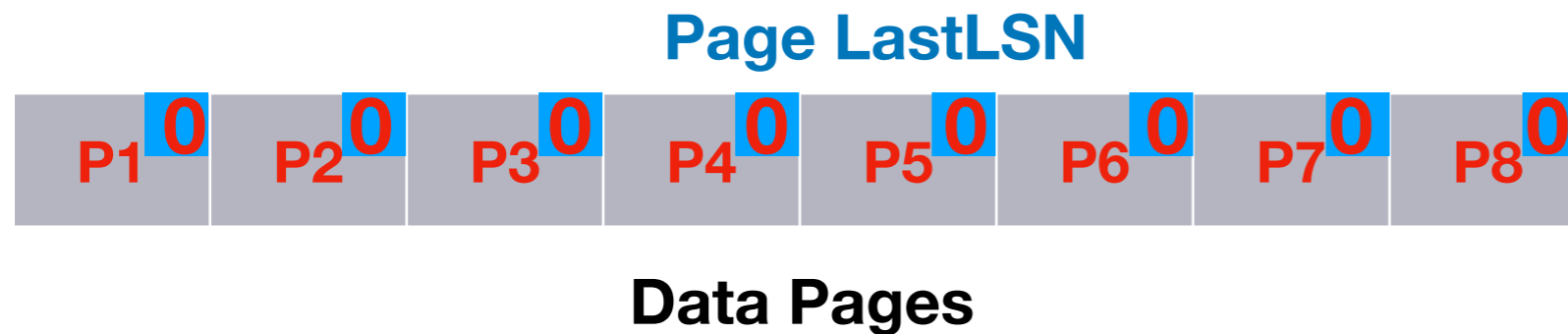
ARIES Example (Run Time)



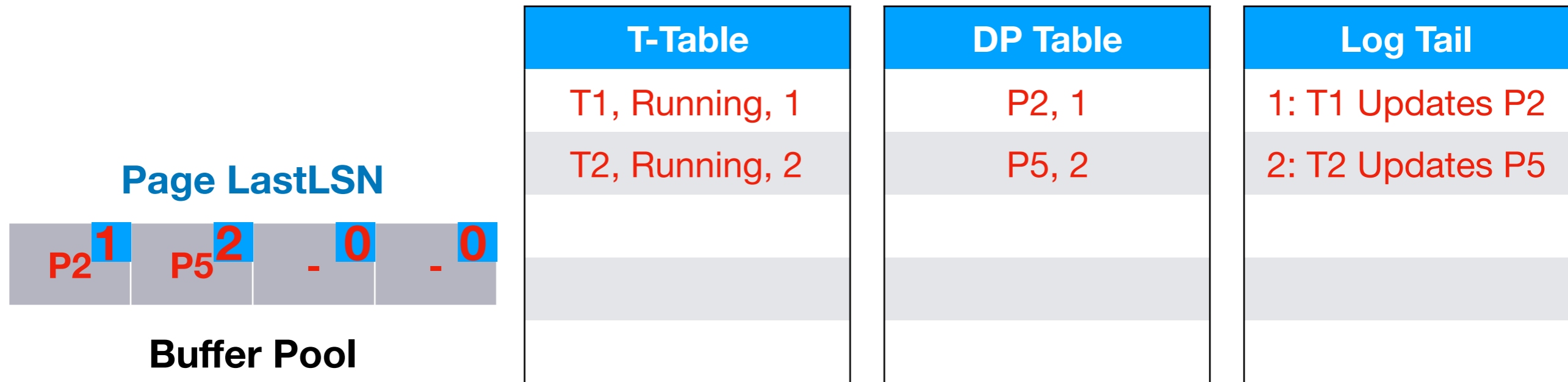
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



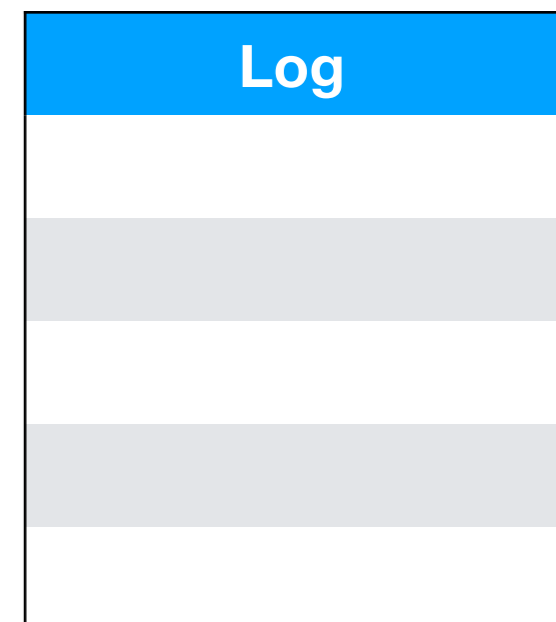
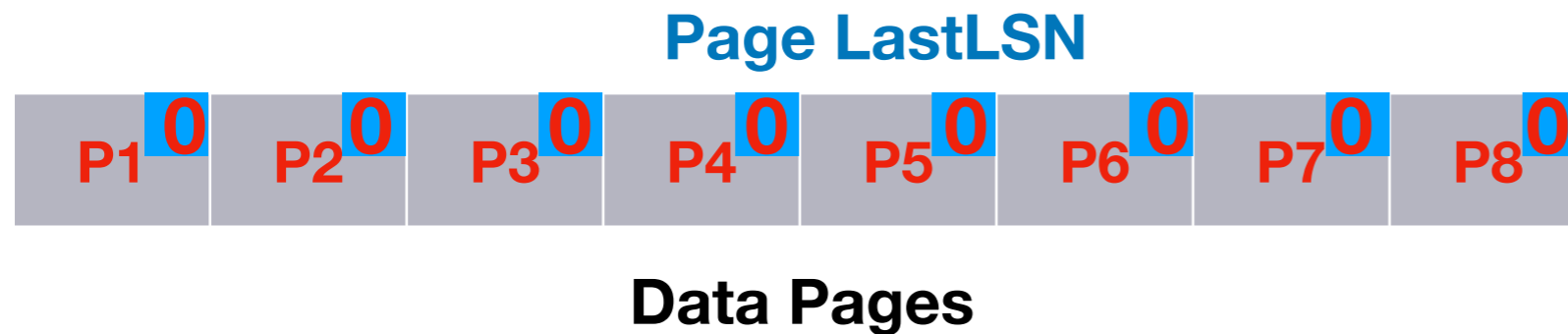
ARIES Example (Run Time)



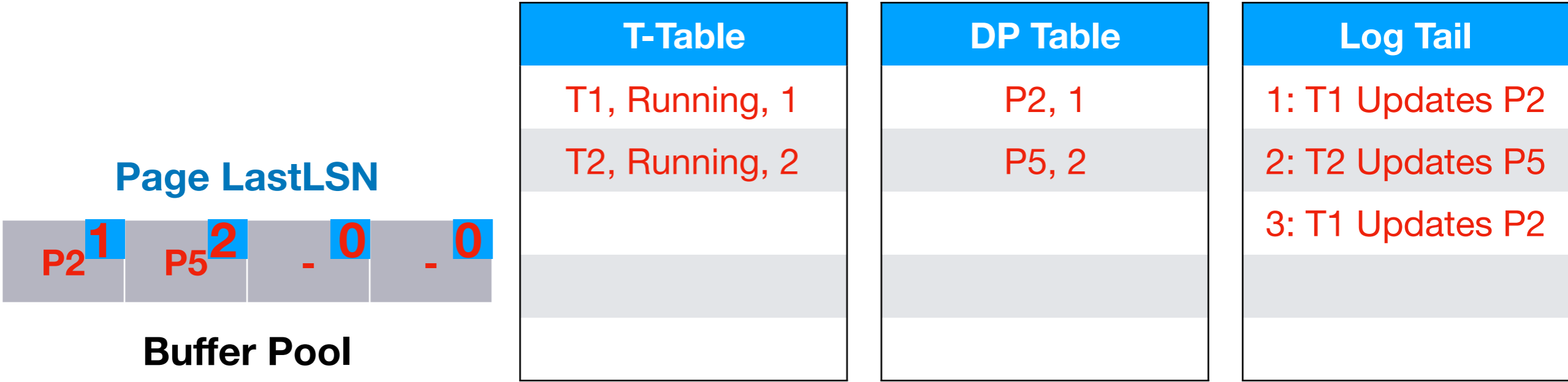
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



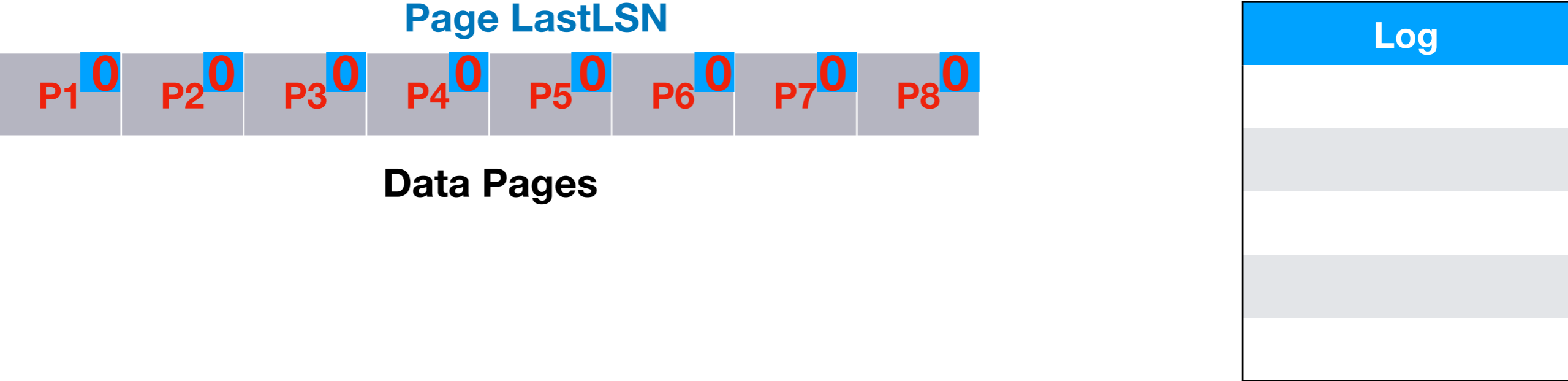
ARIES Example (Run Time)



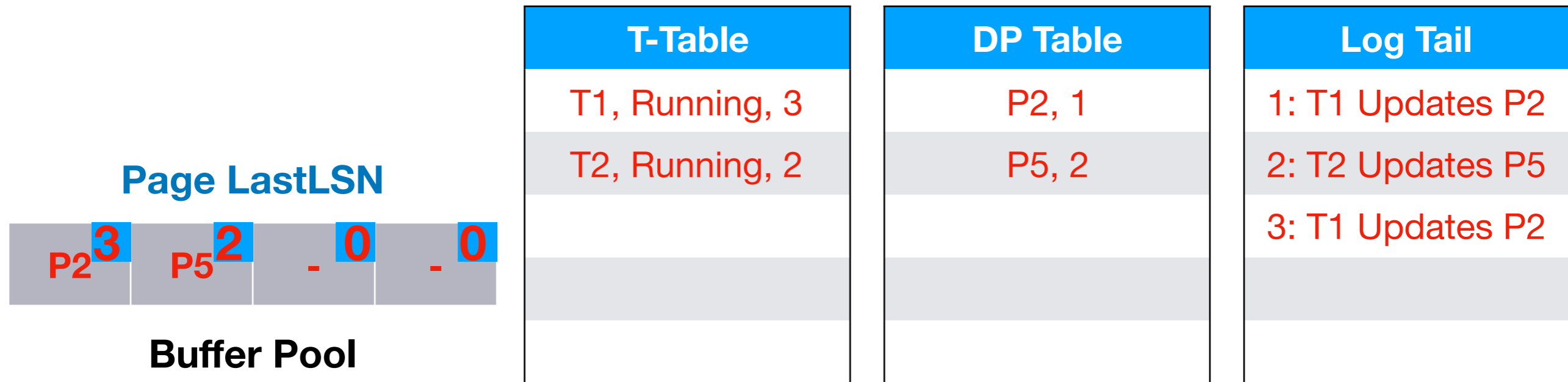
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



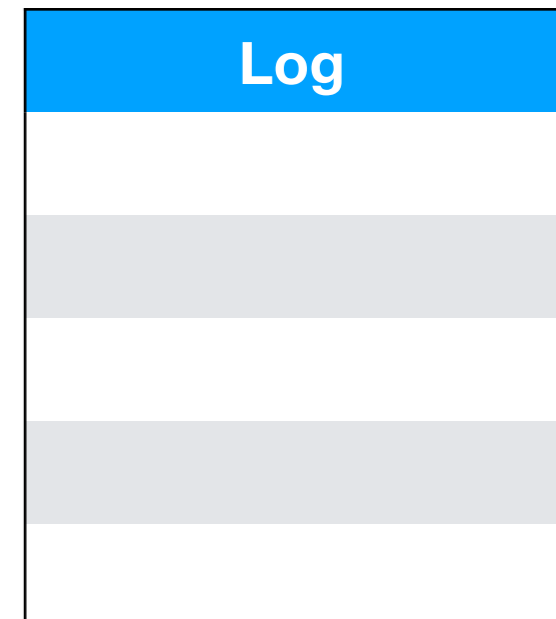
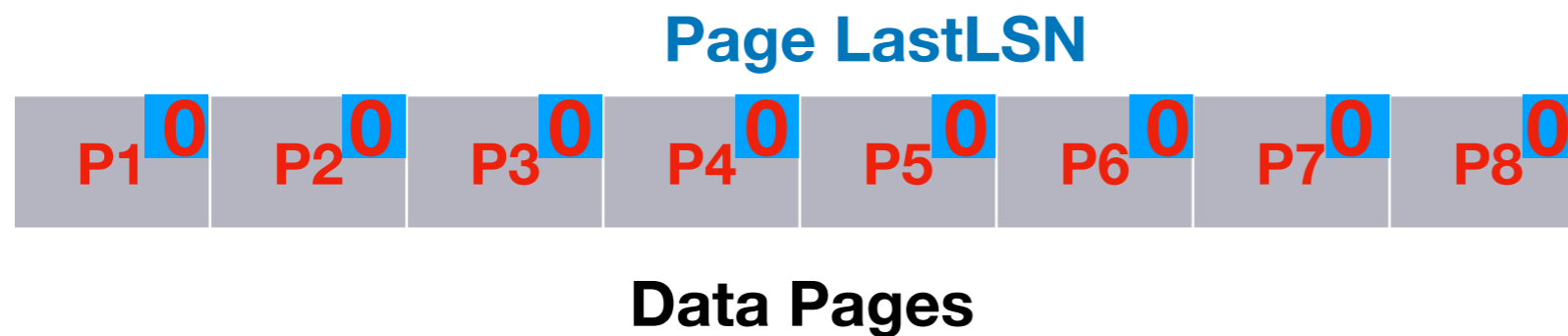
ARIES Example (Run Time)



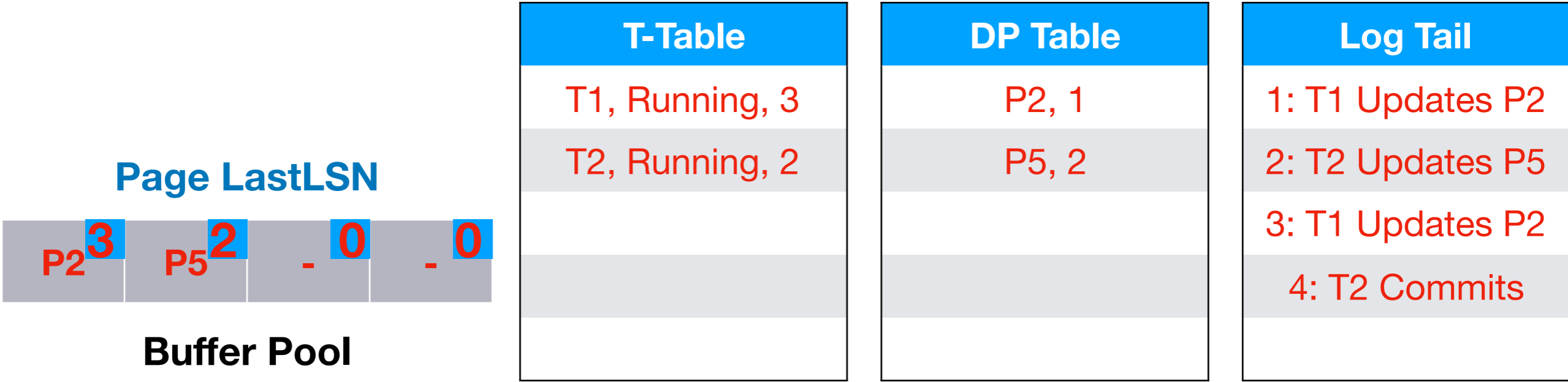
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



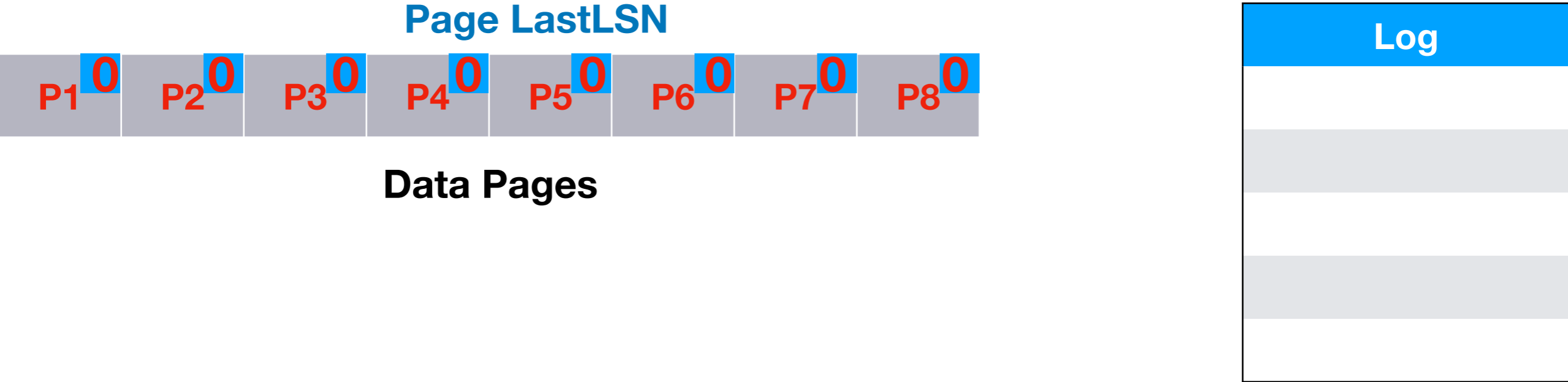
ARIES Example (Run Time)



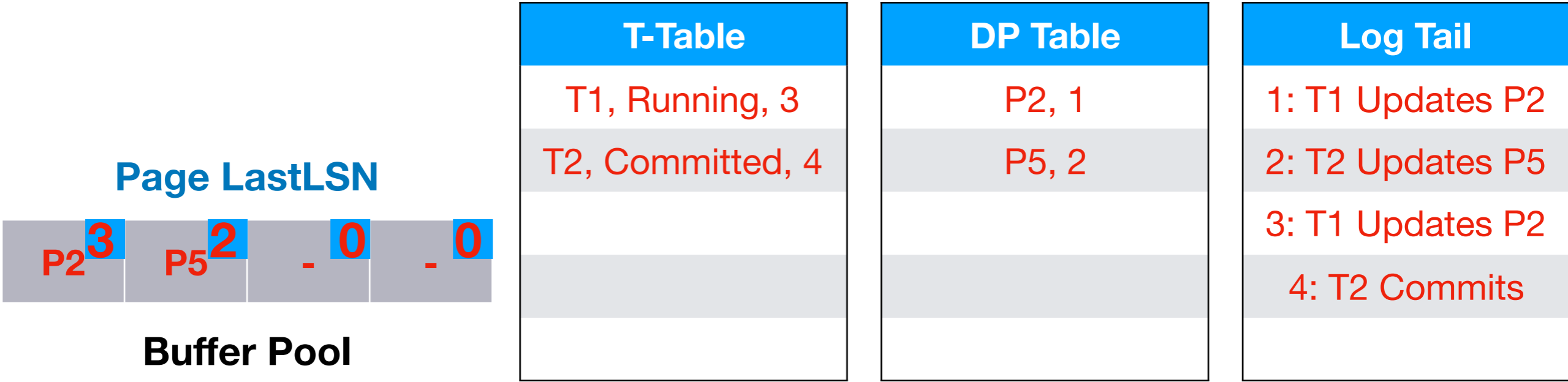
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



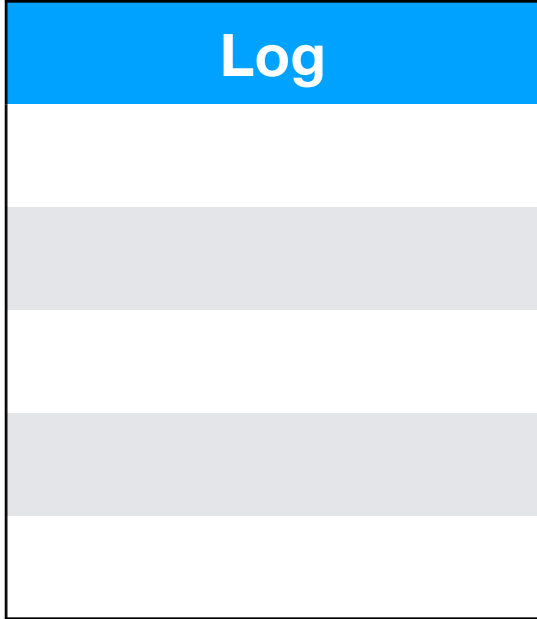
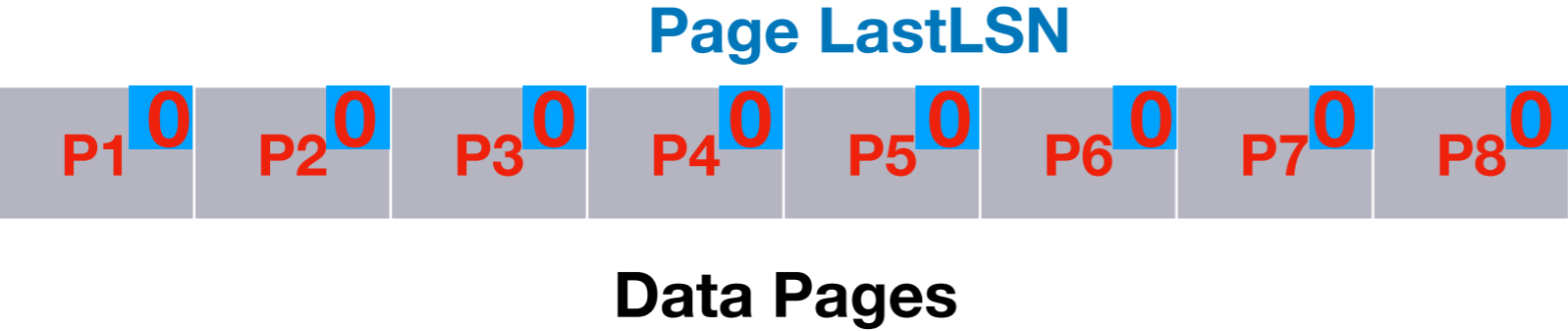
ARIES Example (Run Time)



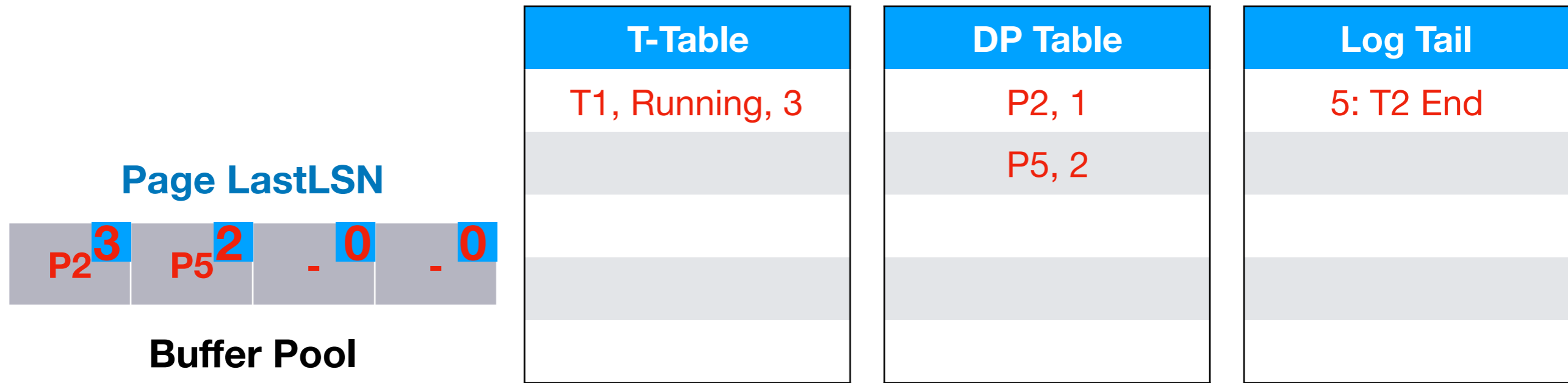
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



