Lecture #13 – Physical Logging

Andy Pavlo // Carnegie Mellon University // Spring 2016

TODAY'S AGENDA

Logging Schemes
Crash Course on ARIES
In-Memory Database Logging & Recovery
Evaluation

LOGGING & RECOVERY

Recovery algorithms are techniques to ensure database **consistency**, txn **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.

LOGGING SCHEMES

Physical Logging

→ Record the changes made to a specific record in the database.
→ Example: Store the original value and after value for an attribute that is changed by a query.

Logical Logging

→ Record the high-level operations executed by txns.
→ Example: The **UPDATE**, **DELETE**, and **INSERT** queries invoked by a txn.

PHYSICAL VS. LOGICAL LOGGING

Logical logging writes less data in each log record than physical logging.

Difficult to implement recovery with logical logging if you have concurrent txns.

→ Hard to determine which parts of the database may have been modified by a query before crash.
→ Also takes longer to recover because you must re-execute every txn all over again.

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$100
El-P	$666
Joy	$888

Logical Log

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = ‘Joy’
```

NAME	SALARY
O.D.B.	$100
El-P	$666
Joy	$888

Logical Log

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$100
El-P	$666
Joy	$888

Logical Log

```
UPDATE employees SET
salary = salary * 1.10
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$100
El-P	$666
Joy	$888

Logical Log

```
UPDATE employees SET
salary = salary * 1.10
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$110
El-P	$732
Joy	$888

Logical Log

```
UPDATE employees SET
salary = salary * 1.10
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$110
El-P	$732
Joy	$888

Logical Log

```
UPDATE employees SET
salary = salary * 1.10

UPDATE employees SET
salary = 900 WHERE
name = 'Joy'
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$110
El-P	$732
Joy	$888

Logical Log

```
UPDATE employees SET
salary = salary * 1.10

UPDATE employees SET
salary = 900 WHERE
name = 'Joy'
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$110
El-P	$732
Joy	$900

Logical Log

```
UPDATE employees SET
salary = salary * 1.10

UPDATE employees SET
salary = 900 WHERE
name = 'Joy'
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = ‘Joy’
```

NAME	SALARY
O.D.B.	$110
El-P	$732
Joy	$990

Logical Log

```
UPDATE employees SET
salary = salary * 1.10

UPDATE employees SET
salary = 900 WHERE
name = ‘Joy’
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B.	$110
El-P	$732
Joy	$990

Logical Log

```
UPDATE employees SET
salary = salary * 1.10

UPDATE employees SET
salary = 900 WHERE
name = 'Joy'
```

LOGICAL LOGGING EXAMPLE

```
UPDATE employees
   SET salary = salary * 1.10
```

```
UPDATE employees
   SET salary = 900
 WHERE name = 'Joy'
```

NAME	SALARY
O.D.B	$110
El-P	$732
Joy	$990

SALARY
$110
$732
$900

Logical Log