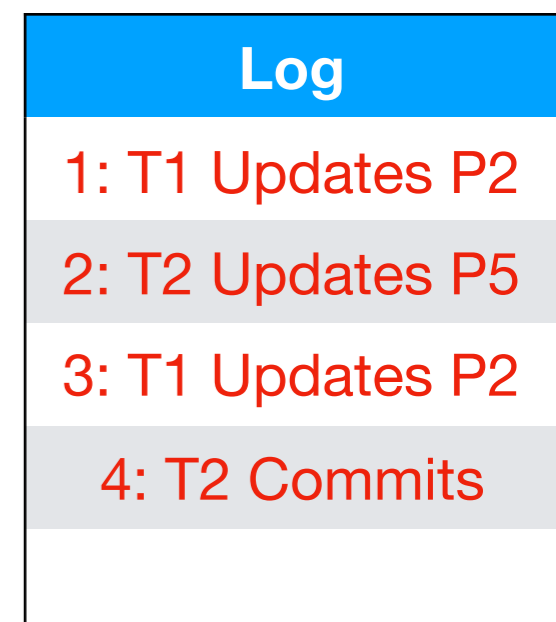
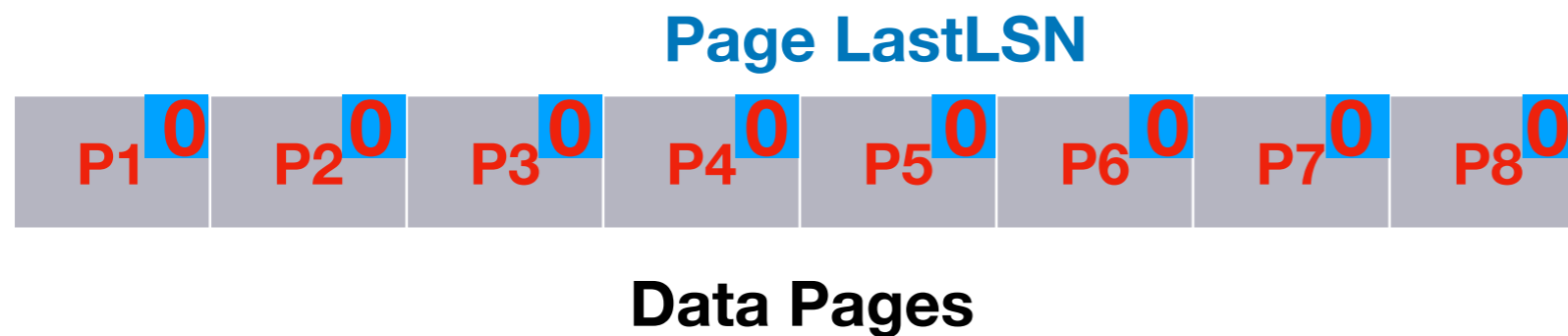
ARIES Example (Run Time)



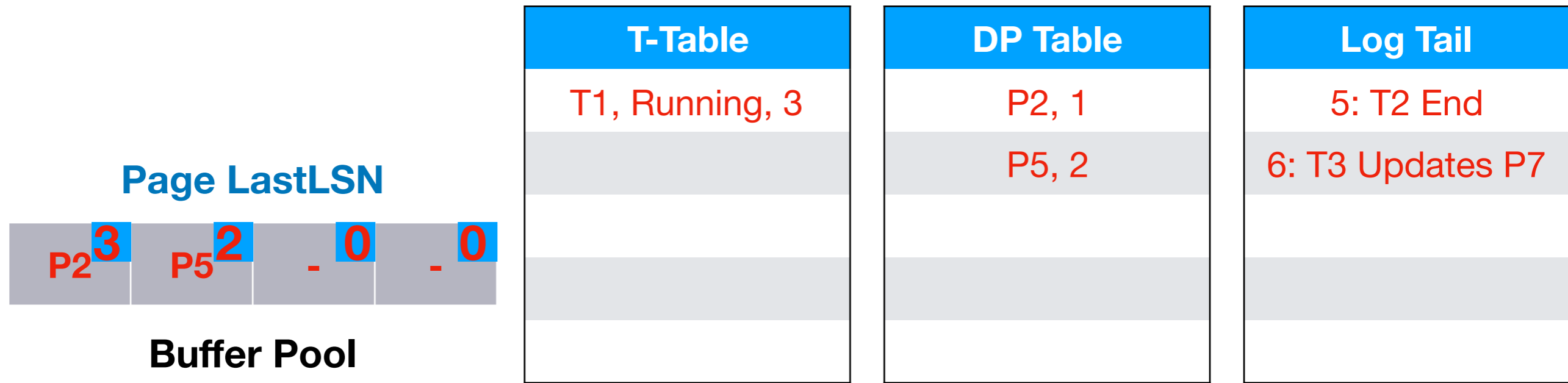
FlushedLSN: 4

Main Memory (Volatile)

Hard Disk (Non-Volatile)



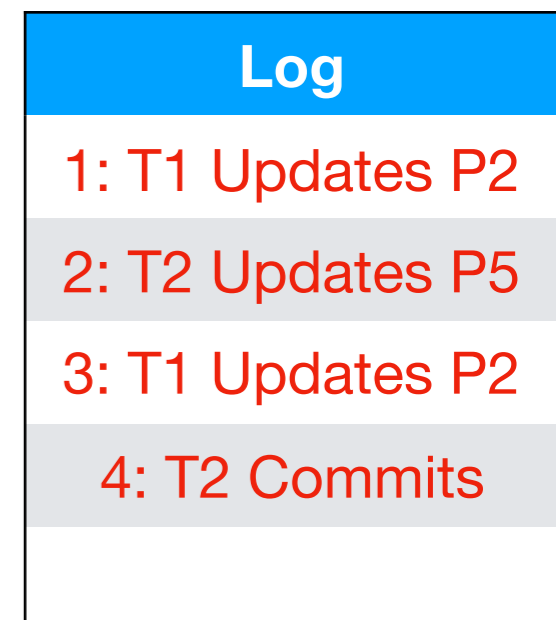
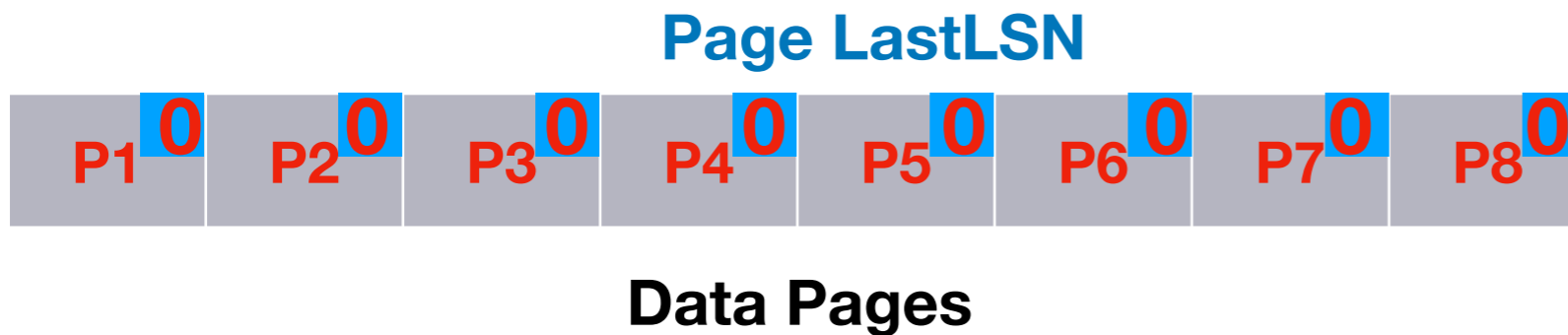
ARIES Example (Run Time)



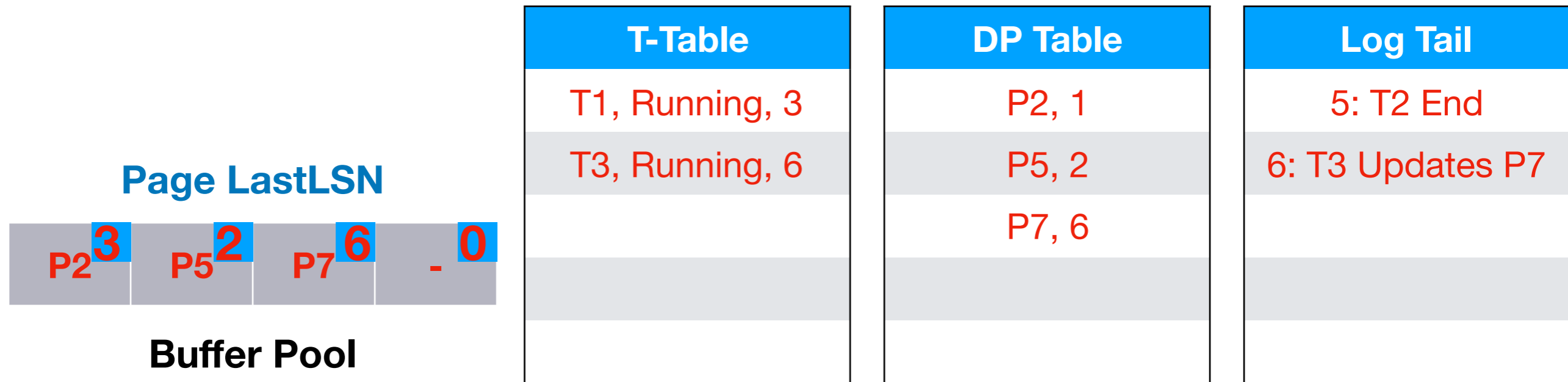
FlushedLSN: 4

Main Memory (Volatile)

Hard Disk (Non-Volatile)



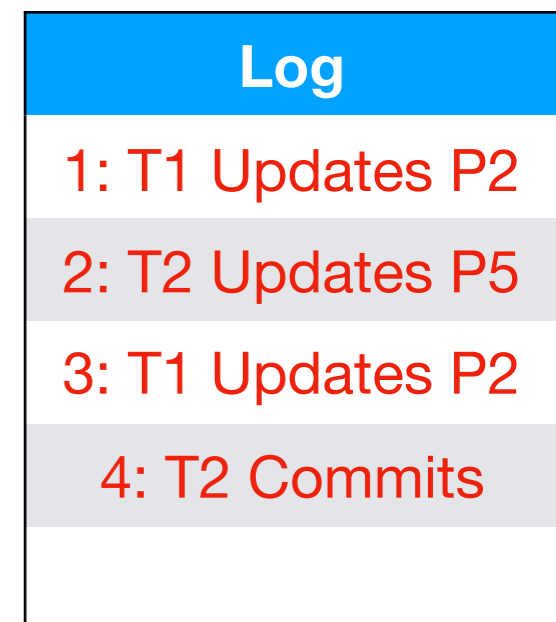
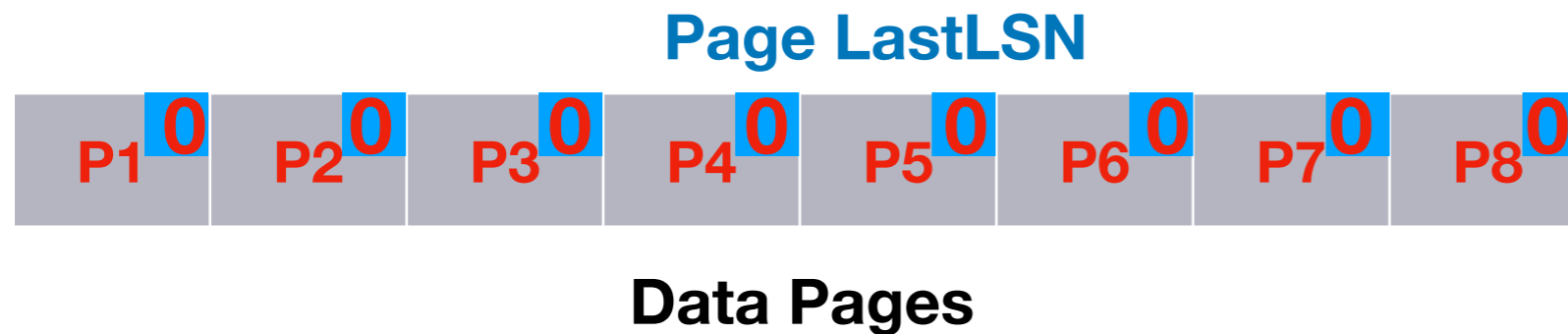
ARIES Example (Run Time)



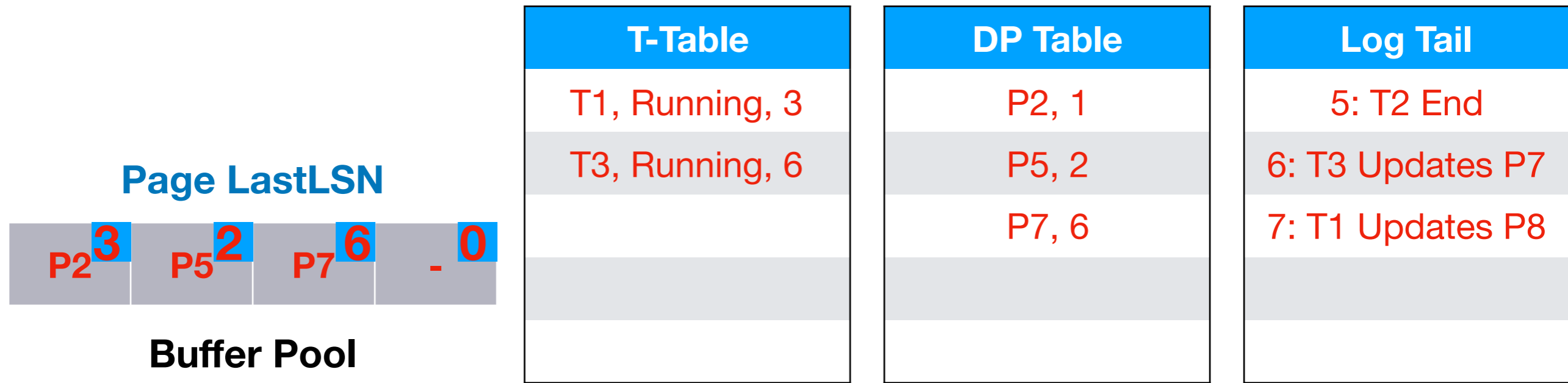
FlushedLSN: 4

Main Memory (Volatile)

Hard Disk (Non-Volatile)



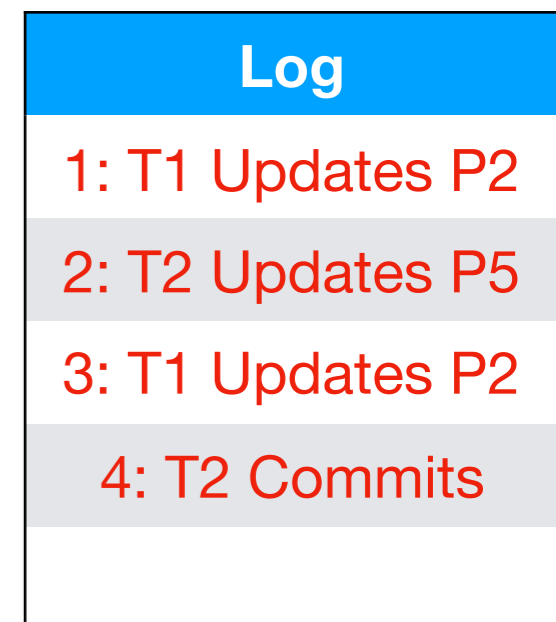
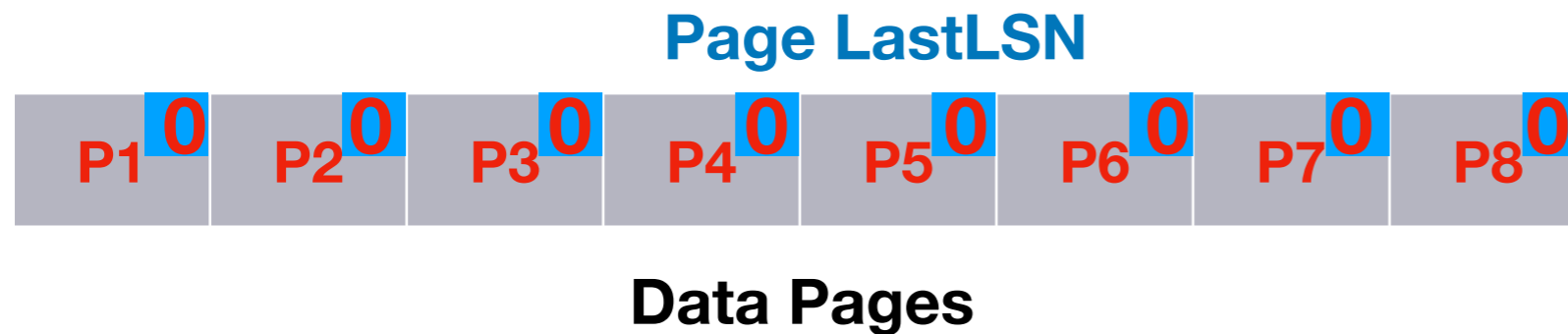
ARIES Example (Run Time)



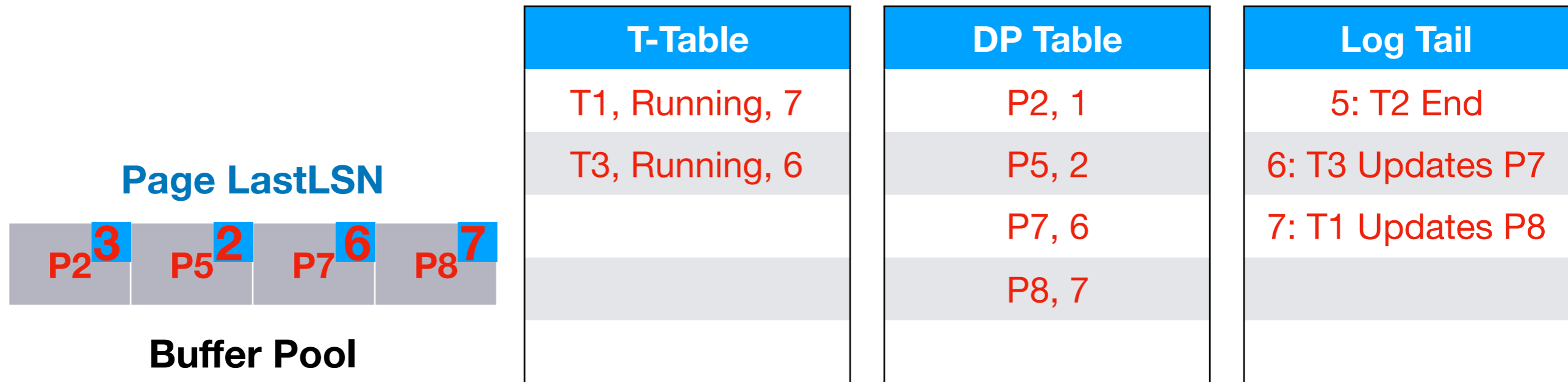
FlushedLSN: 4

Main Memory (Volatile)

Hard Disk (Non-Volatile)



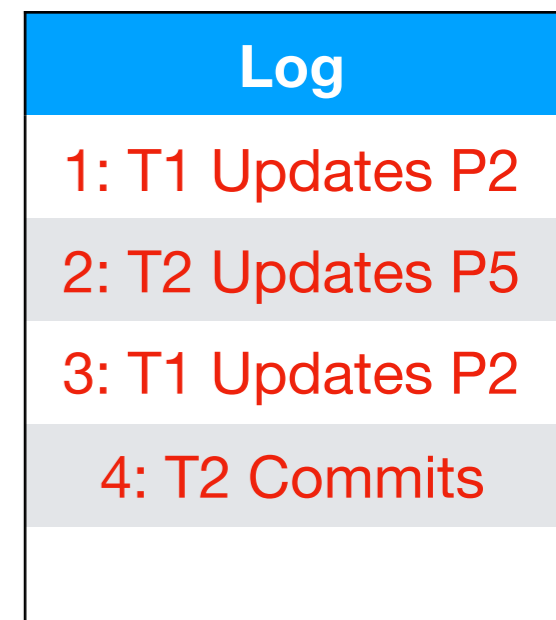
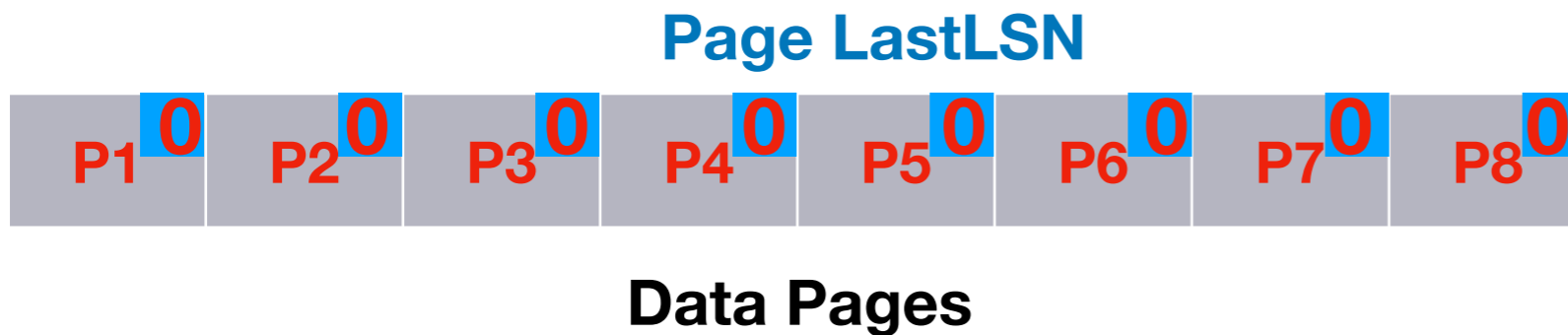
ARIES Example (Run Time)



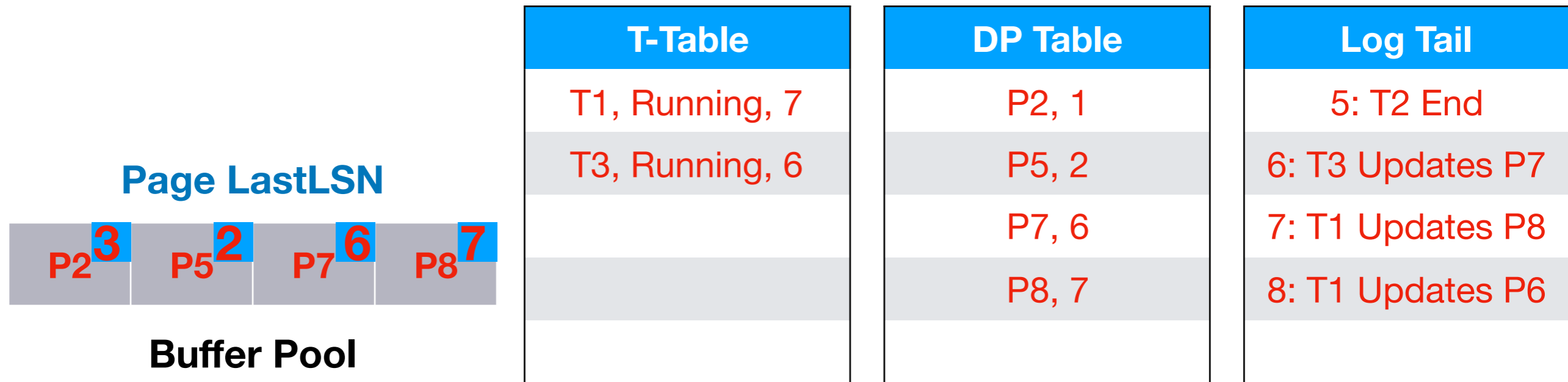
FlushedLSN: 4

Main Memory (Volatile)

Hard Disk (Non-Volatile)



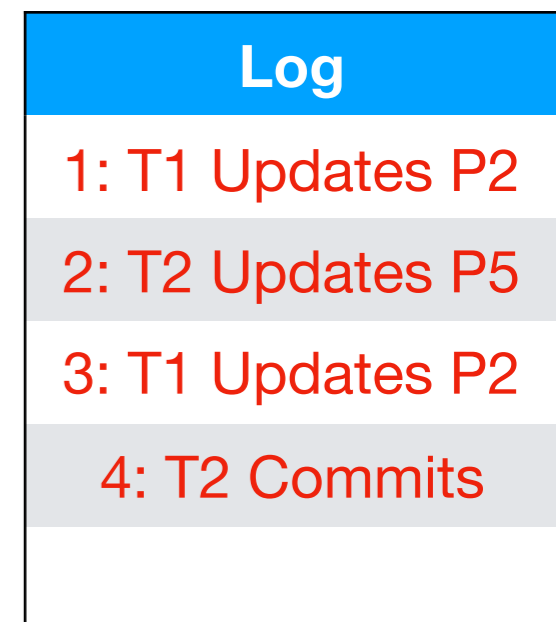
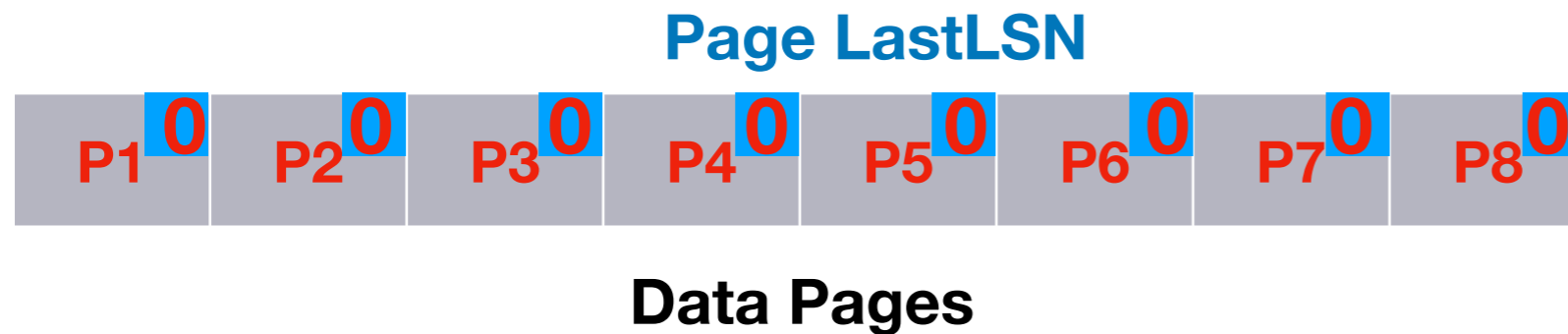
ARIES Example (Run Time)