```
UPDATE employees SET
salary = salary * 1.10

UPDATE employees SET
salary = 900 WHERE
name = 'Joy'
```

DISK-ORIENTED LOGGING & RECOVERY

The "gold standard" for physical logging & recovery in a disk-oriented DBMS is **ARIES**.

- → **A**lgorithms for **R**ecovery and **I**solation **E**xploiting **S**emantics
- → Invented by IBM Research in the early 1990s.

Relies on STEAL and NO-FORCE buffer pool management policies.

ARIES: A TRANSACTION RECOVERY METHOD SUPPORTING FINE-GRANULARITY LOCKING AND PARTIAL ROLLBACKS USING WRITE-AHEAD LOGGING
ACM Transactions on Database Systems 1992


CHOP HOUSE
RED ROOSTER ALE
HARD CIDER


CHOP HOUSE
RED ROOSTER ALE
INDIANA PALE ALE
HARD CIDER

ARIES – MAIN IDEAS

Write-Ahead Logging:

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

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.

ARIES – RUNTIME LOGGING

For each modification to the database, the DBMS appends a record to the tail of the log.

When a txn commits, its log records are flushed to durable storage.

ARIES – RUNTIME CHECKPOINTS

Use fuzzy checkpoints to allow txns to keep on running while writing checkpoint.

→ The checkpoint may contain updates from txns that have not committed and may abort later on.

The DBMS records internal system state as of the beginning of the checkpoint.

→ Active Transaction Table (ATT)

→ Dirty Page Table (DPT)

LOG SEQUENCE NUMBERS

Every log record has a globally unique ***log sequence number*** (LSN) that is used to determine the serial order of those records.

The DBMS keeps track of various LSNs in both volatile and non-volatile storage to determine the order of almost **<u>everything</u>** in the system…

LOG SEQUENCE NUMBERS

Each page contains a ***pageLSN*** that represents the LSN of the most recent update to that page.

The DBMS keeps track of the max log record written to disk (***flushedLSN***).

For a page ***i*** to be written, the DBMS must flush log at least to the point where $\textbf{\textit{pageLSN}}_i \leq \textbf{\textit{flushedLSN}}$

LOG SEQUENCE NUMBERS

WAL (Tail)

```
015:<T5 begin>
016:<T5,A,99,88>
017:<T5,B,5,10>
018:<T5 commit>
⋮
```

Non-Volatile Storage

```
001:<T1 begin>
002:<T1, A, 1, 2>
003:<T1 commit>
004:<T2 begin>
005:<T2, A, 2, 3>
006:<T3 begin>
007:<CHECKPOINT>
008:<T2 commit>
009:<T4 begin>
010:<T4, X, 5, 6>
011:<T3, B, 4, 2>
012:<T3 commit>
013:<T4, B, 2, 3>
014:<T4, C, 1, 2>
```

Buffer Pool

pageLSN

A=99	B=5	C=12

flushedLSN

pageLSN

A=99	B=5	C=12

Master Record

LOG SEQUENCE NUMBERS

WAL (Tail)

```
015 <T5 begin>
016 <T5,A,99,88>
017 <T5,B,5,10>
018 <T5 commit>
⋮
```

Non-Volatile Storage

```
001 <T1 begin>
002 <T1, A, 1, 2>
003 <T1 commit>
004 <T2 begin>
005 <T2, A, 2, 3>
006 <T3 begin>
007 <CHECKPOINT>
008 <T2 commit>
009 <T4 begin>
010 <T4, X, 5, 6>
011 <T3, B, 4, 2>
012 <T3 commit>
013 <T4, B, 2, 3>
014 <T4, C, 1, 2>
```

pageLSN

A=99	B=5	C=12

Master Record

Buffer Pool

pageLSN

A=99	B=5	C=12

flushedLSN

LOG SEQUENCE NUMBERS

WAL (Tail)

```
015:<T5 begin>
016:<T5,A,99,88>
017:<T5,B,5,10>
018:<T5 commit>
⋮
```

Non-Volatile Storage

```
001:<T1 begin>
002:<T1, A, 1, 2>
003:<T1 commit>
004:<T2 begin>
005:<T2, A, 2, 3>
006:<T3 begin>
007:<CHECKPOINT>
008:<T2 commit>
009:<T4 begin>
010:<T4, X, 5, 6>
011:<T3, B, 4, 2>
012:<T3 commit>
013:<T4, B, 2, 3>
014:<T4, C, 1, 2>
```

pageLSN

A=99	B=5	C=12

Master Record

Buffer Pool

pageLSN

A=99	B=5	C=12

flushedLSN

LOG SEQUENCE NUMBERS

WAL (Tail)

```
015:<T5 begin>
016:<T5,A,99,88>
017:<T5,B,5,10>
018:<T5 commit>
⋮
```

Non-Volatile Storage