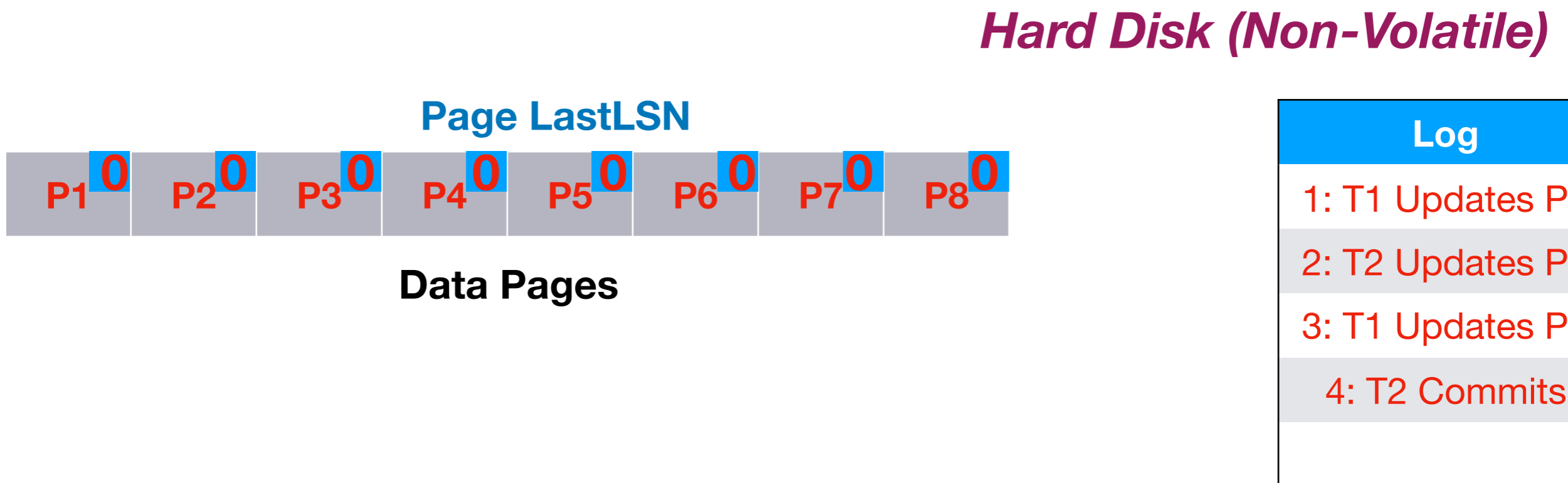
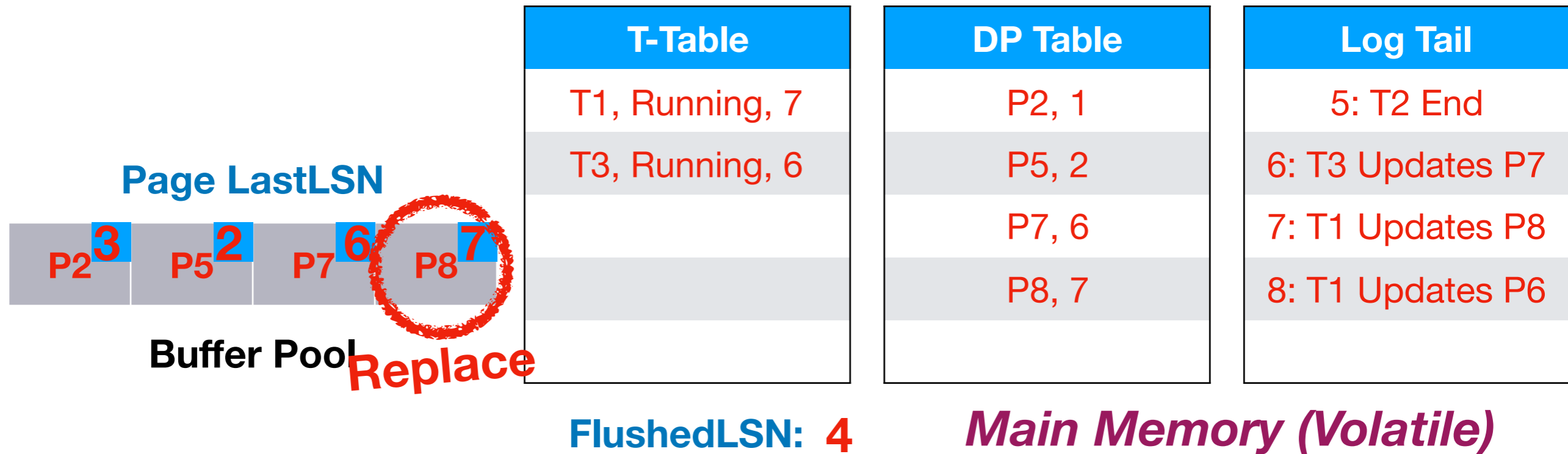
FlushedLSN: 4

Main Memory (Volatile)

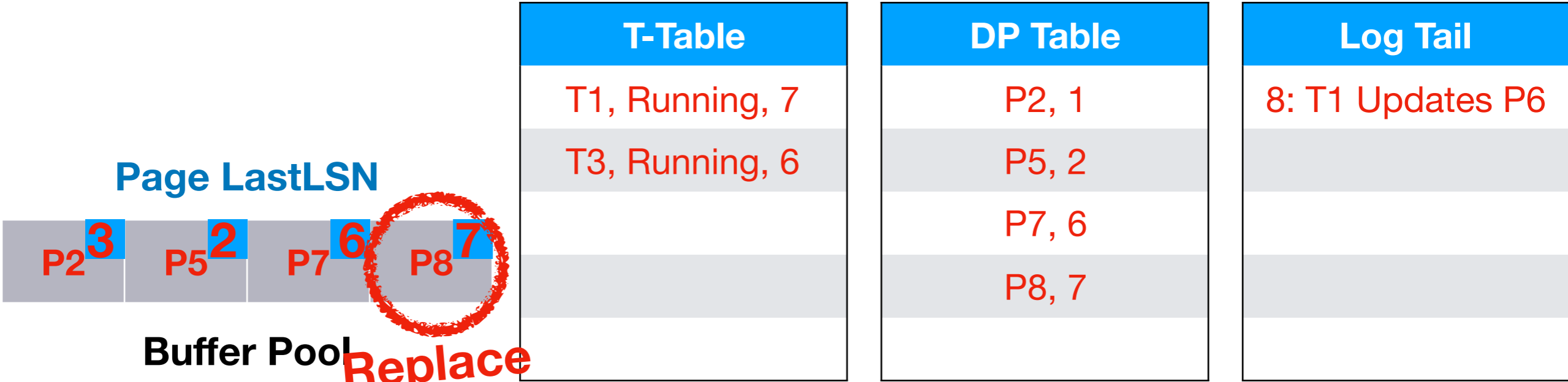
Hard Disk (Non-Volatile)



ARIES Example (Run Time)



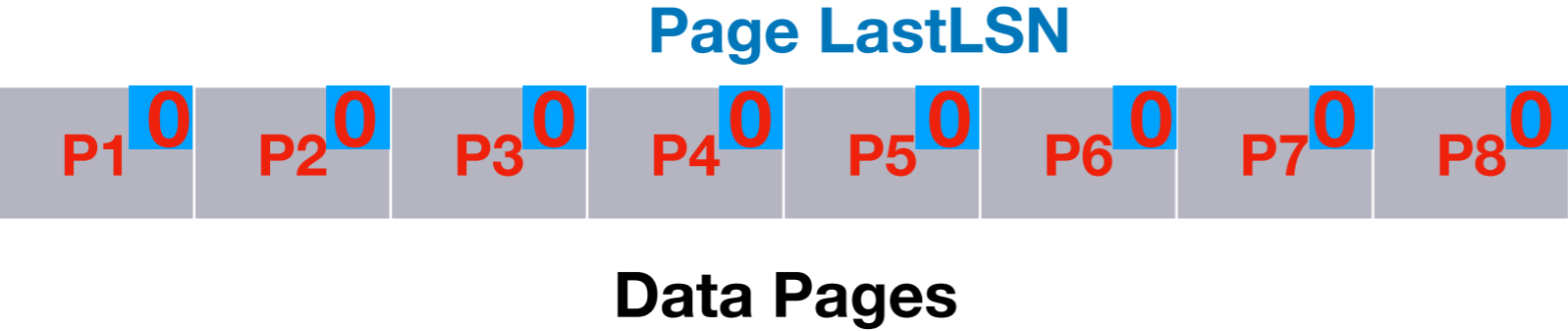
ARIES Example (Run Time)



FlushedLSN: 7

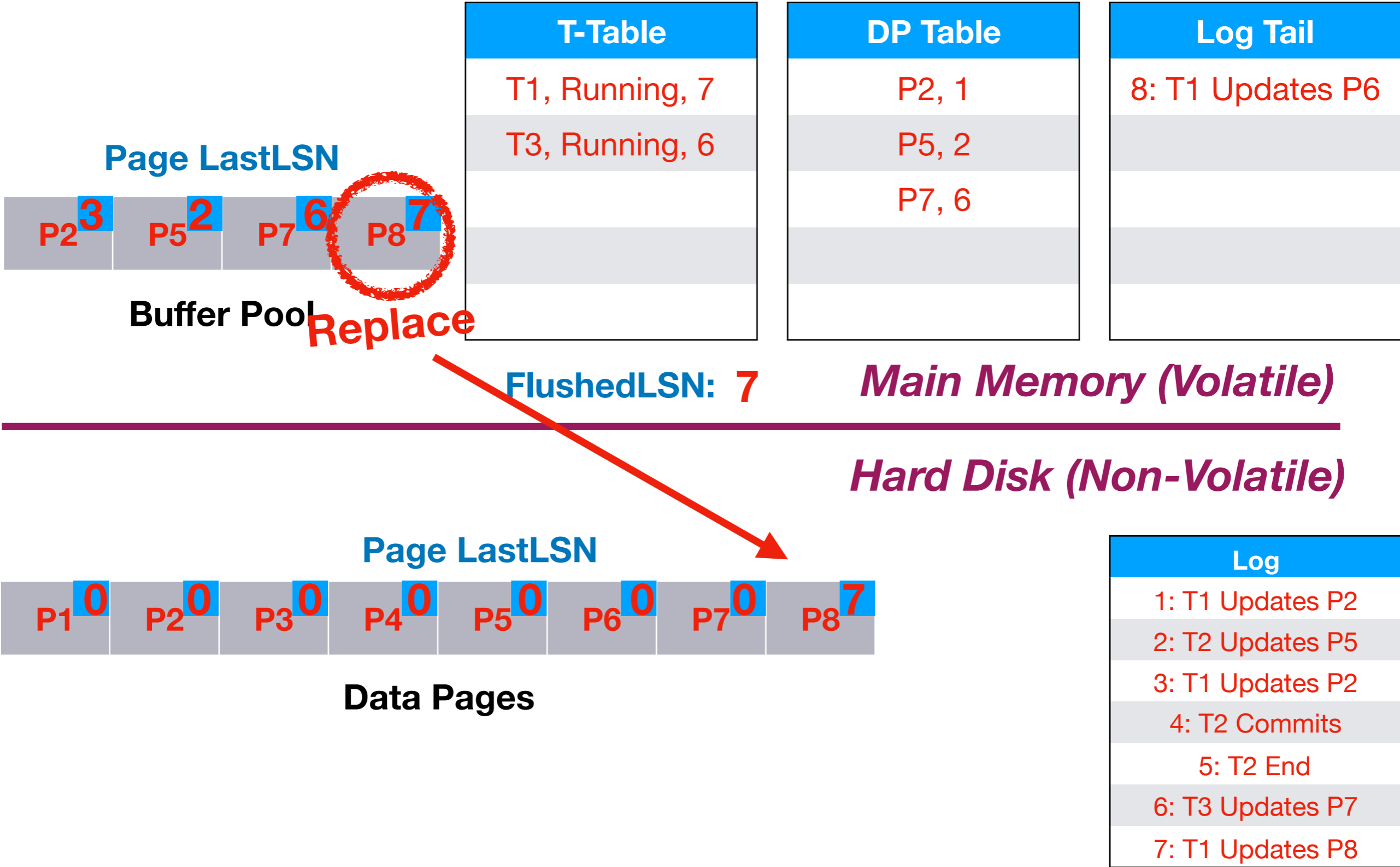
Main Memory (Volatile)

Hard Disk (Non-Volatile)

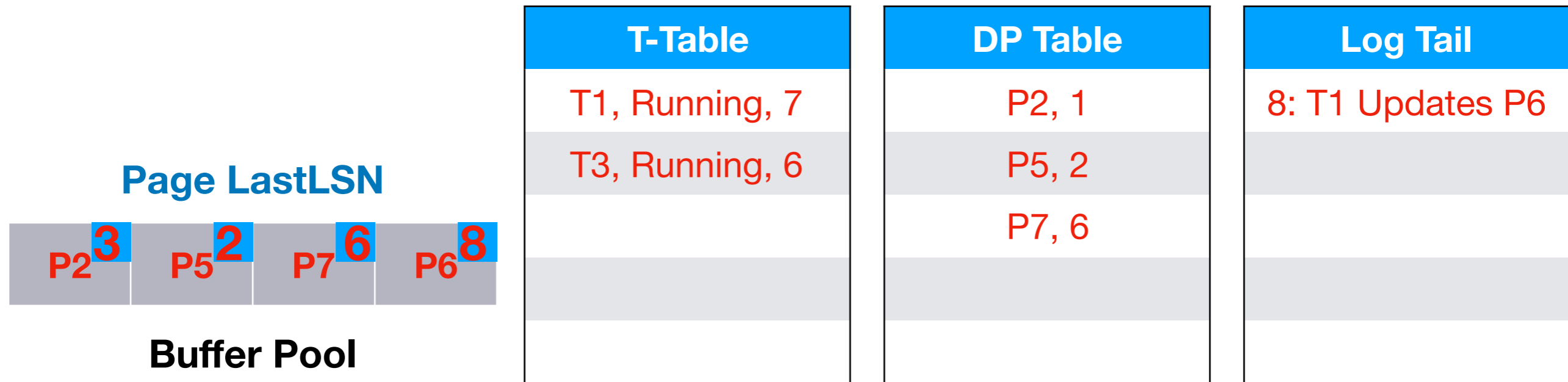


Log
1: T1 Updates P2
2: T2 Updates P5
3: T1 Updates P2
4: T2 Commits
5: T2 End
6: T3 Updates P7
7: T1 Updates P8

ARIES Example (Run Time)



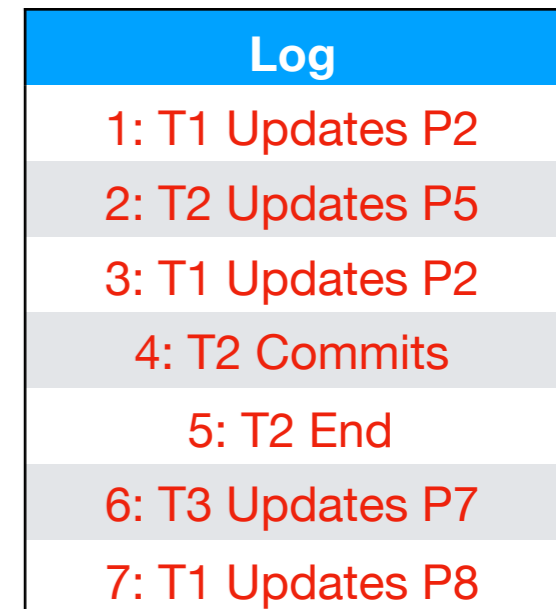
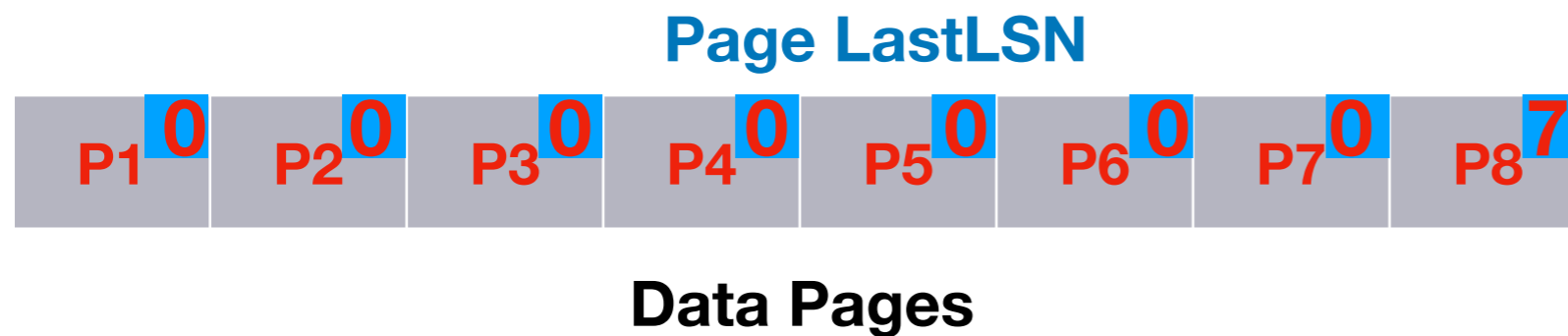
ARIES Example (Run Time)



FlushedLSN: 7

Main Memory (Volatile)

Hard Disk (Non-Volatile)



Outlook

- ARIES data structures
- ARIES run time behavior
- **ARIES recovery algorithm**

Recovery Phase Overview

- **Analysis phase**
 - Read log to restore transaction & dirty page tables
- **Redo phase**
 - Redo non-persisted changes of all transactions
- **Undo phase**
 - Undo changes of aborted transactions

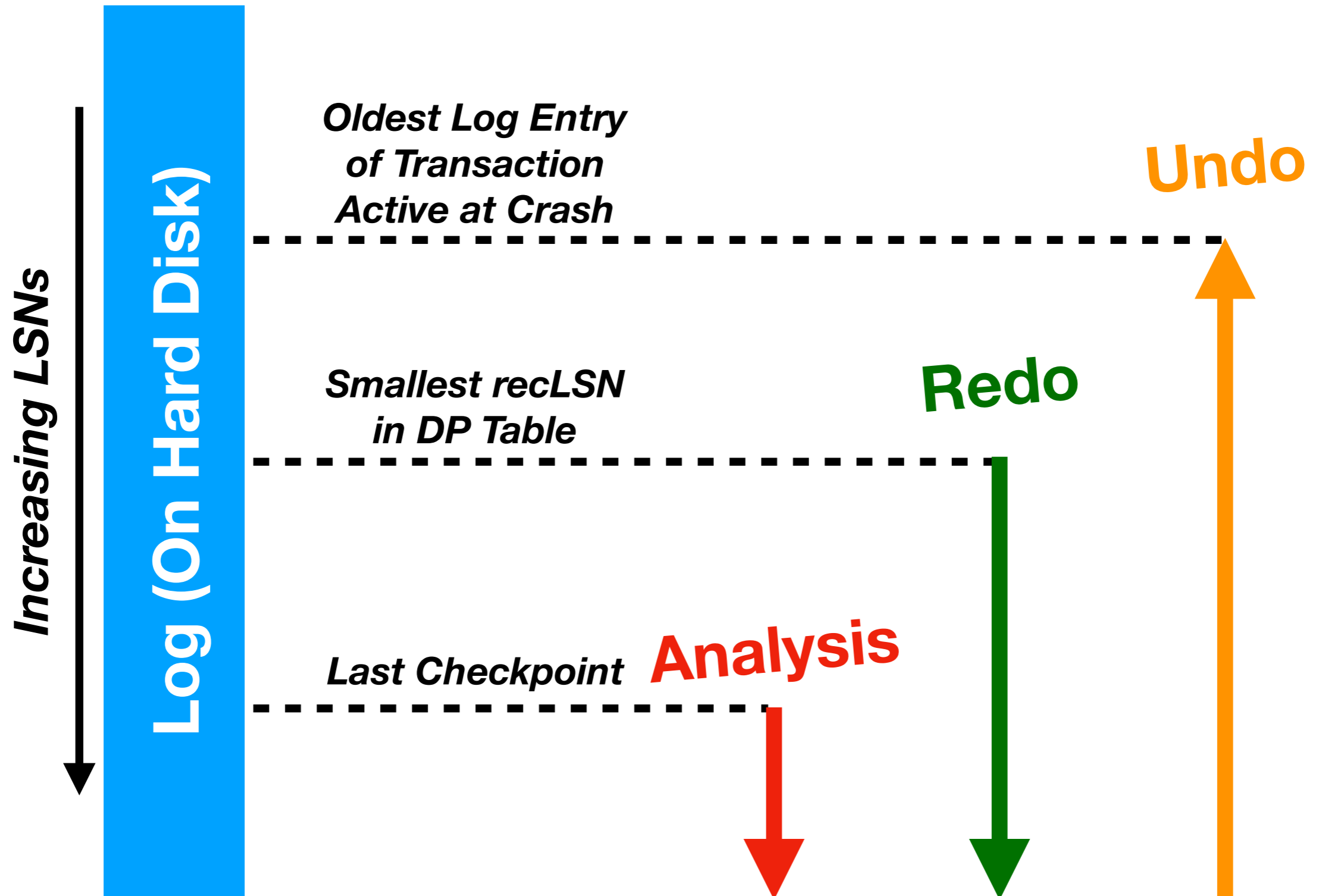
Analysis Refinement: Checkpoints

- Restore **transaction & dirty page** table during analysis
 - Without refinements, would have to read **entire log**
- **Checkpoints save time** during the analysis phase
 - Checkpoint captures **transaction & DP table** state
 - Checkpoint written to **log**
 - **Master** stores LSN of last checkpoint

Writing Checkpoints

- Want to write checkpoints **without stopping** transactions
- Log "**Begin_Checkpoint**" at checkpoint start
- Now start writing **Transaction & DP table** to log
- Log "**End_Checkpoint**" once this is done
- Logged tables represent state **at checkpoint start**

Log Scans During Recovery



Analysis Phase I

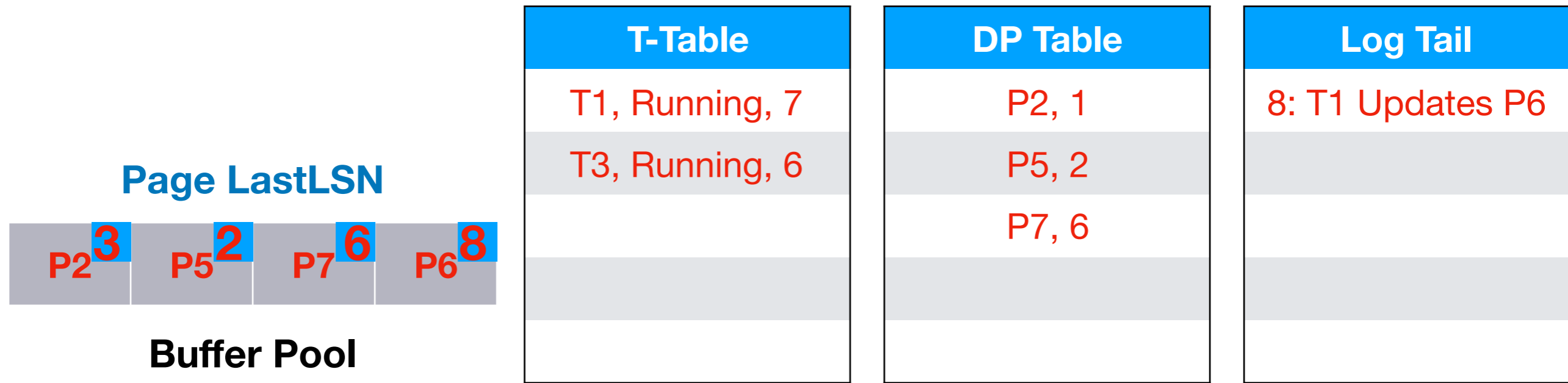
- Initialize **transaction & DP table** to last checkpoint
- Scan log forward starting from **Begin_Checkpoint**
- LSN **Lx**: Transaction **Ty** updates page **Pz**
 - If Pz not in **DP table**, add with recLSN = Lx
 - If Ty not in **transaction table**, add as running
 - Set lastLSN = Lx in **transaction table**

Analysis Phase II

- LSN Lx: transaction Ty **commits**
 - Mark Ty as **committed** in transaction table
- LSN Lx: transaction Ty **aborts**
 - Mark Ty as **aborted** in transaction table
- LSN Lx: **end record** for transaction Ty
 - **Remove** Ty from transaction table

*Do We Get Precise
Transaction Table & DP
Table State Before Crash?*

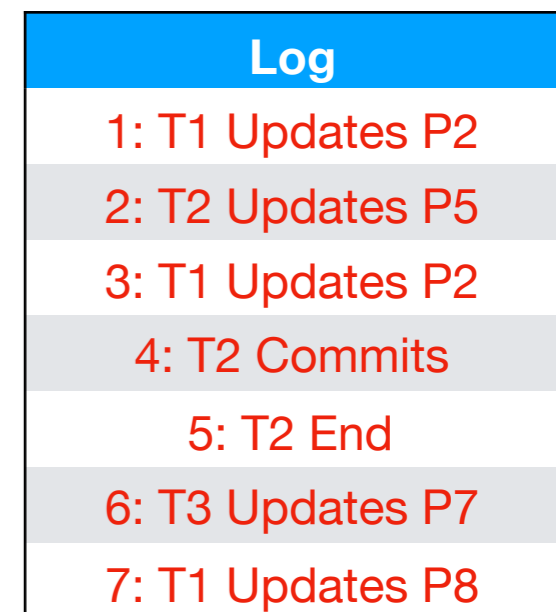
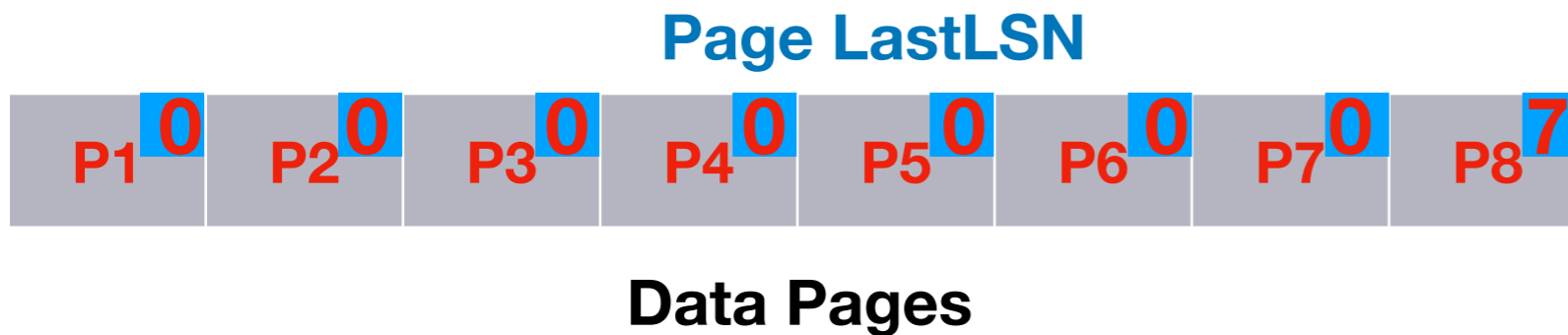
ARIES Example (Analysis)



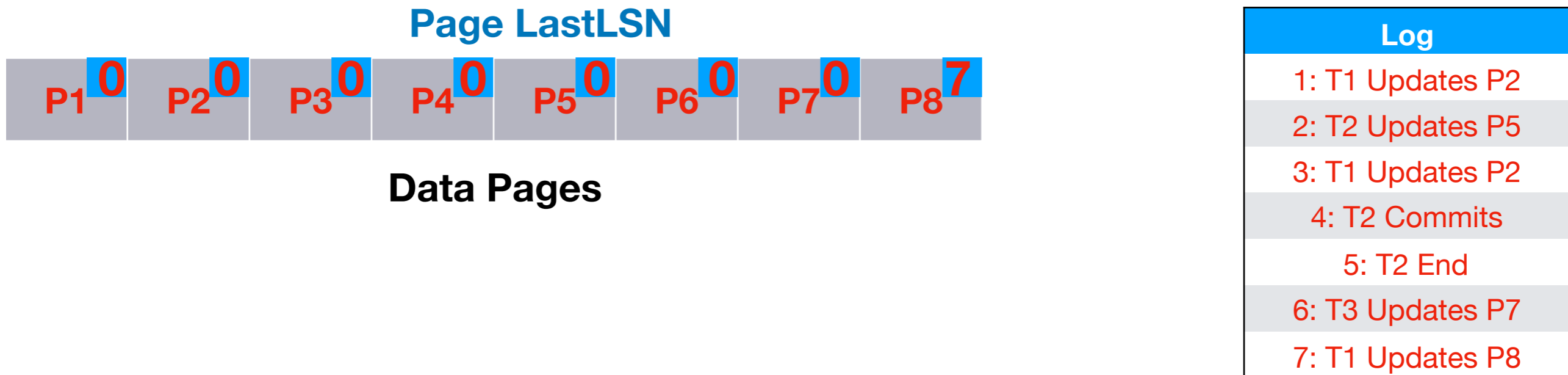
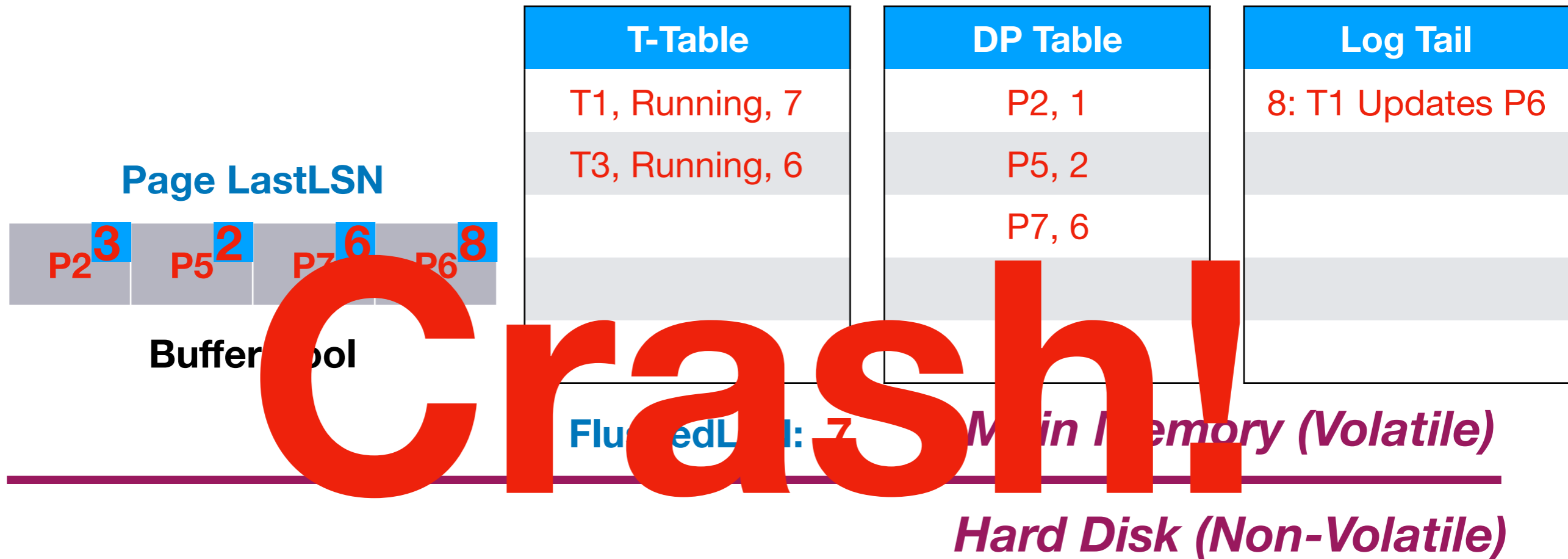
FlushedLSN: 7

Main Memory (Volatile)

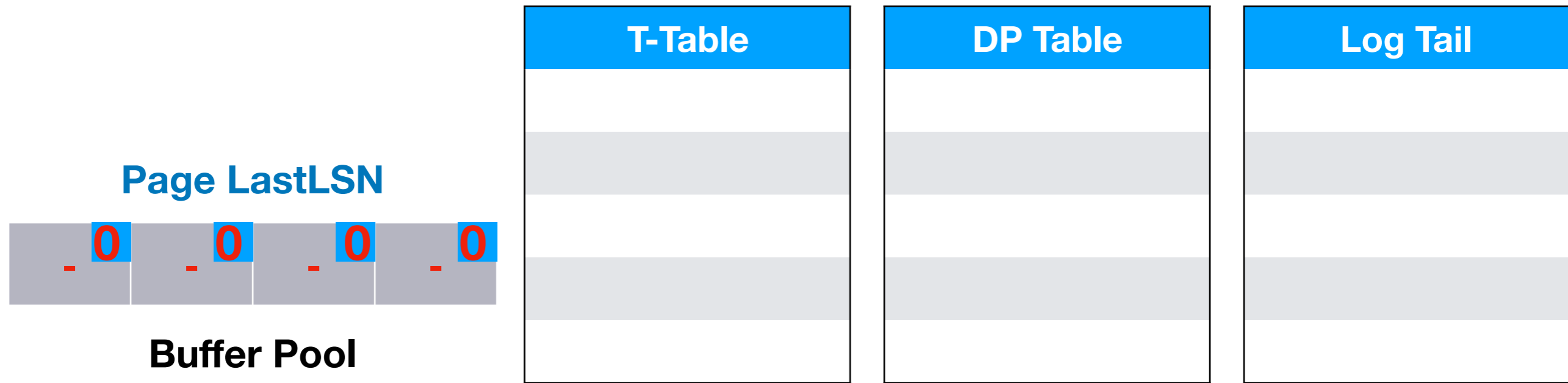
Hard Disk (Non-Volatile)



ARIES Example (Analysis)

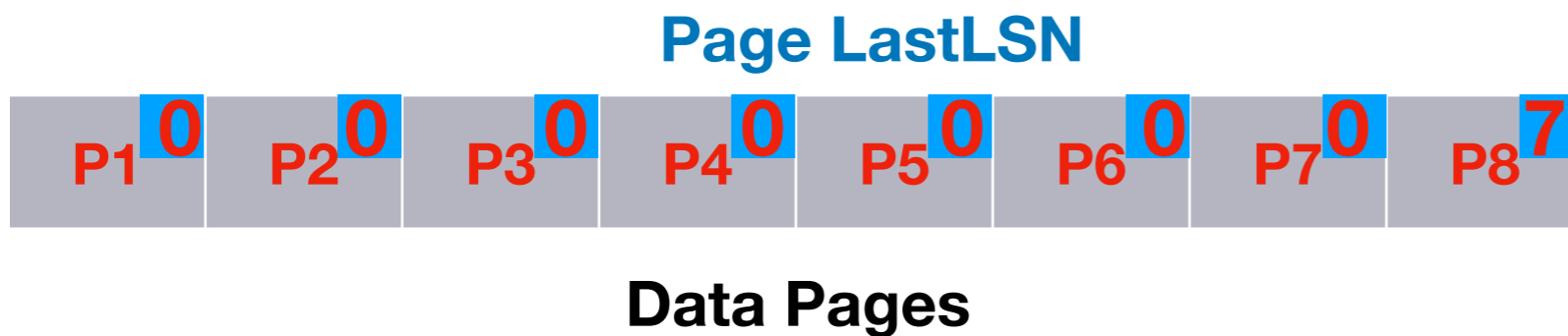


ARIES Example (Analysis)



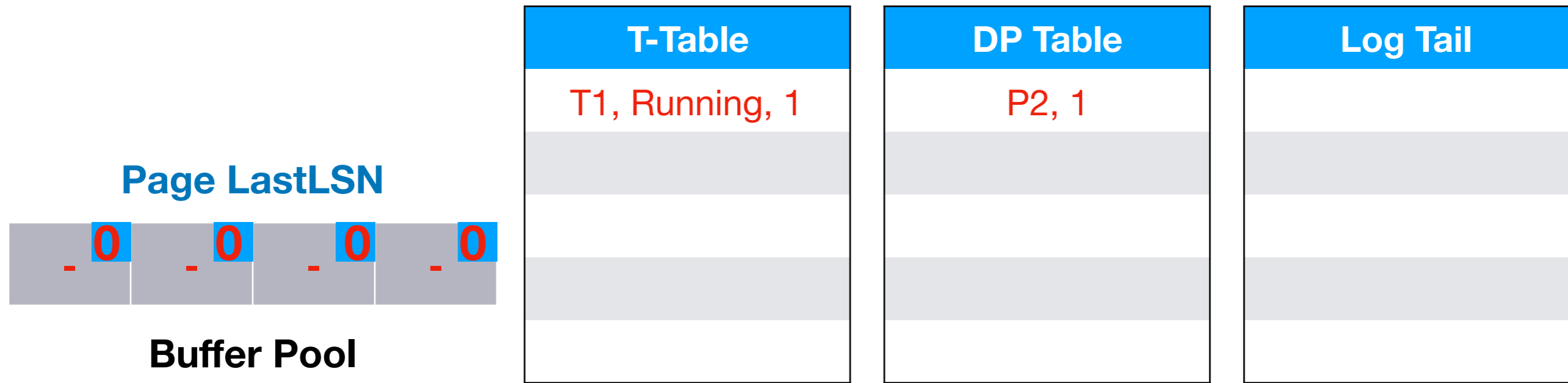
Main Memory (Volatile)

Hard Disk (Non-Volatile)



Log
1: T1 Updates P2
2: T2 Updates P5
3: T1 Updates P2
4: T2 Commits
5: T2 End
6: T3 Updates P7
7: T1 Updates P8

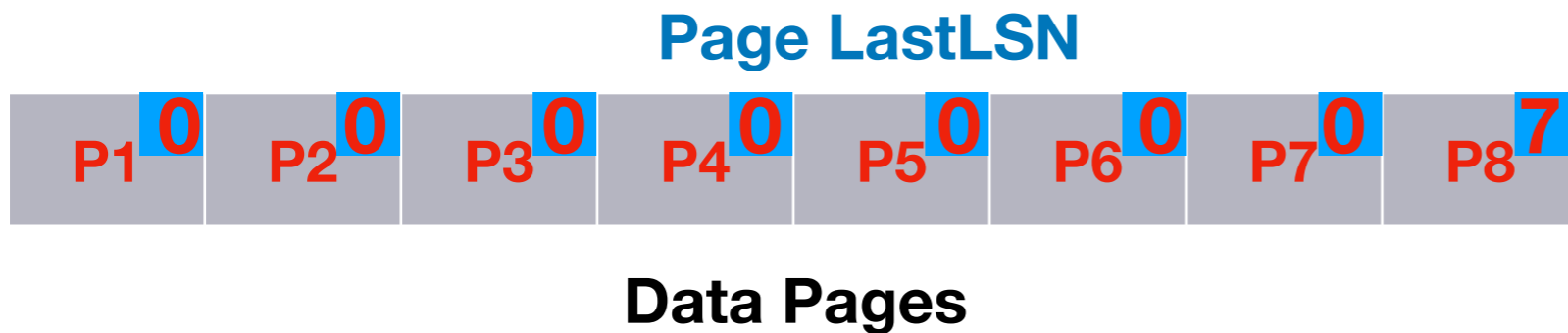
ARIES Example (Analysis)



FlushedLSN: 0

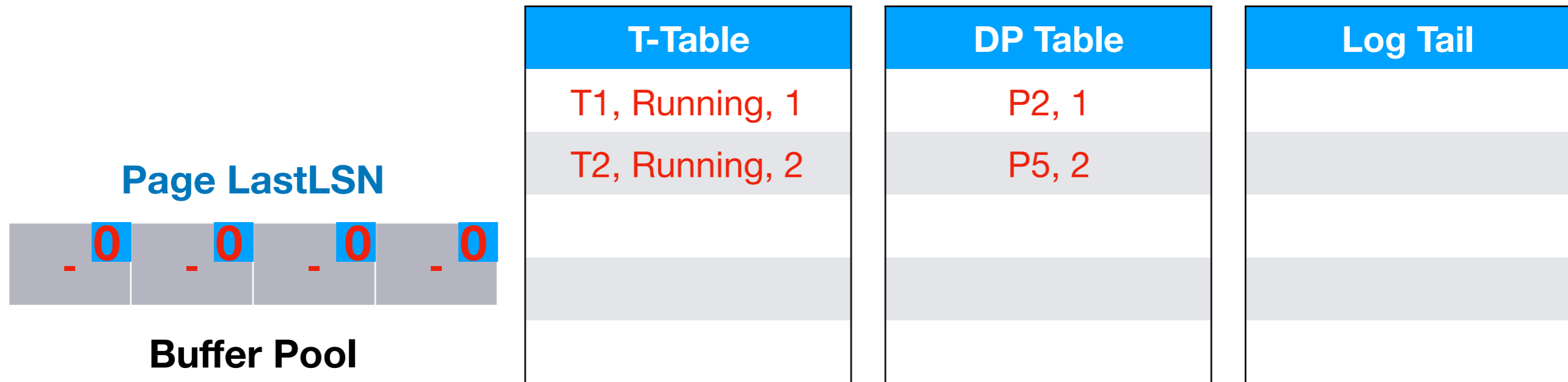
Main Memory (Volatile)

Hard Disk (Non-Volatile)



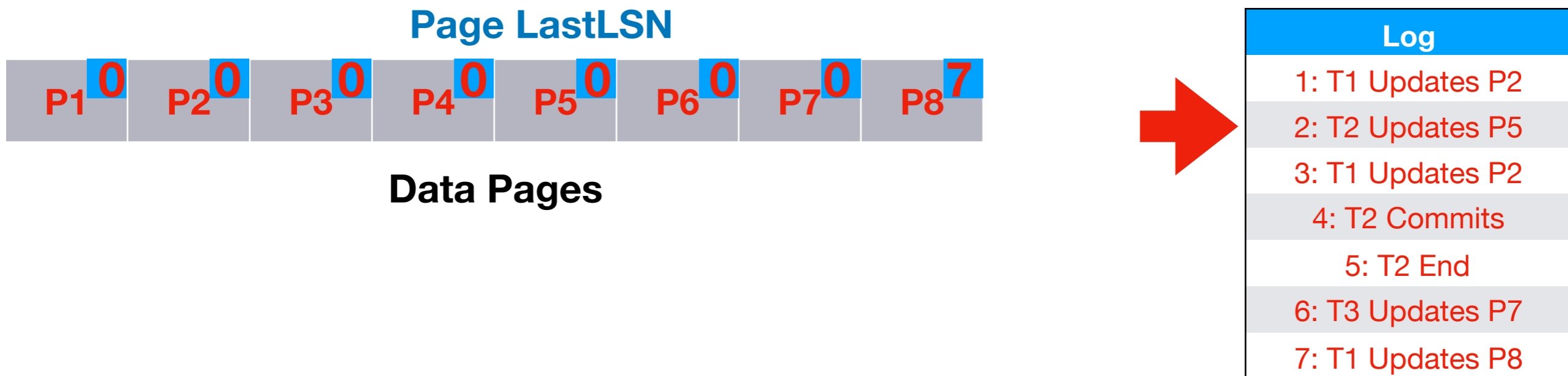
Log
1: T1 Updates P2
2: T2 Updates P5
3: T1 Updates P2
4: T2 Commits
5: T2 End
6: T3 Updates P7
7: T1 Updates P8

ARIES Example (Analysis)

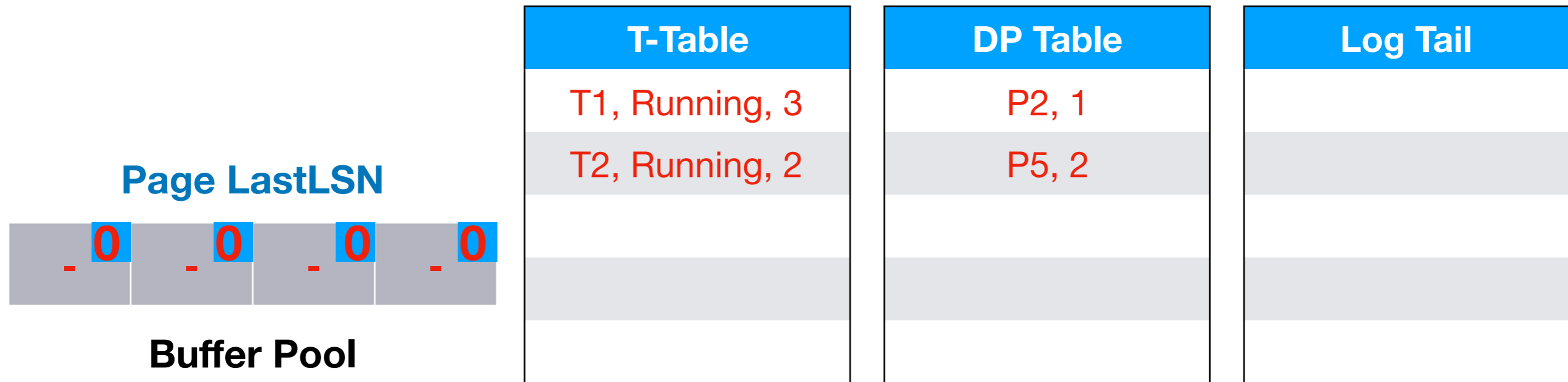


Main Memory (Volatile)

Hard Disk (Non-Volatile)

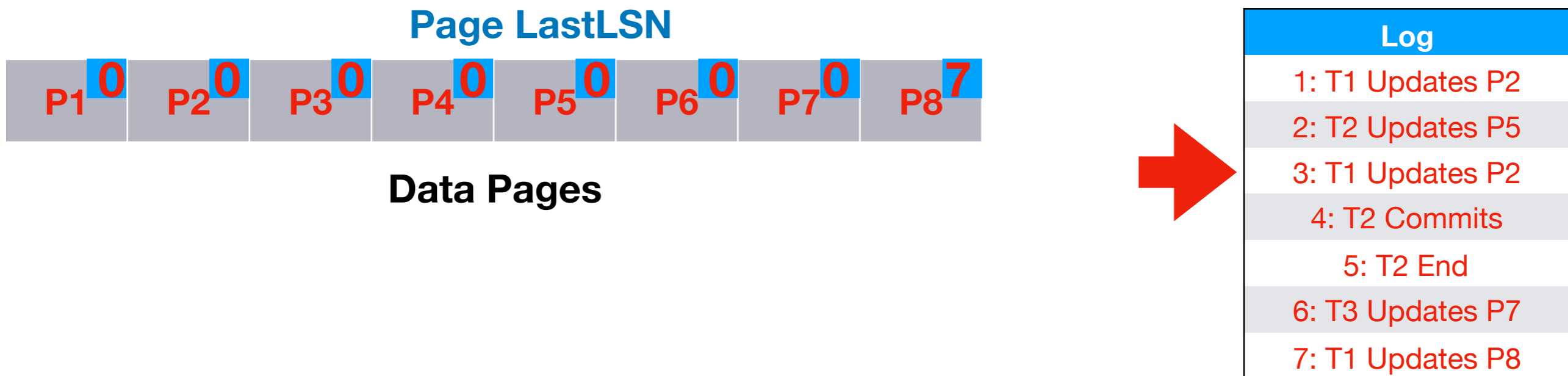


ARIES Example (Analysis)

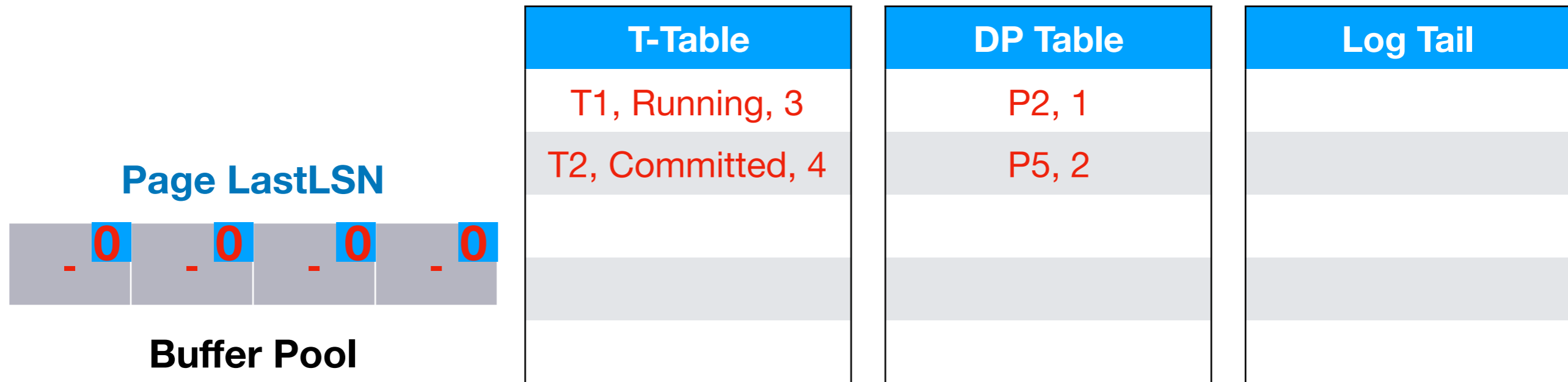


Main Memory (Volatile)

Hard Disk (Non-Volatile)



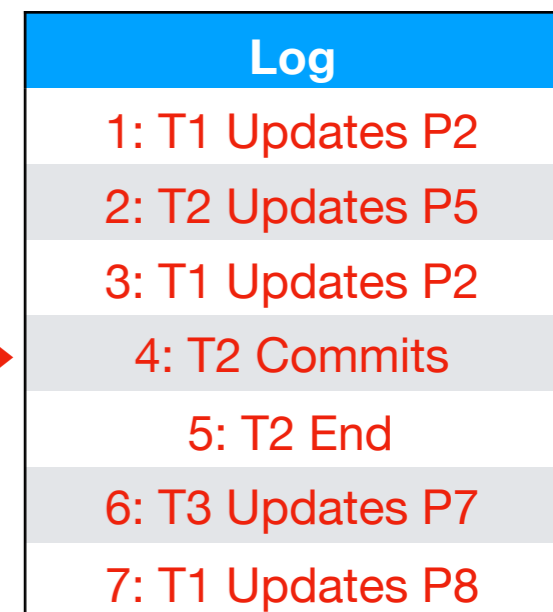
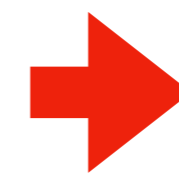
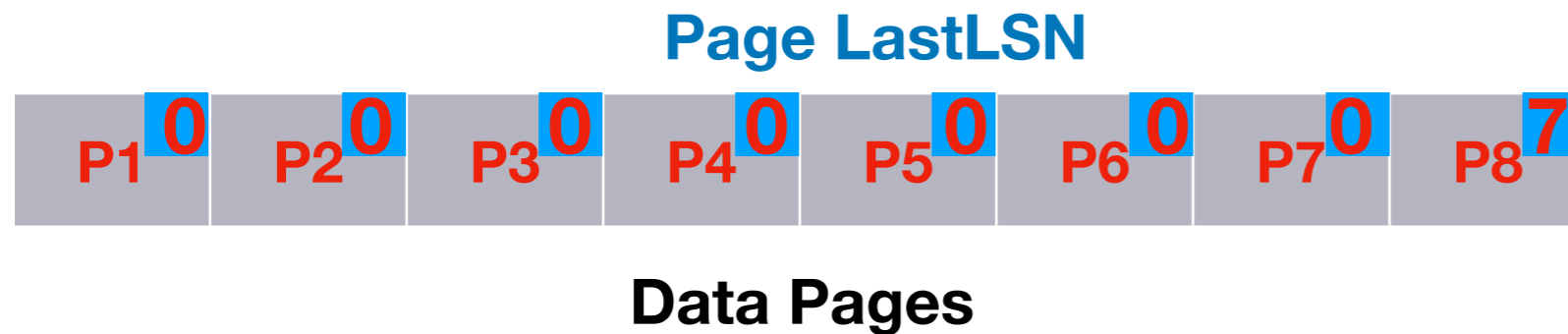
ARIES Example (Analysis)



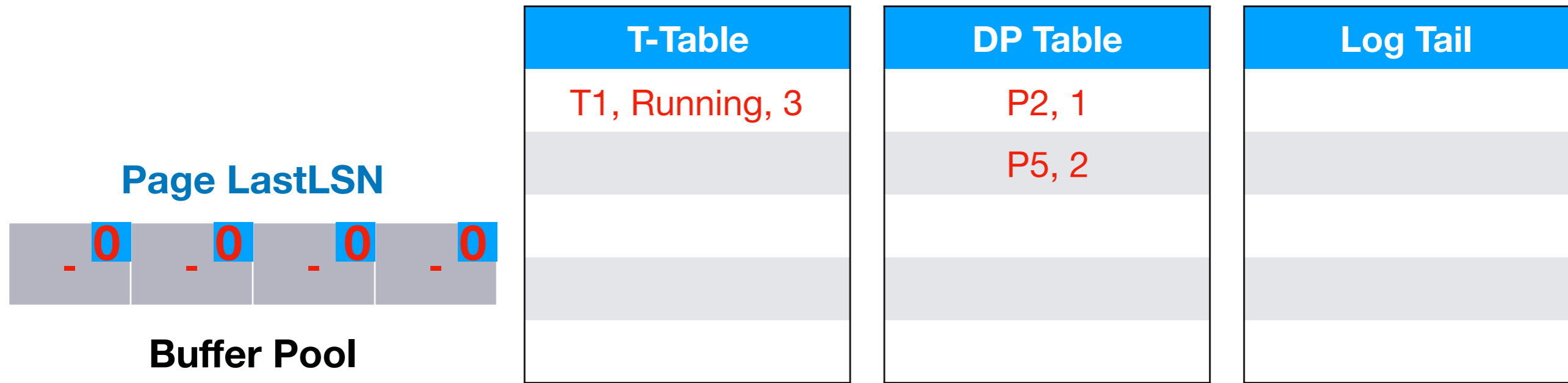
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