```
001:<T1 begin>
002:<T1, A, 1, 2>
003:<T1 commit>
004:<T2 begin>
005:<T2, A, 2, 3>
006:<T3 begin>
007:<CHECKPOINT>
008:<T2 commit>
009:<T4 begin>
010:<T4, X, 5, 6>
011:<T3, B, 4, 2>
012:<T3 commit>
013:<T4, B, 2, 3>
014:<T4, C, 1, 2>
```

pageLSN

A=99	B=5	C=12

Master Record

Buffer Pool

pageLSN

A=99	B=5	C=12

flushedLSN

LOG SEQUENCE NUMBERS

WAL (Tail)

015:<T5 begin>
016:<T5,A,99,88>
017:<T5,B,5,10>
018:<T5 commit>
⋮

Buffer Pool

pageLSN

A=99	B=5	C=12

flushedLSN

Non-Volatile Storage

001:<T1 begin>
002:<T1, A, 1, 2>
003:<T1 commit>
004:<T2 begin>
005:<T2, A, 2, 3>
006:<T3 begin>
007:<CHECKPOINT>
008:<T2 commit>
009:<T4 begin>
010:<T4, X, 5, 6>
011:<T3, B, 4, 2>
012:<T3 commit>
013:<T4, B, 2, 3>
014:<T4, C, 1, 2>

pageLSN

A=99	B=5	C=12

Master Record

LOG SEQUENCE NUMBERS

WAL (Tail)

```
015:<T5 begin>
016:<T5,A,99,88>
017:<T5,B,5,10>
018:<T5 commit>
  ⋮
```

Non-Volatile Storage