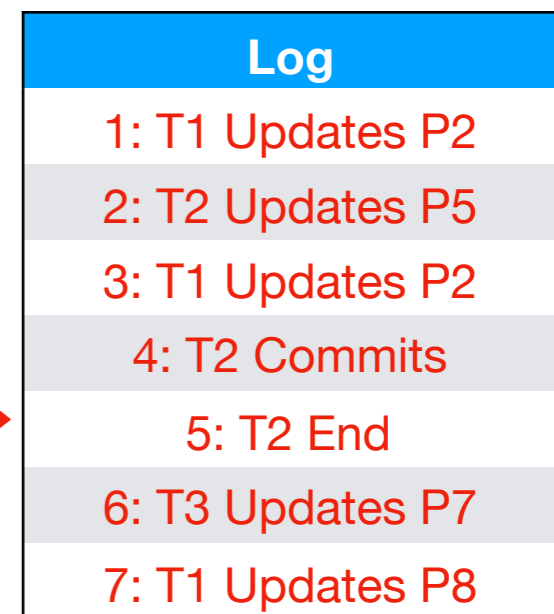
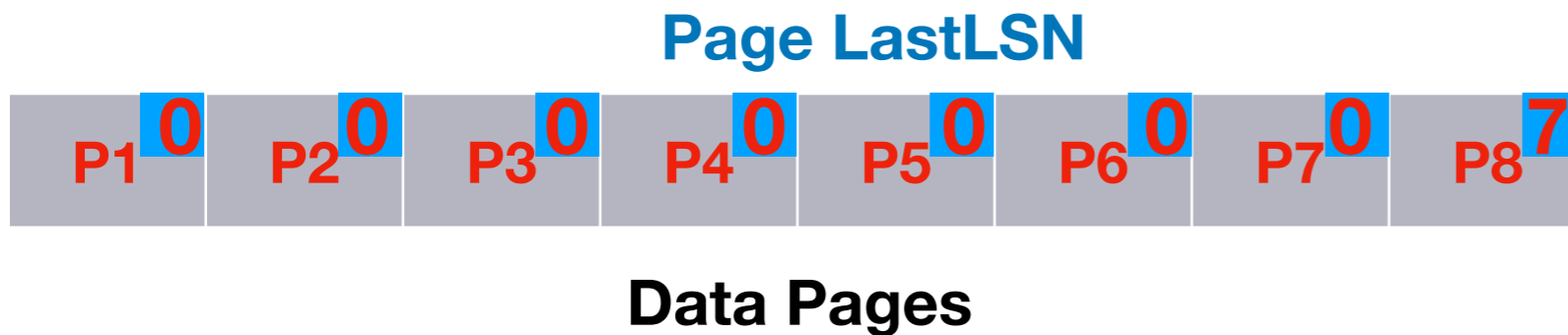
ARIES Example (Analysis)



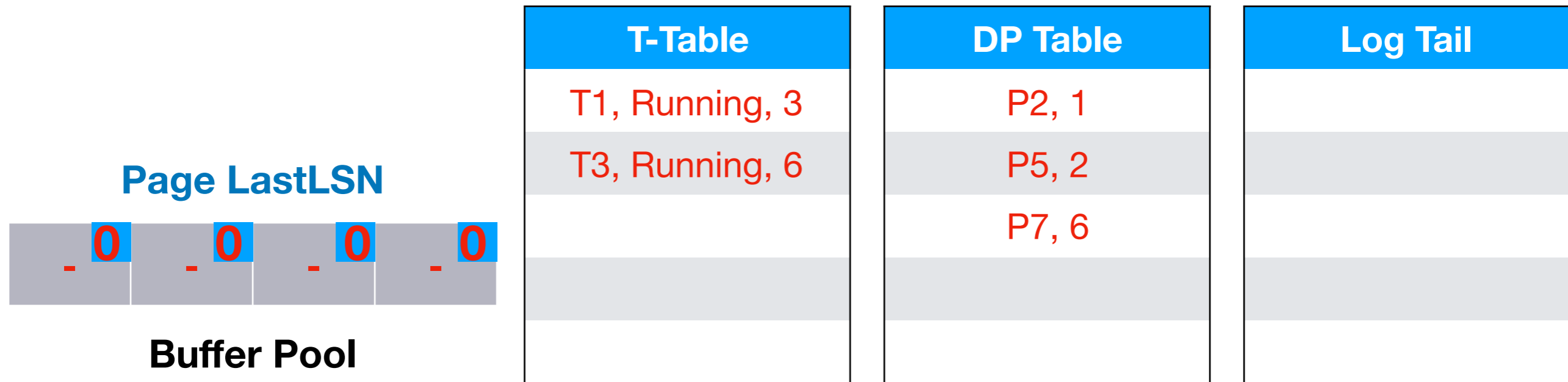
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



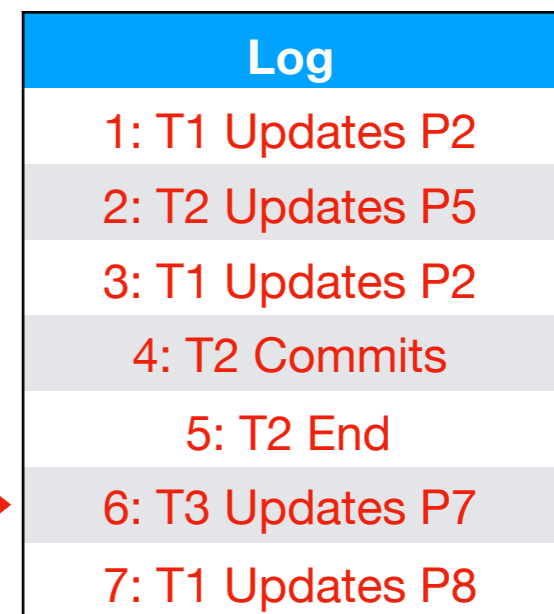
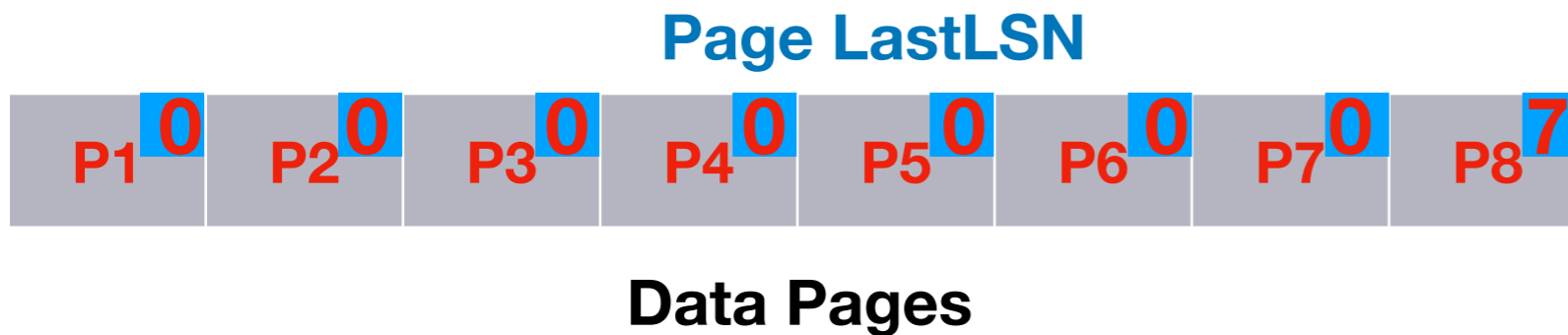
ARIES Example (Analysis)



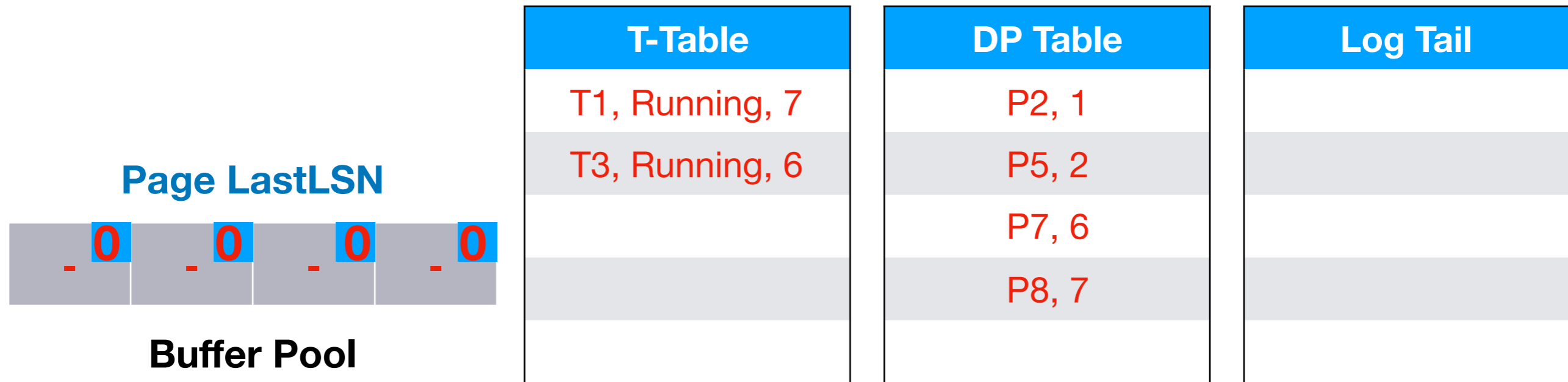
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



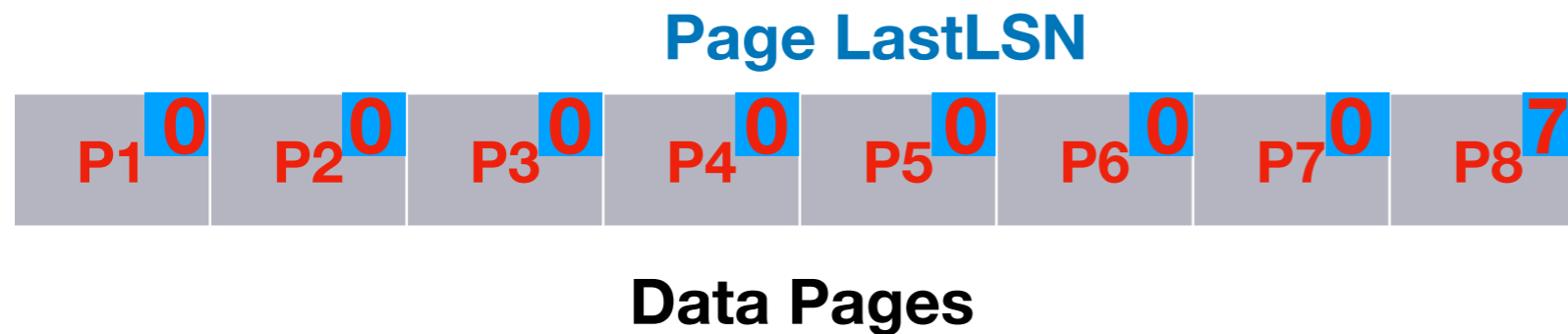
ARIES Example (Analysis)



FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



Data Pages



Log
1: T1 Updates P2
2: T2 Updates P5
3: T1 Updates P2
4: T2 Commits
5: T2 End
6: T3 Updates P7
7: T1 Updates P8

Redo Phase

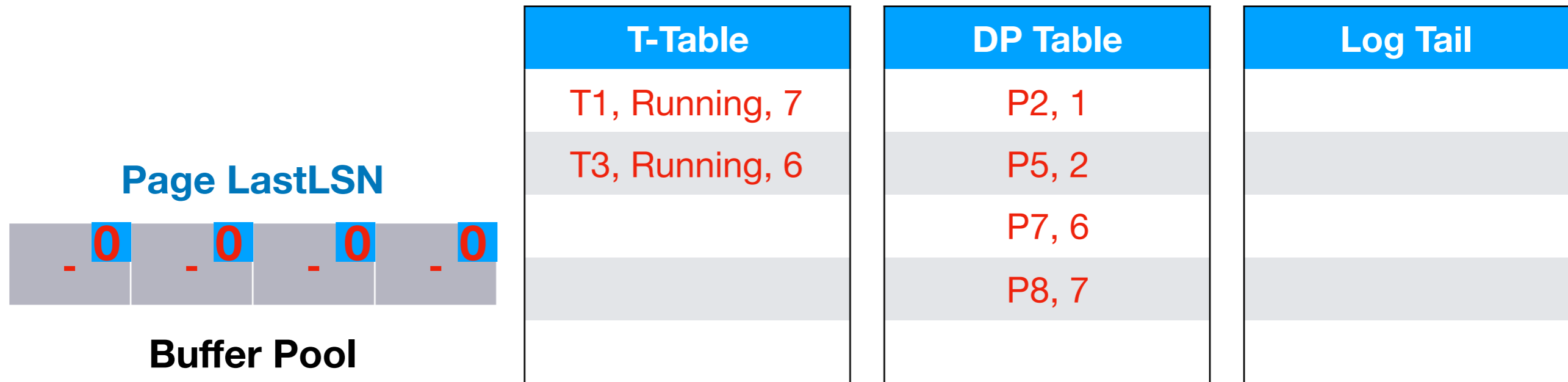
- Start scanning log from **earliest recLSN** in DP table
 - This is **first log entry** that made some page dirty
- Focus on **data updates** and **compensation log records**
 - **Redo change**, update pageLSN (no new log entry)
 - May **avoid redo** if certain conditions hold (see next)

Avoiding Redo Operations

- We can avoid a redo if one of those **conditions** holds
 1. Page affected by update is **not in dirty page table**
 2. Affected page in DP table but with **recLSN > LSN**
 3. PageLSN of page on hard disk with **pageLSN \geq LSN**

*Why Do We Check
Conditions in This Order?*

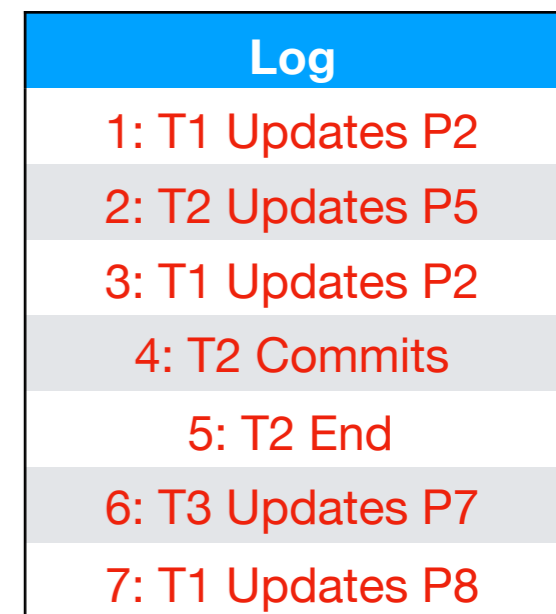
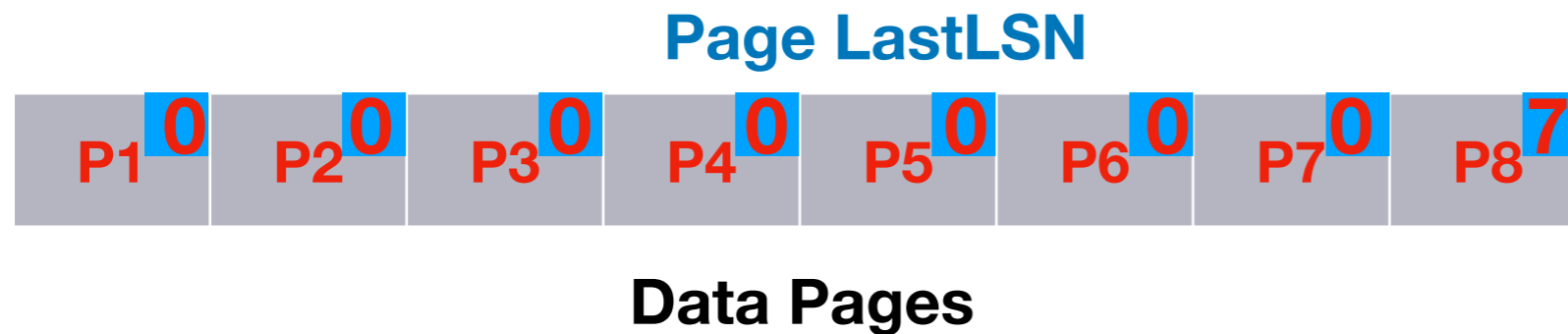
ARIES Example (Redo)



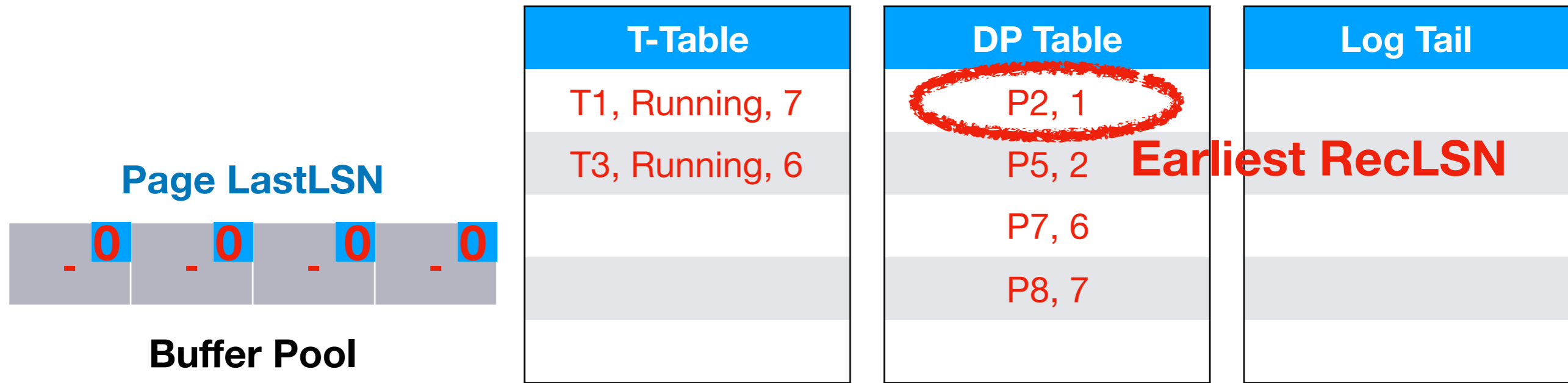
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



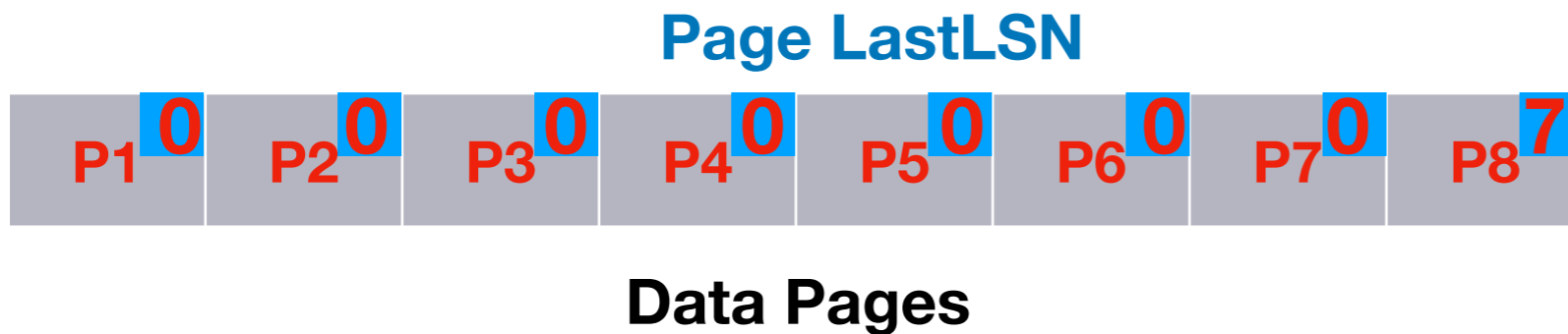
ARIES Example (Redo)



FlushedLSN: 0

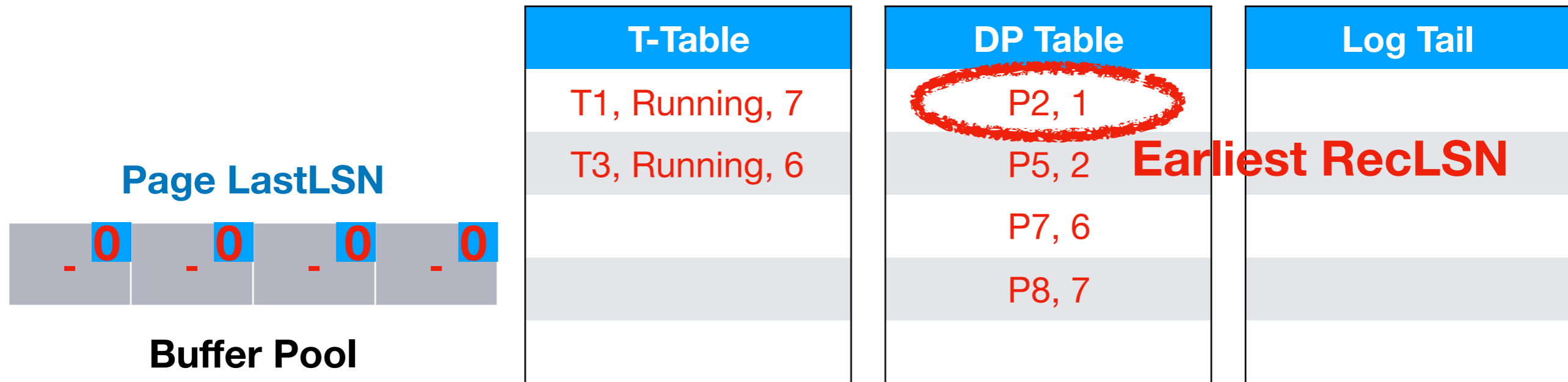
Main Memory (Volatile)

Hard Disk (Non-Volatile)



Log
1: T1 Updates P2
2: T2 Updates P5
3: T1 Updates P2
4: T2 Commits
5: T2 End
6: T3 Updates P7
7: T1 Updates P8

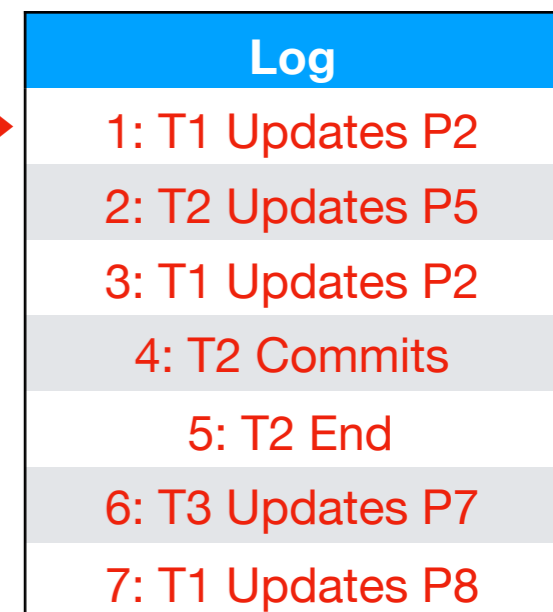
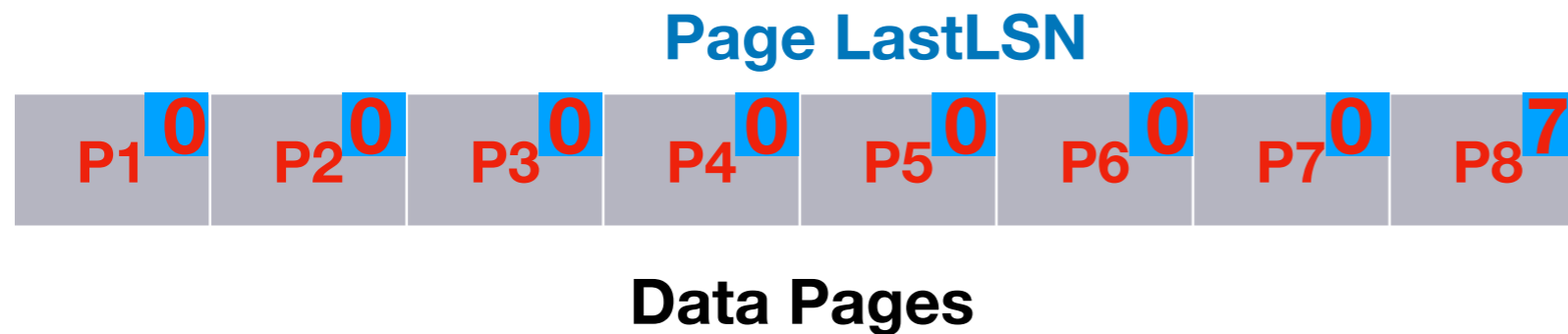
ARIES Example (Redo)



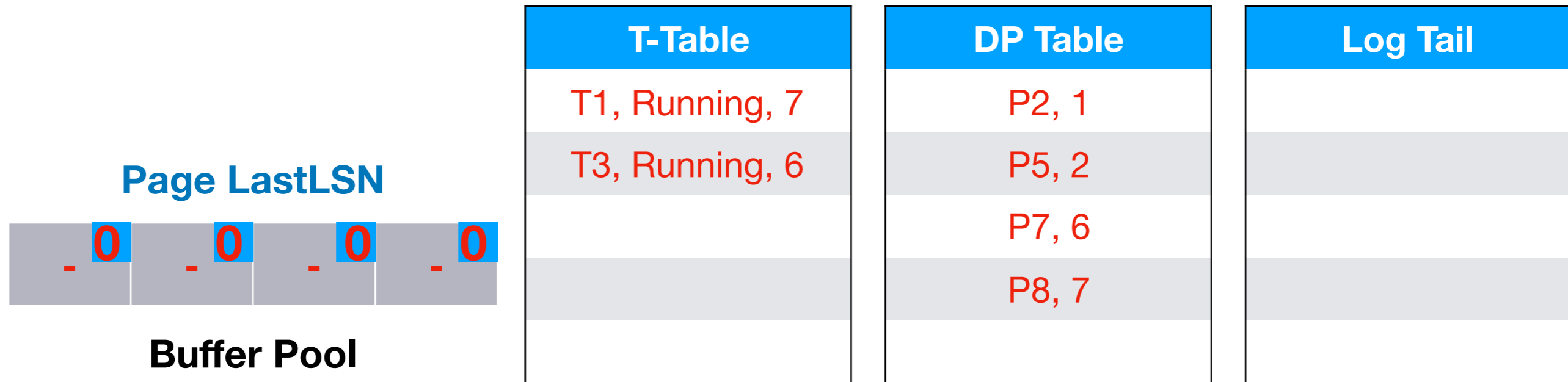
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



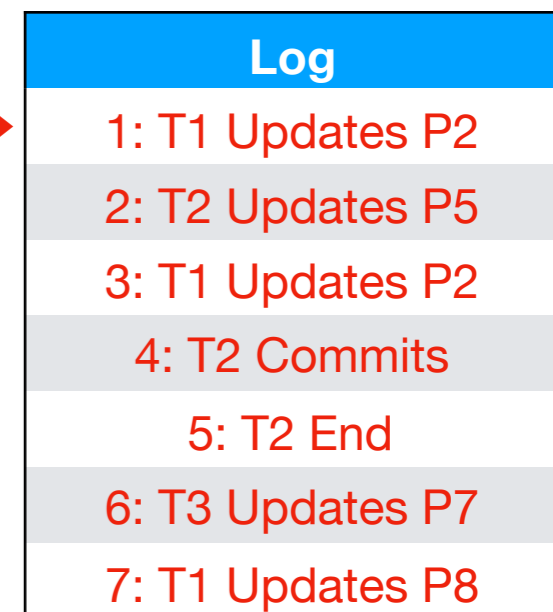
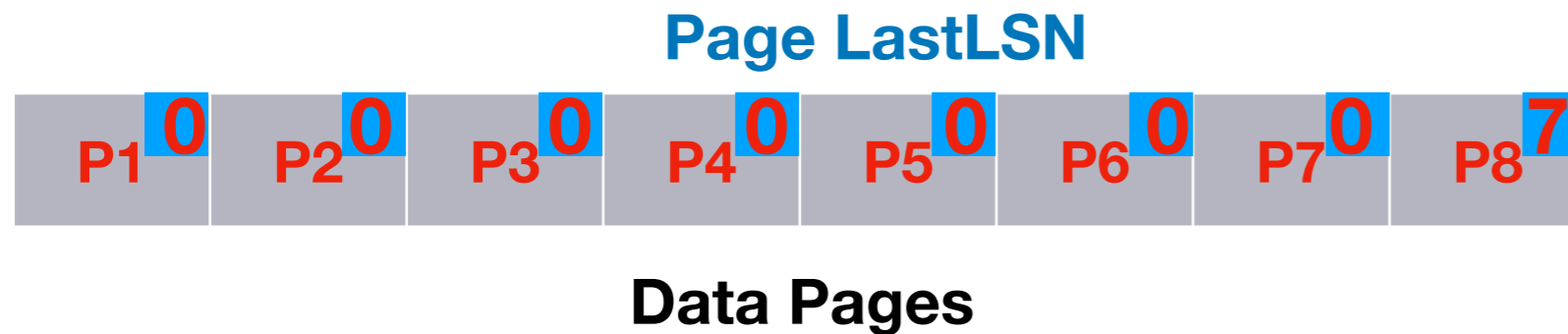
ARIES Example (Redo)



FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



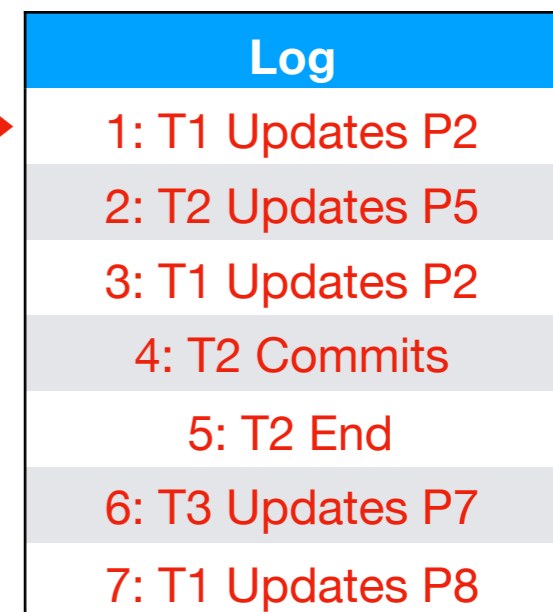
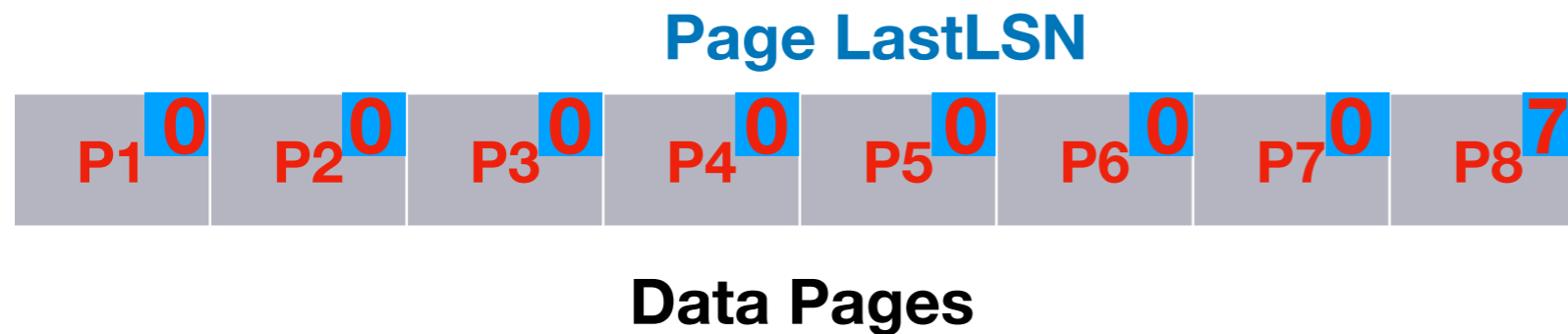
ARIES Example (Redo)



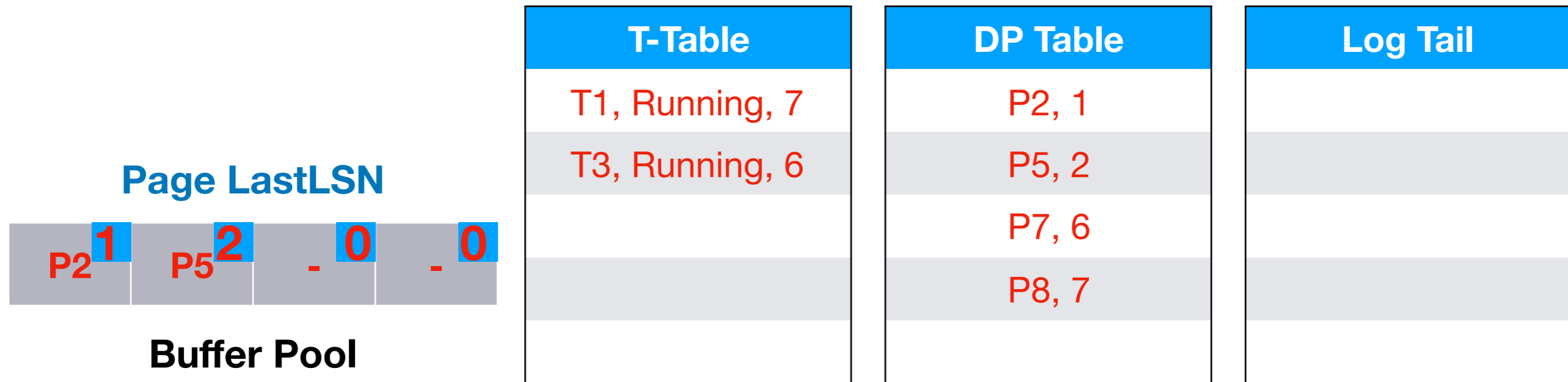
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



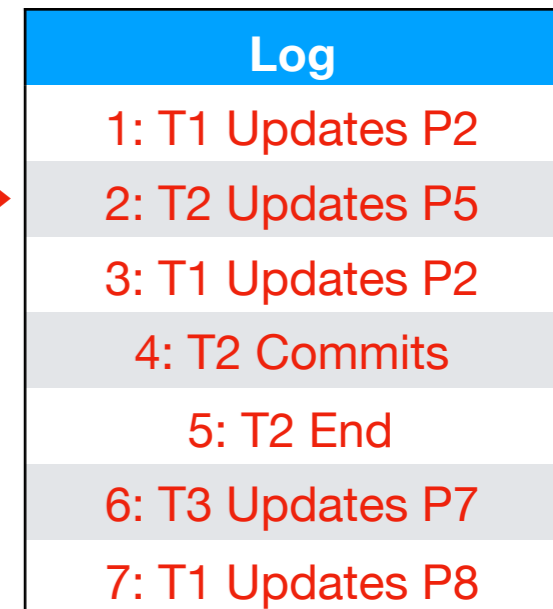
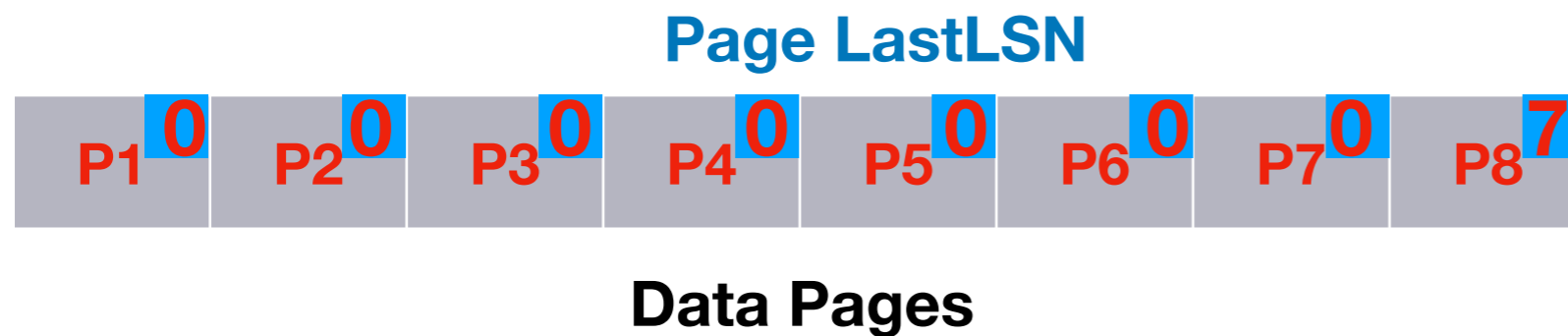
ARIES Example (Redo)



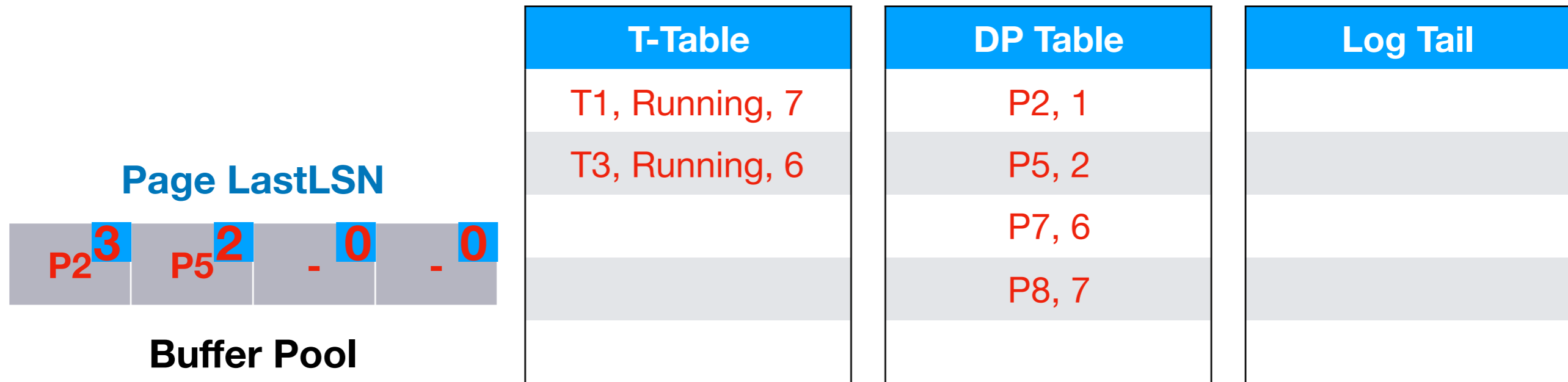
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



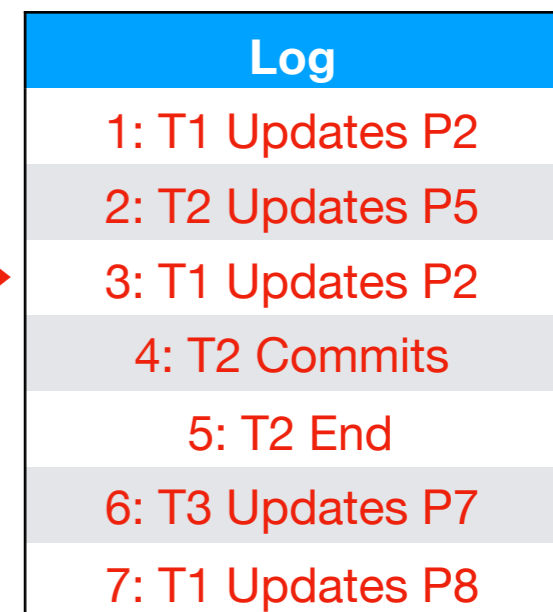
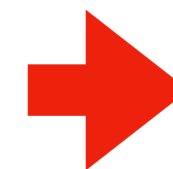
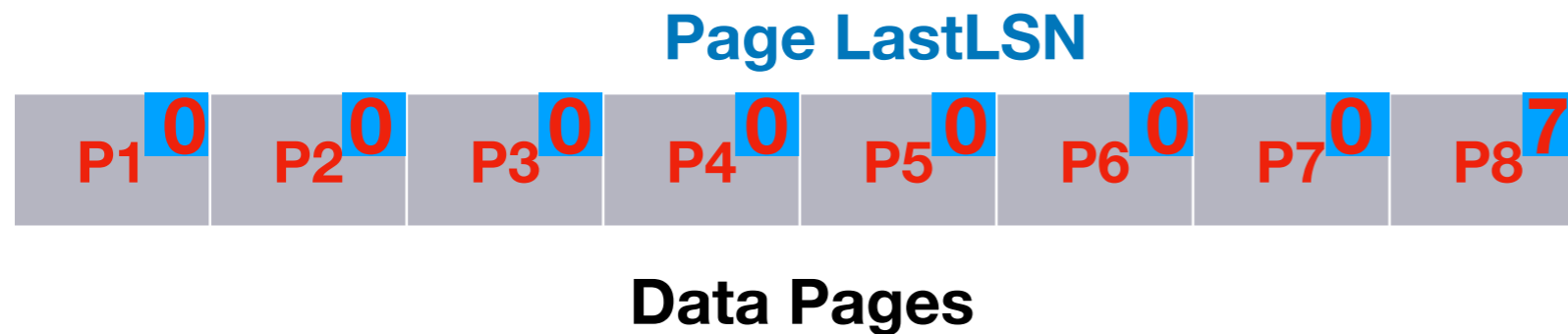
ARIES Example (Redo)



FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



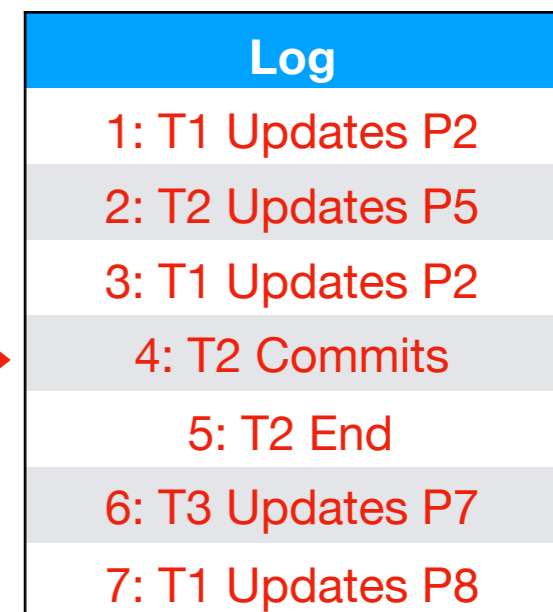
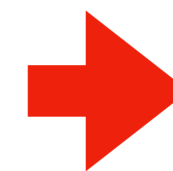
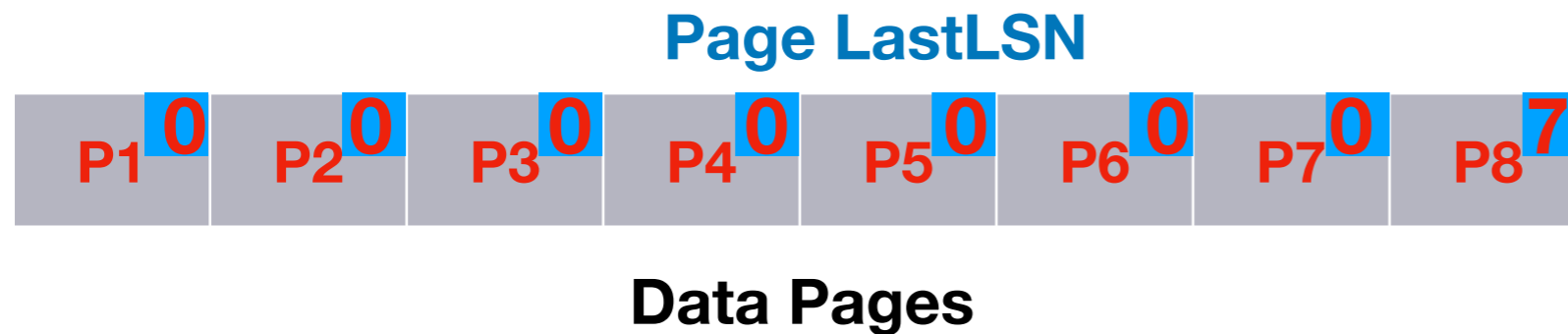
ARIES Example (Redo)



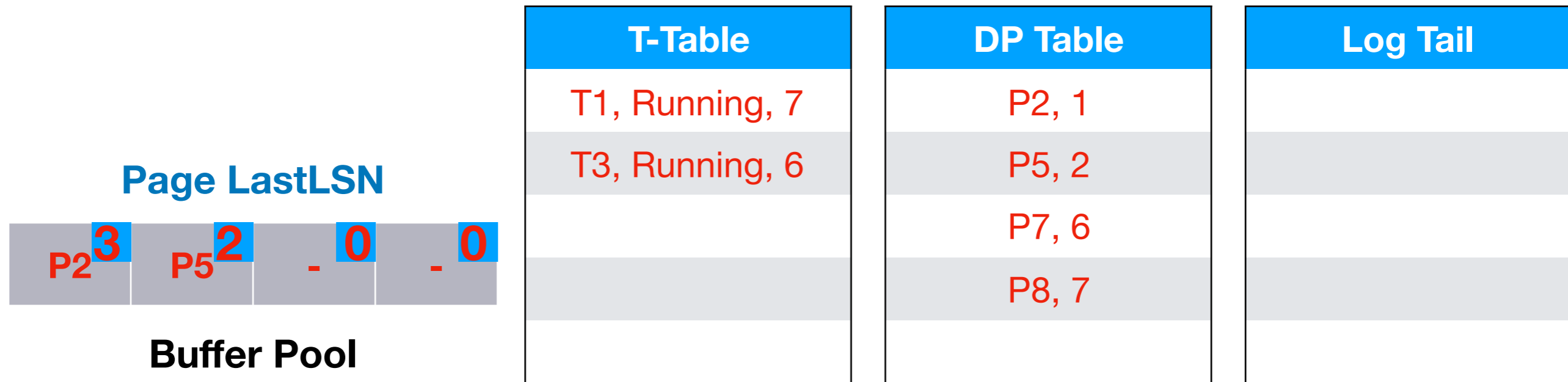
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



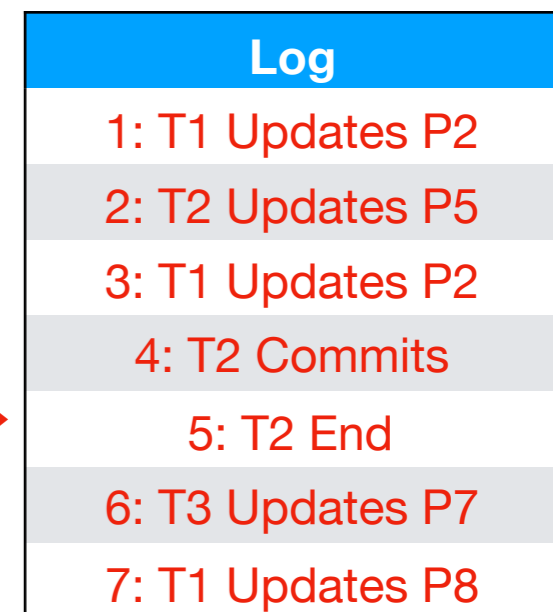
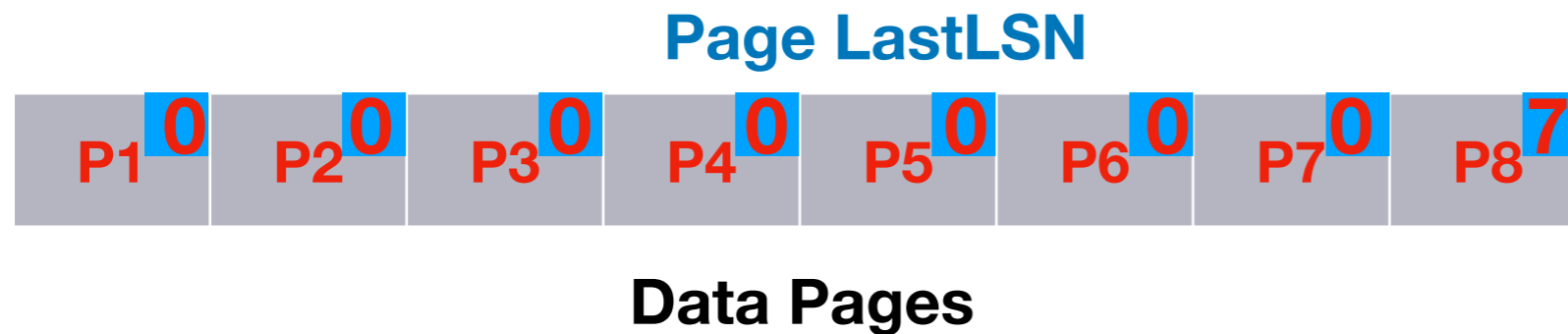
ARIES Example (Redo)



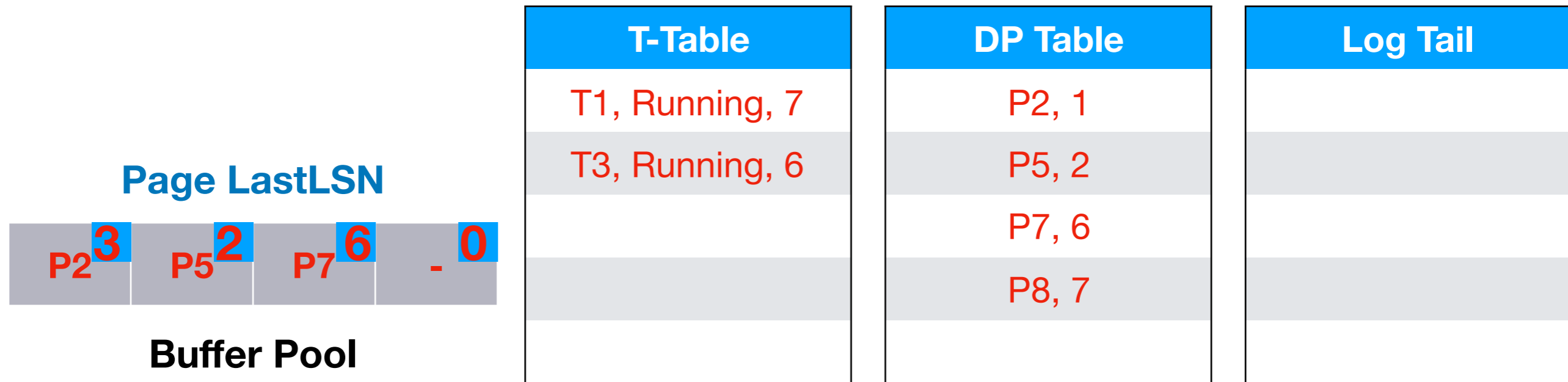
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



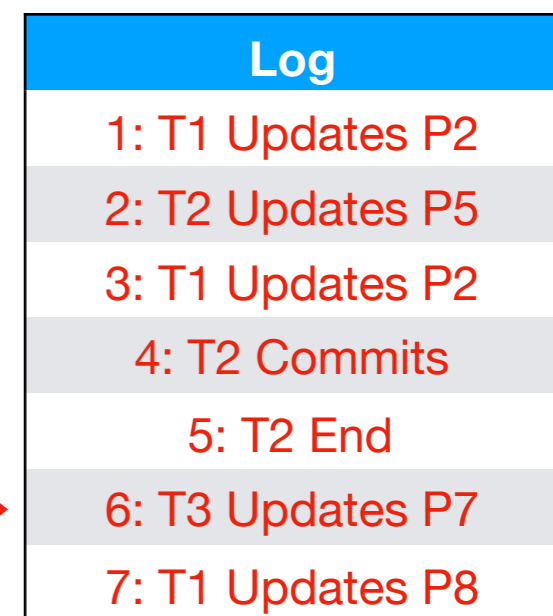
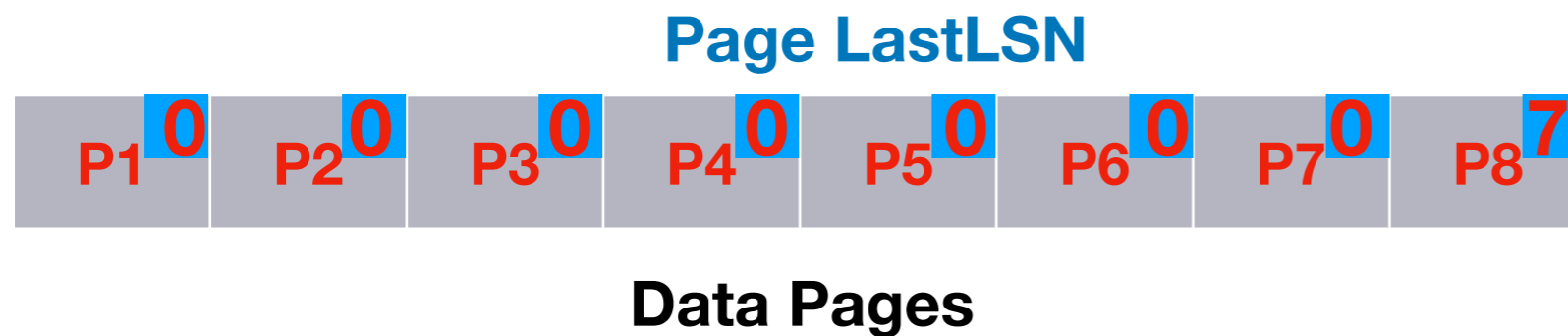
ARIES Example (Redo)



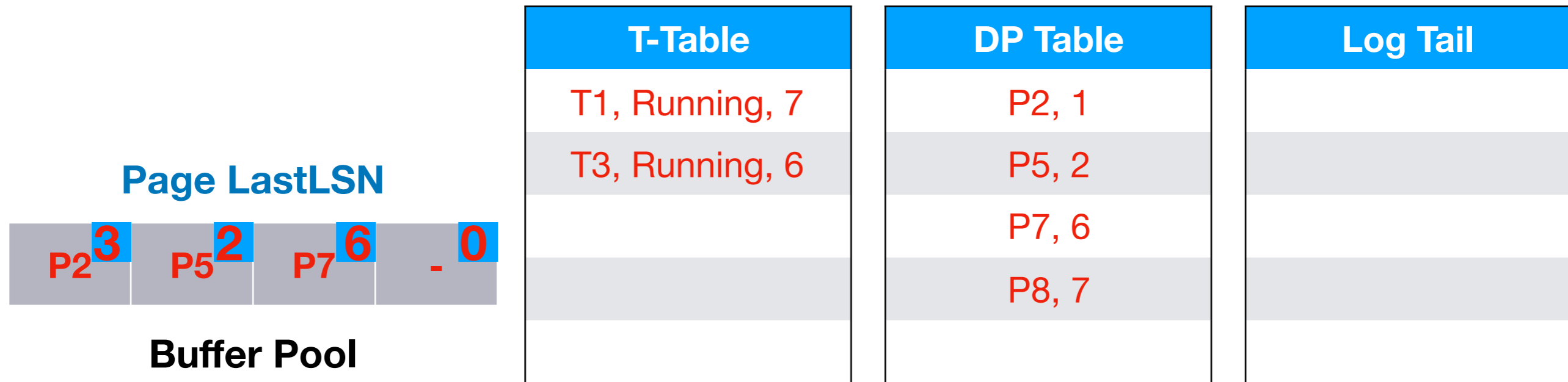
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



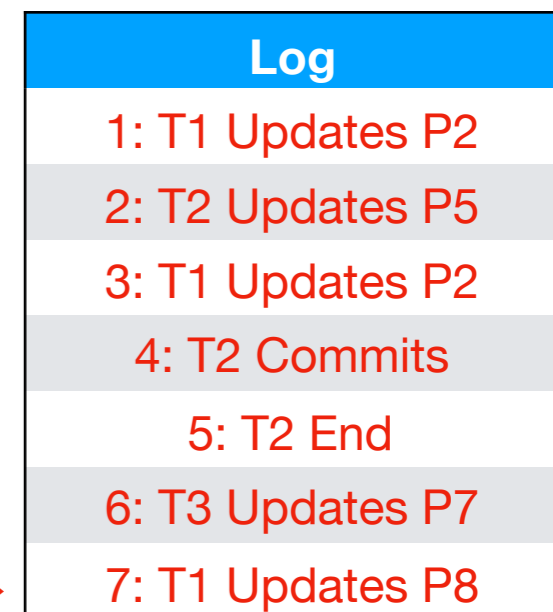
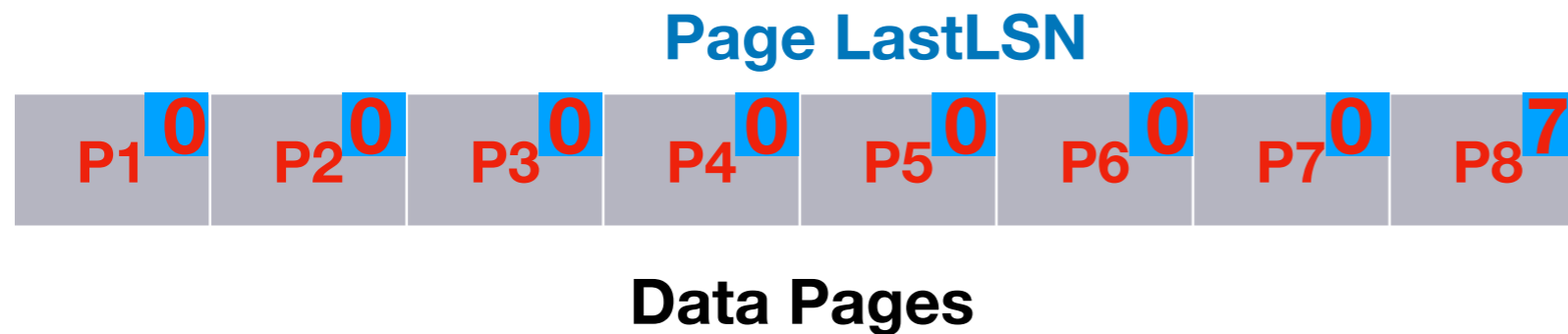
ARIES Example (Redo)



FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



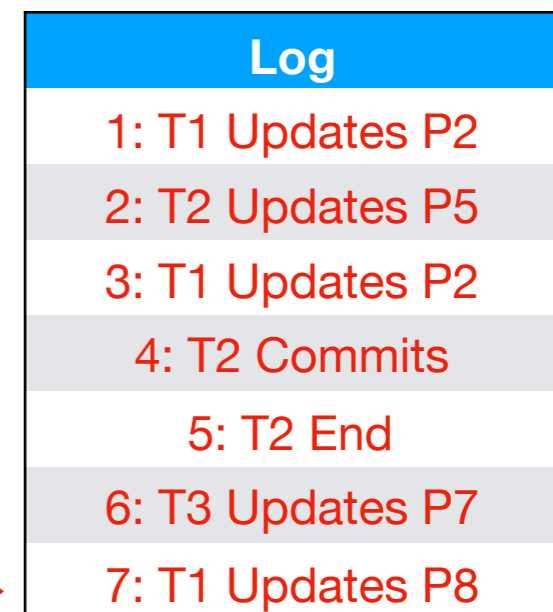
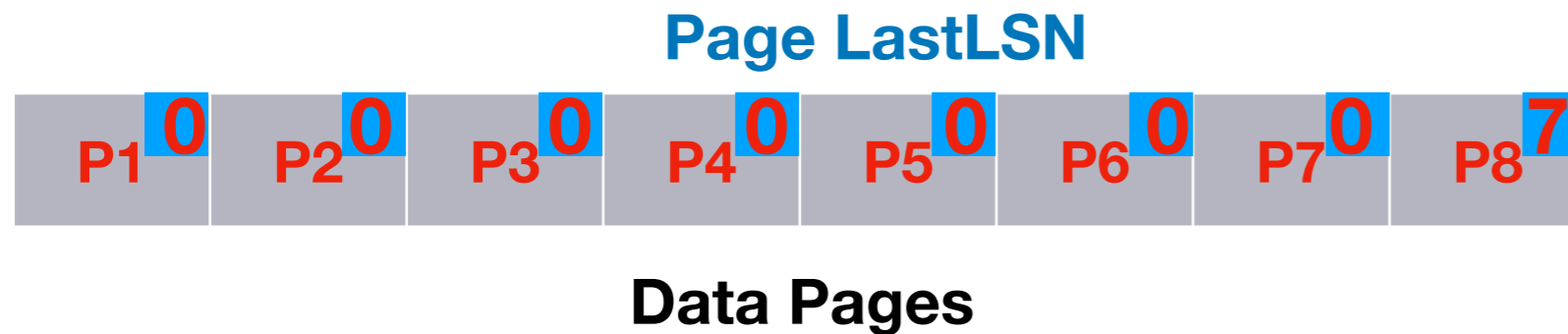
ARIES Example (Redo)



FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



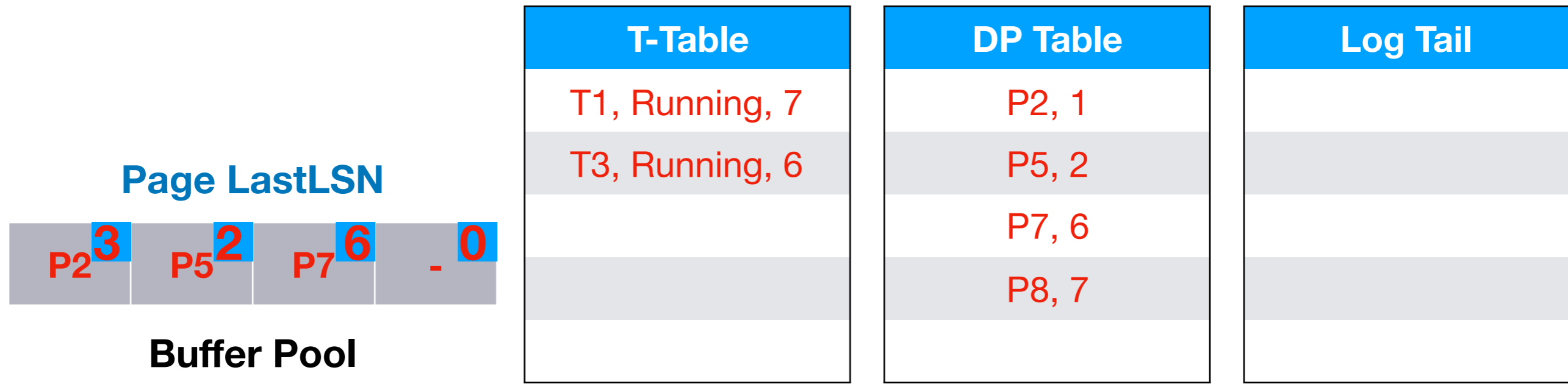
Do We Need To Redo This ... ?



Undo Phase

- Undo changes by transactions **running at crash**
- Undo changes in **reverse log order**
 - Follow **prevLSN pointers** backwards in log
 - Write **compensation log records** while undoing

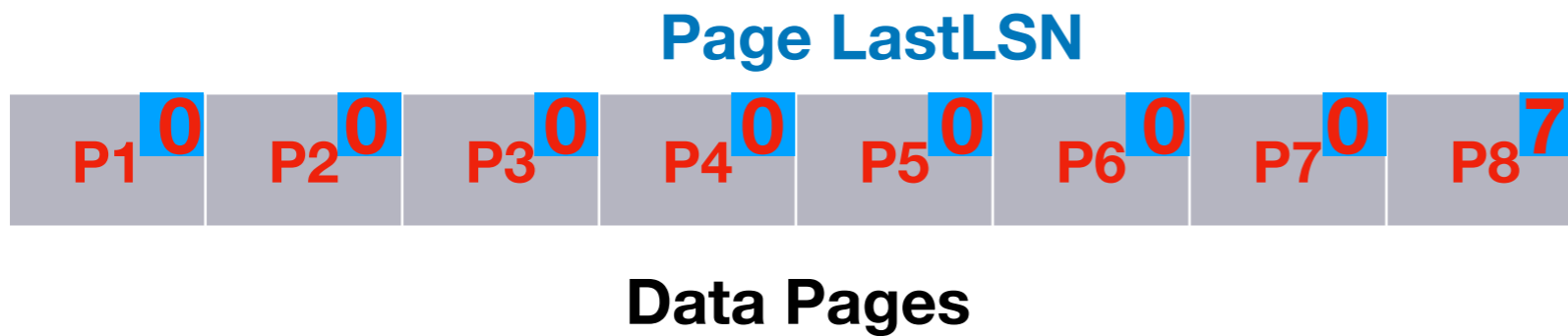
ARIES Example (Undo)



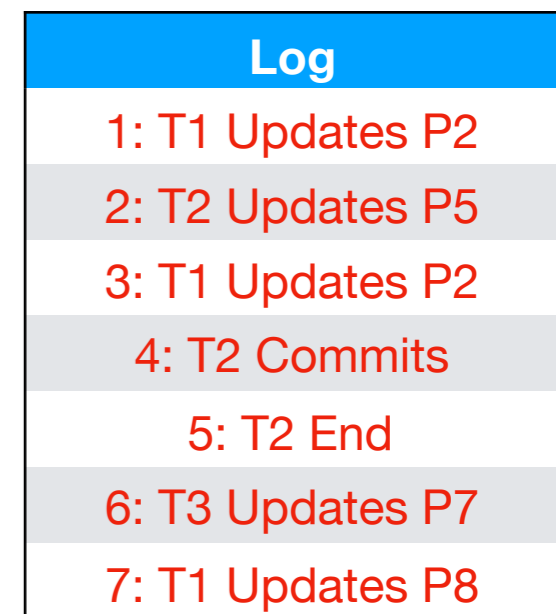
FlushedLSN: 0

Main Memory (Volatile)

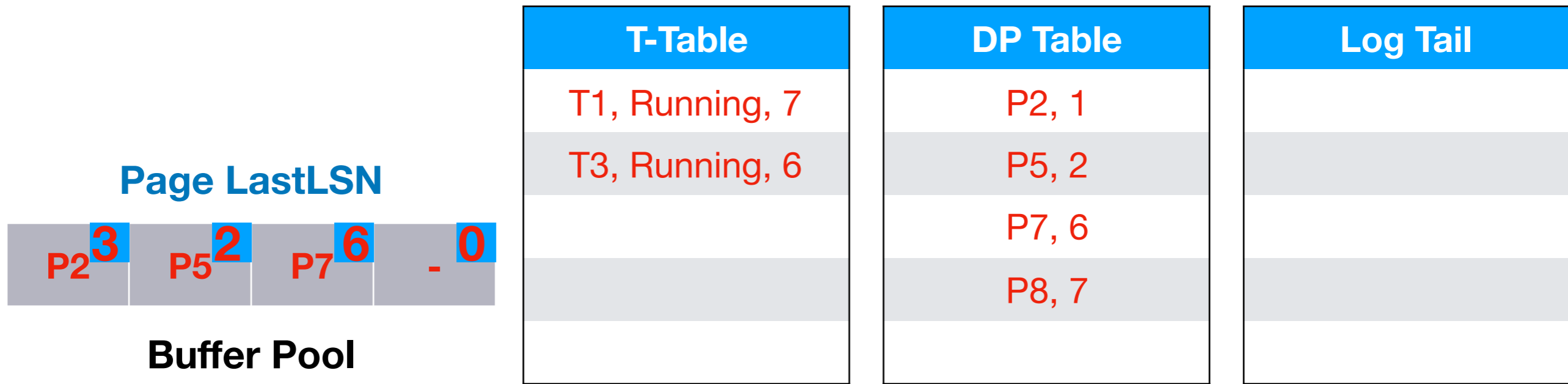
Hard Disk (Non-Volatile)



PrevLSNs



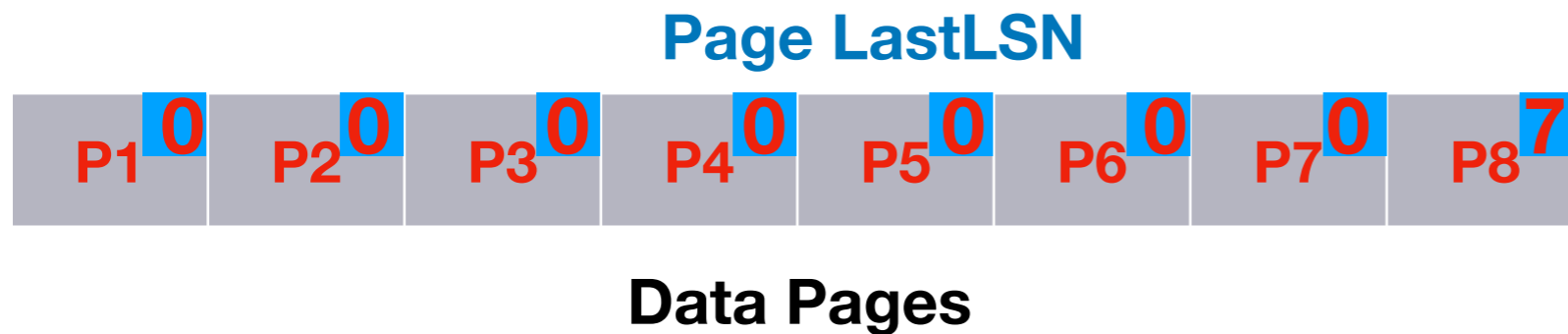
ARIES Example (Undo)



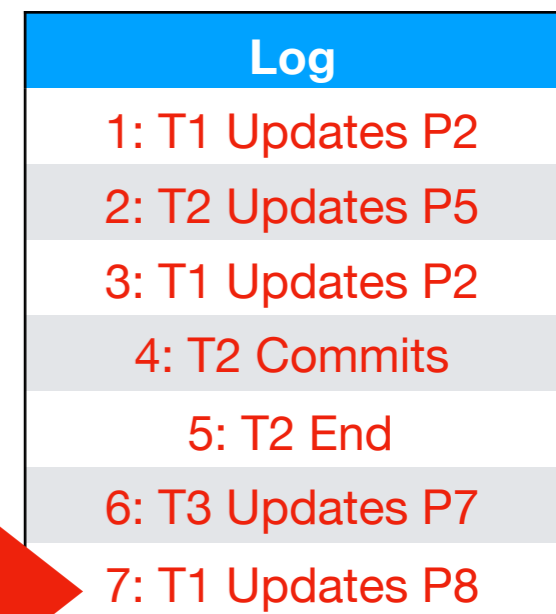
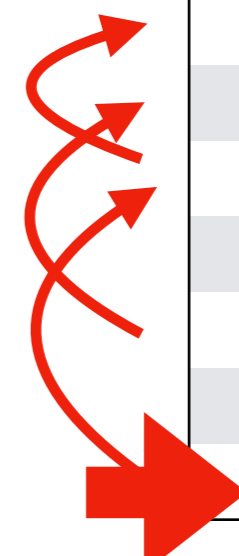
FlushedLSN: 0

Main Memory (Volatile)

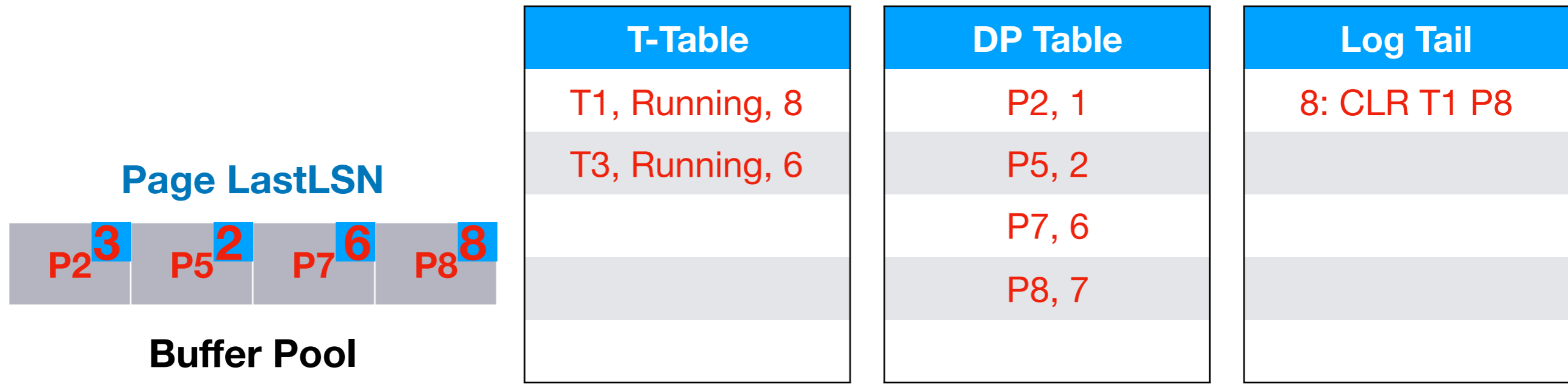
Hard Disk (Non-Volatile)



PrevLSNs



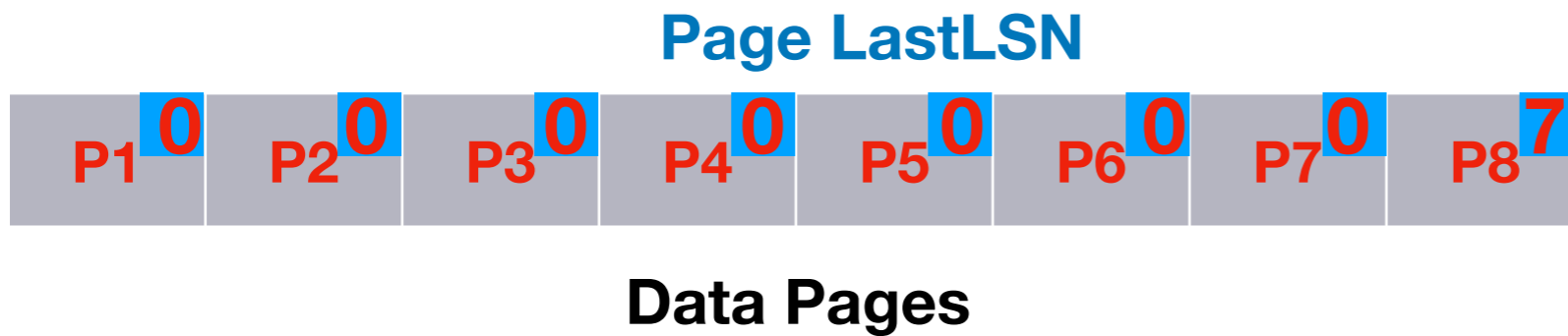
ARIES Example (Undo)



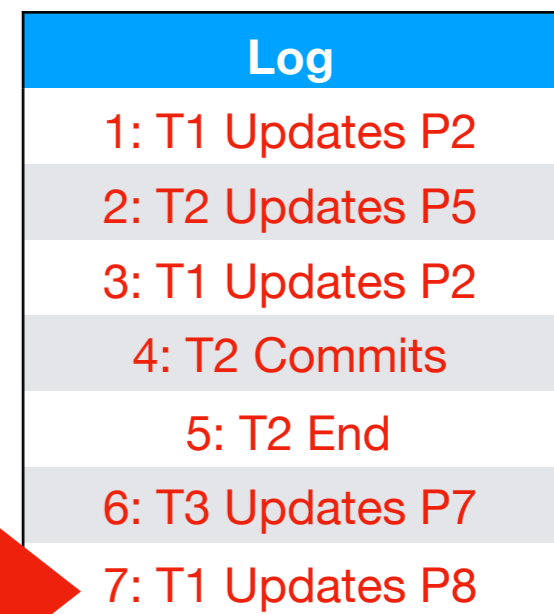
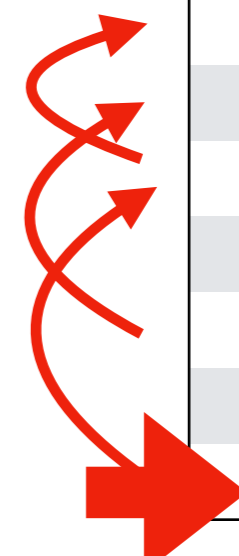
FlushedLSN: 0

Main Memory (Volatile)

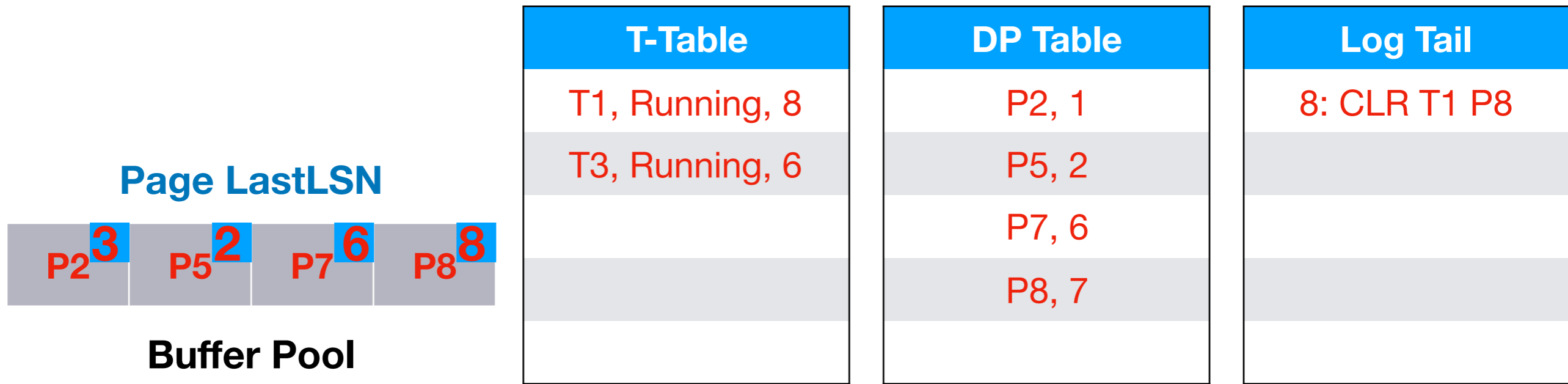
Hard Disk (Non-Volatile)



PrevLSNs



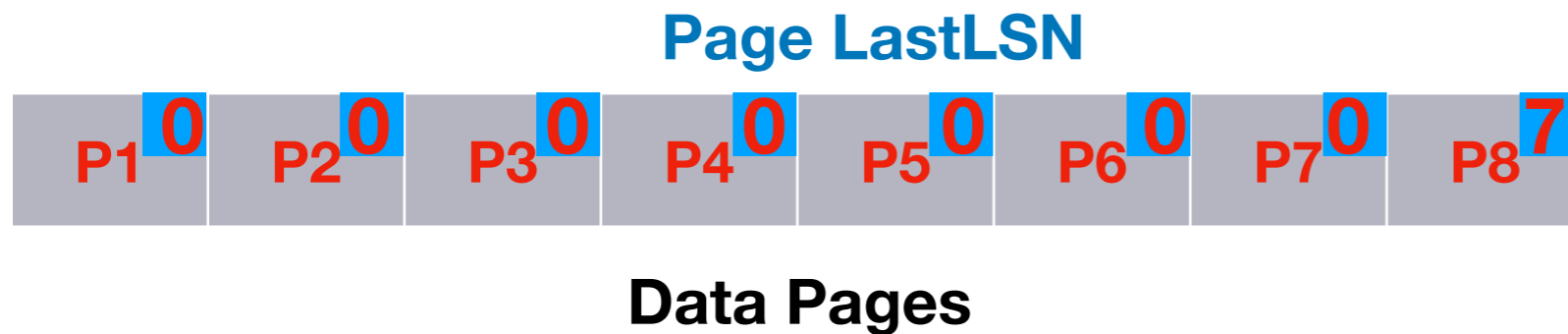
ARIES Example (Undo)



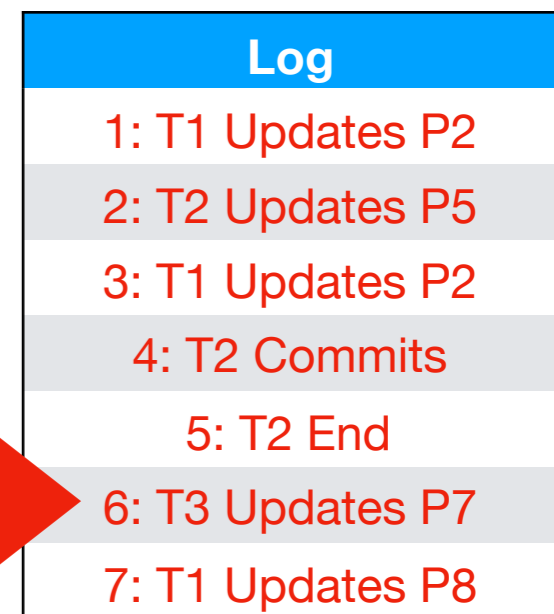
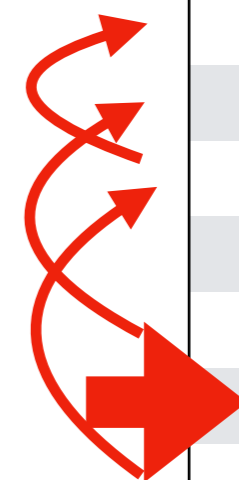
FlushedLSN: 0

Main Memory (Volatile)

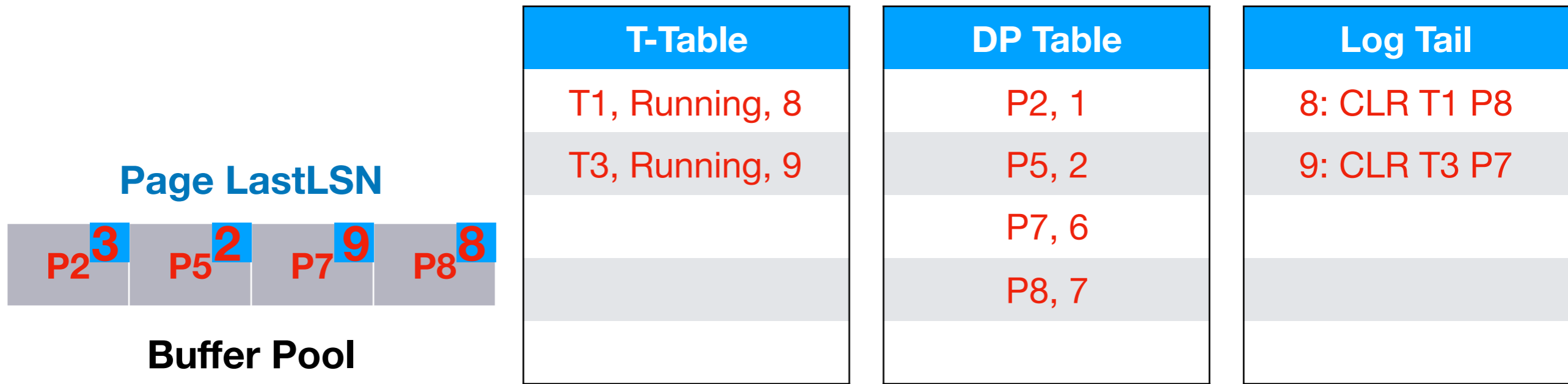
Hard Disk (Non-Volatile)



PrevLSNs



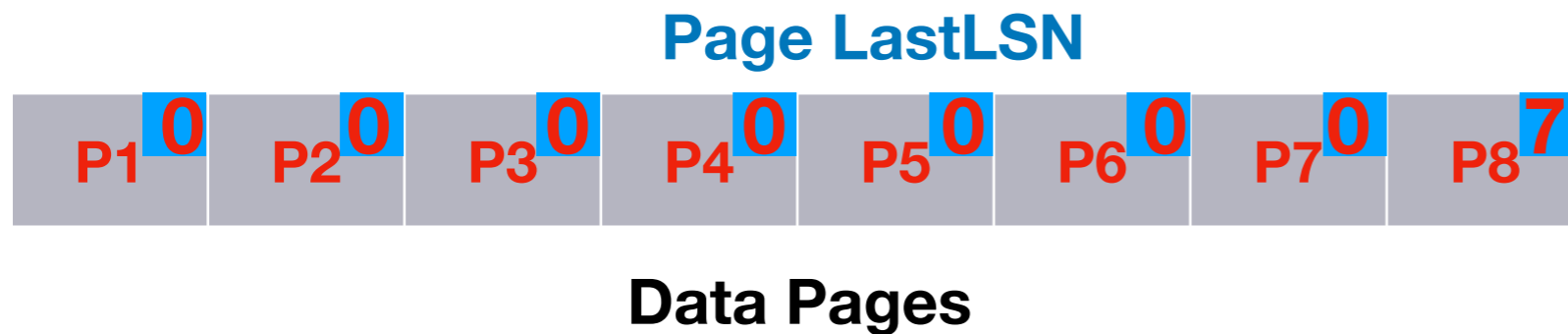
ARIES Example (Undo)



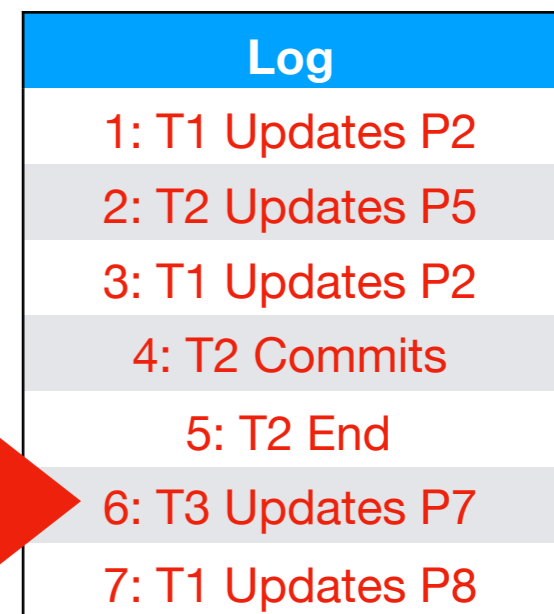
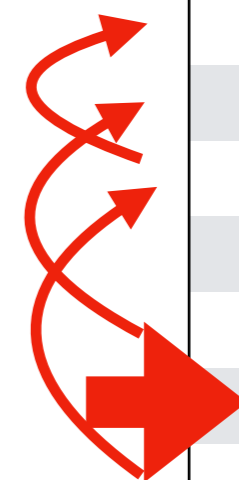
FlushedLSN: 0

Main Memory (Volatile)

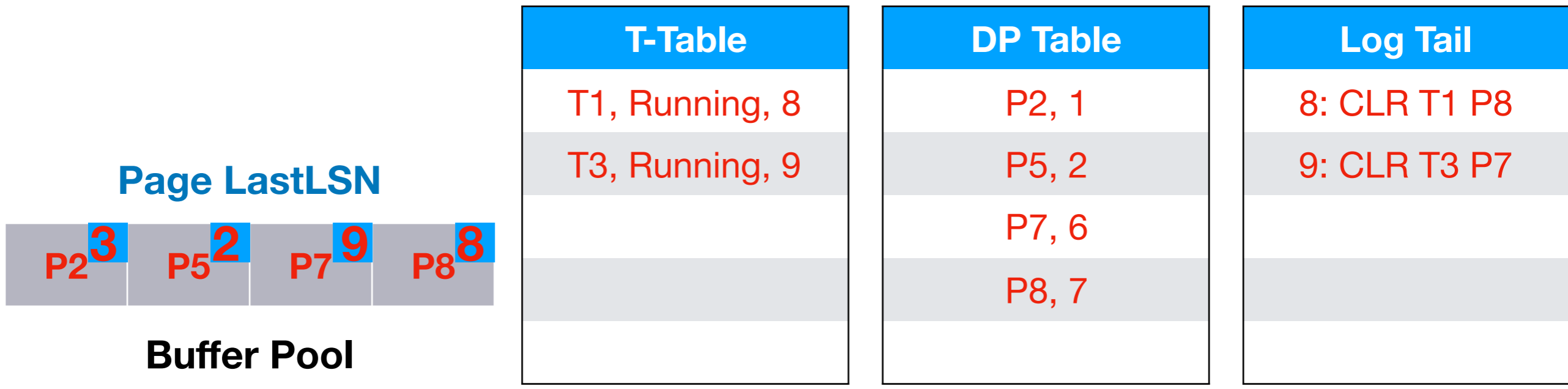
Hard Disk (Non-Volatile)



PrevLSNs



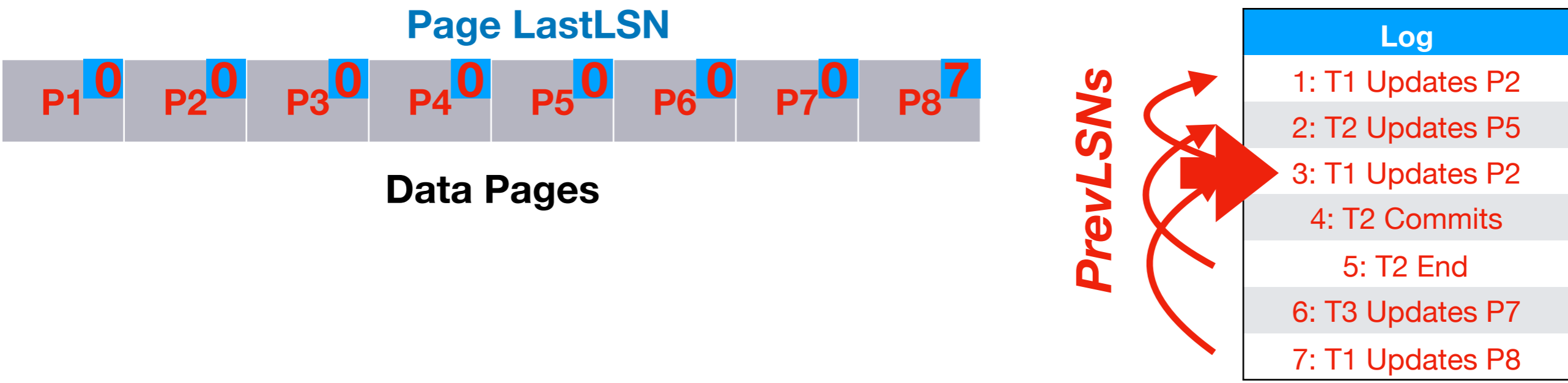
ARIES Example (Undo)



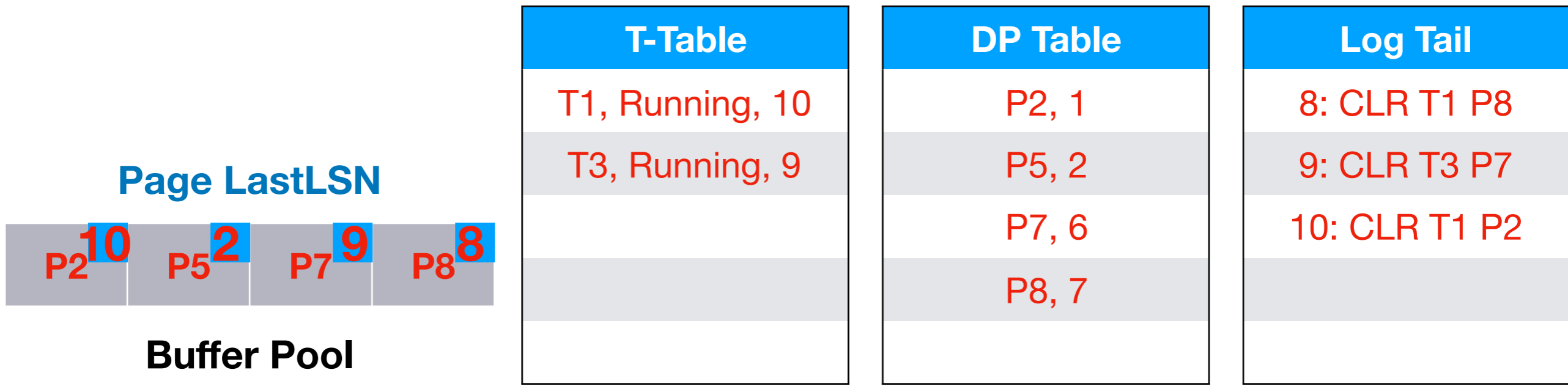
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



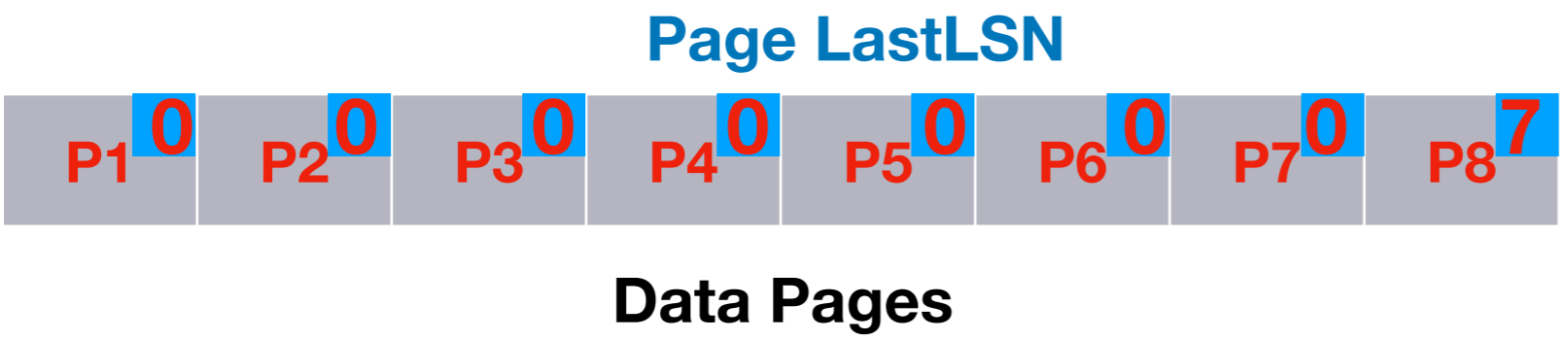
ARIES Example (Undo)



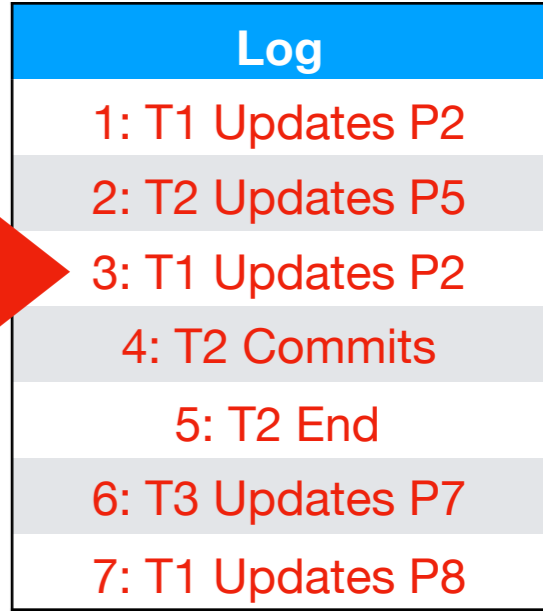
FlushedLSN: 0

Main Memory (Volatile)

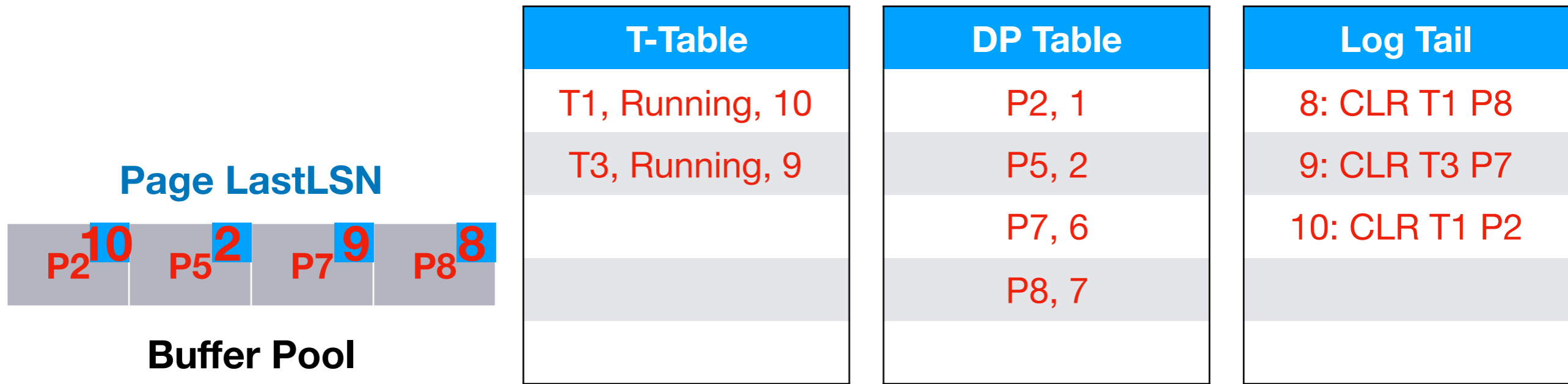
Hard Disk (Non-Volatile)



PrevLSNs

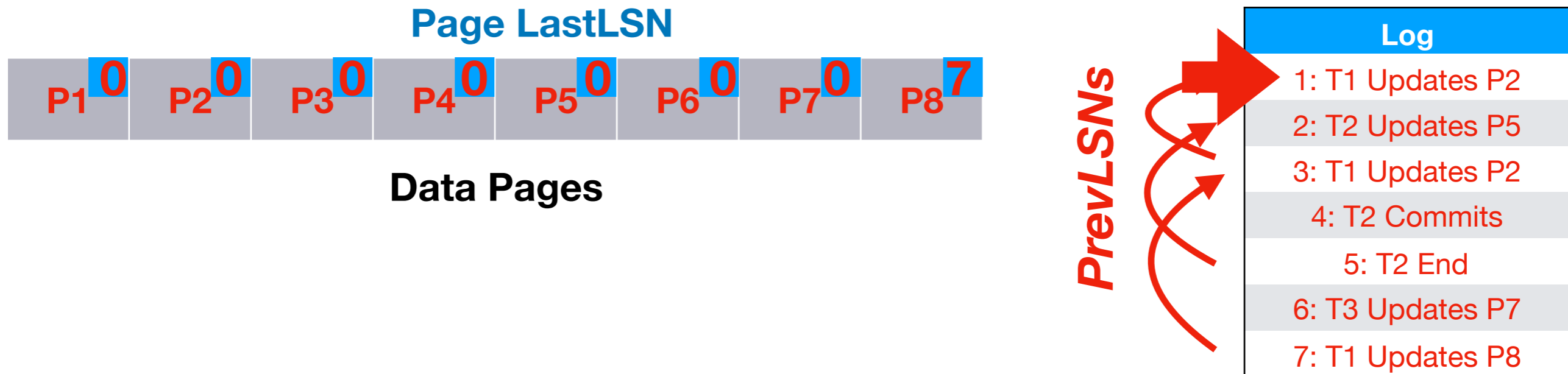


ARIES Example (Undo)

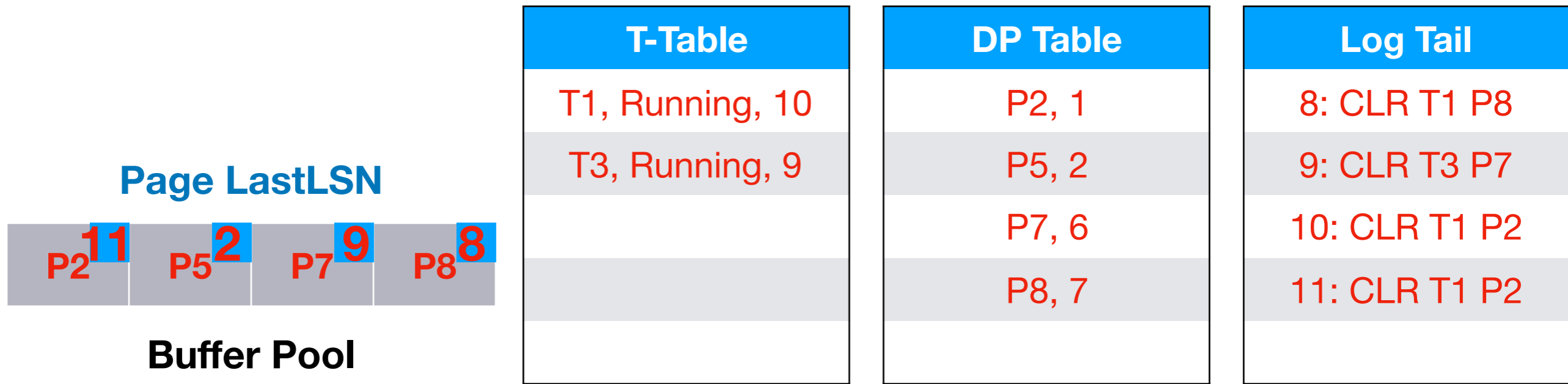


Main Memory (Volatile)

Hard Disk (Non-Volatile)



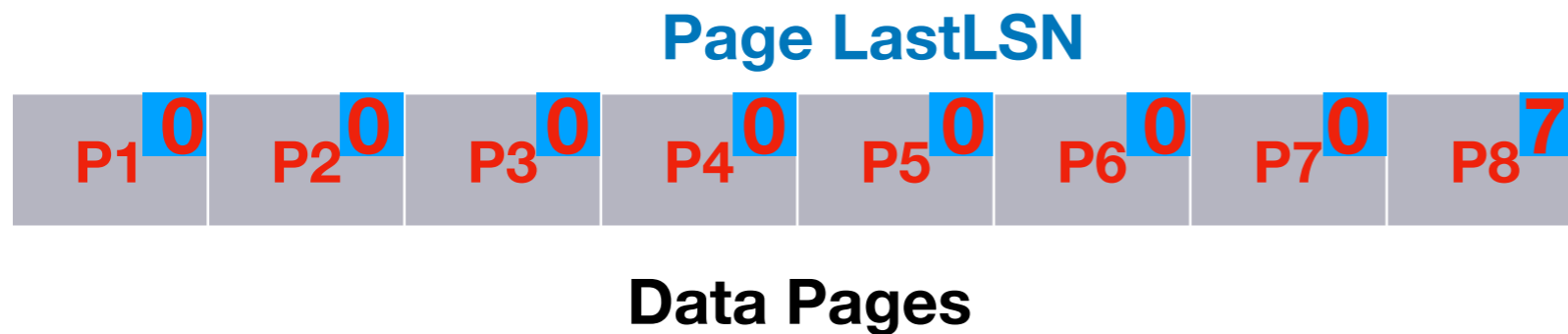
ARIES Example (Undo)



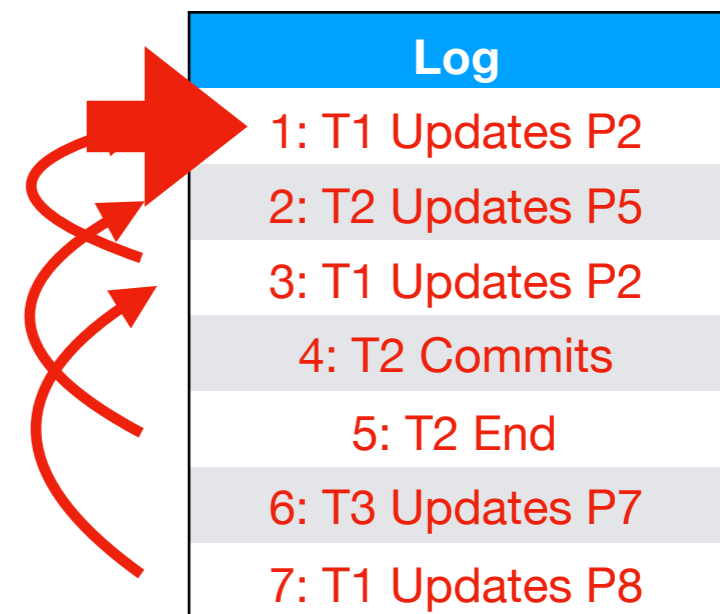
FlushedLSN: 0

Main Memory (Volatile)

Hard Disk (Non-Volatile)



PrevLSNs



Crash During Recovery?

- ARIES can deal with failures during **any** recovery phase
- Crash during analysis? Simply **restart analysis**.
- Crash during redo? Restart **analysis and redo**.
 - May not have to redo everything if **changes persisted**
- Crash during undo? **Restart all phases**.
 - Note: **CLRs** are redone but never undone!