```
001:<T1 begin>
002:<T1, A, 1, 2>
003:<T1 commit>
004:<T2 begin>
005:<T2, A, 2, 3>
006:<T3 begin>
007:<CHECKPOINT>
008:<T2 commit>
009:<T4 begin>
010:<T4, X, 5, 6>
011:<T3, B, 4, 2>
012:<T3 commit>
013:<T4, B, 2, 3>
014:<T4, C, 1, 2>
```

pageLSN

A=99	B=5	C=12

Master Record

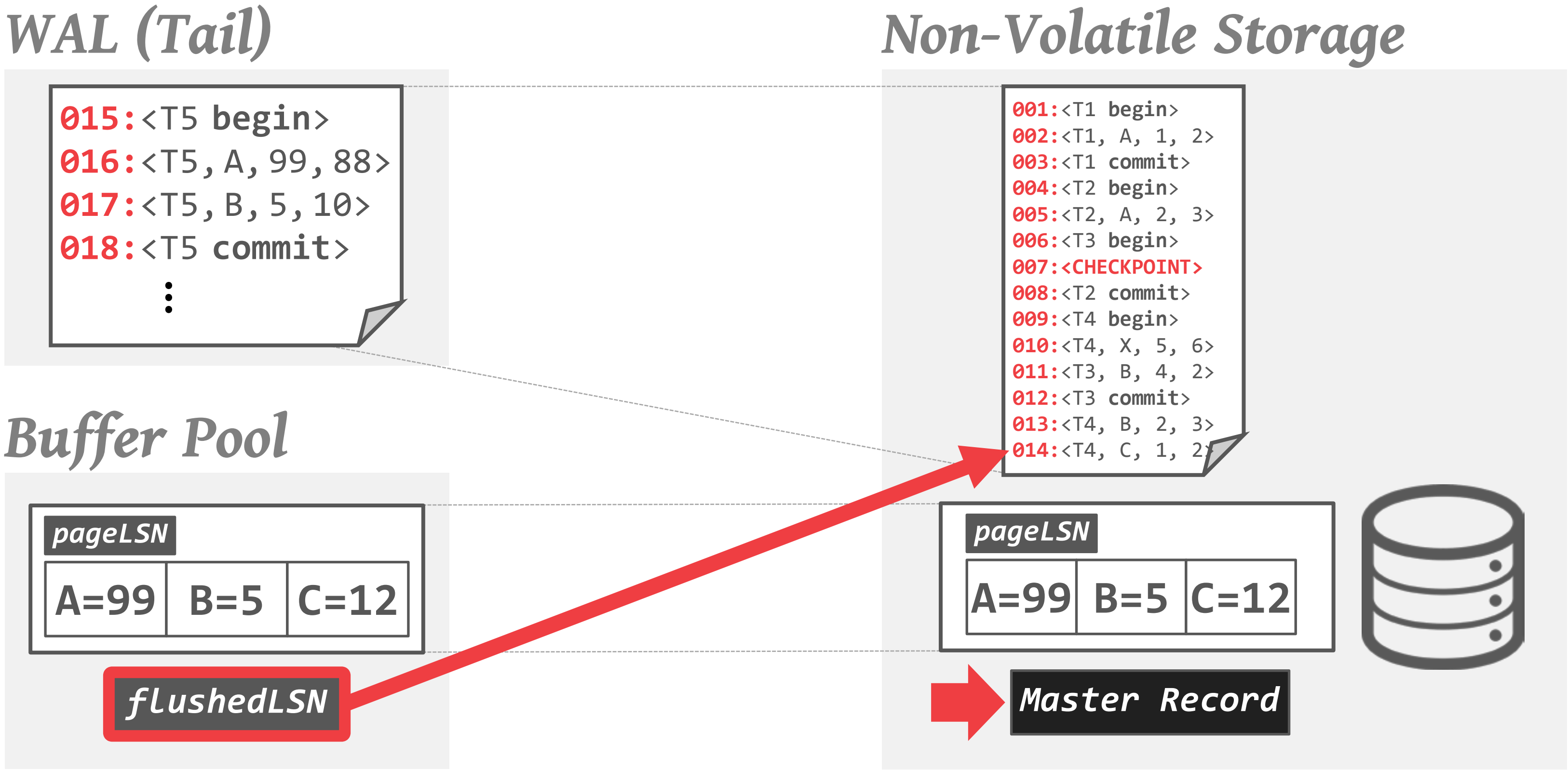
Buffer Pool

pageLSN

A=99	B=5	C=12

flushedLSN

LOG SEQUENCE NUMBERS



LOG SEQUENCE NUMBERS

WAL (Tail)

```
015:<T5 begin>
016:<T5,A,99,88>
017:<T5,B,5,10>
018:<T5 commit>
⋮
```

Buffer Pool

pageLSN

A=99	B=5	C=12

flushedLSN

Non-Volatile Storage

```
001:<T1 begin>
002:<T1, A, 1, 2>
003:<T1 commit>
004:<T2 begin>
005:<T2, A, 2, 3>
006:<T3 begin>
007:<CHECKPOINT>
008:<T2 commit>
009:<T4 begin>
010:<T4, X, 5, 6>
011:<T3, B, 4, 2>
012:<T3 commit>
013:<T4, B, 2, 3>
014:<T4, C, 1, 2>
```

pageLSN

A=99	B=5	C=12

Master Record

LOG SEQUENCE NUMBERS

WAL (Tail)

```
015:<T5 begin>
016:<T5,A,99,88>
017:<T5,B,5,10>
018:<T5 commit>
⋮
```

Non-Volatile Storage

```
001:<T1 begin>
002:<T1, A, 1, 2>
003:<T1 commit>
004:<T2 begin>
005:<T2, A, 2, 3>
006:<T3 begin>
007:<CHECKPOINT>
008:<T2 commit>
009:<T4 begin>
010:<T4, X, 5, 6>
011:<T3, B, 4, 2>
012:<T3 commit>
013:<T4, B, 2, 3>
014:<T4, C, 1, 2>
```

pageLSN

A=99	B=5	C=12

Master Record

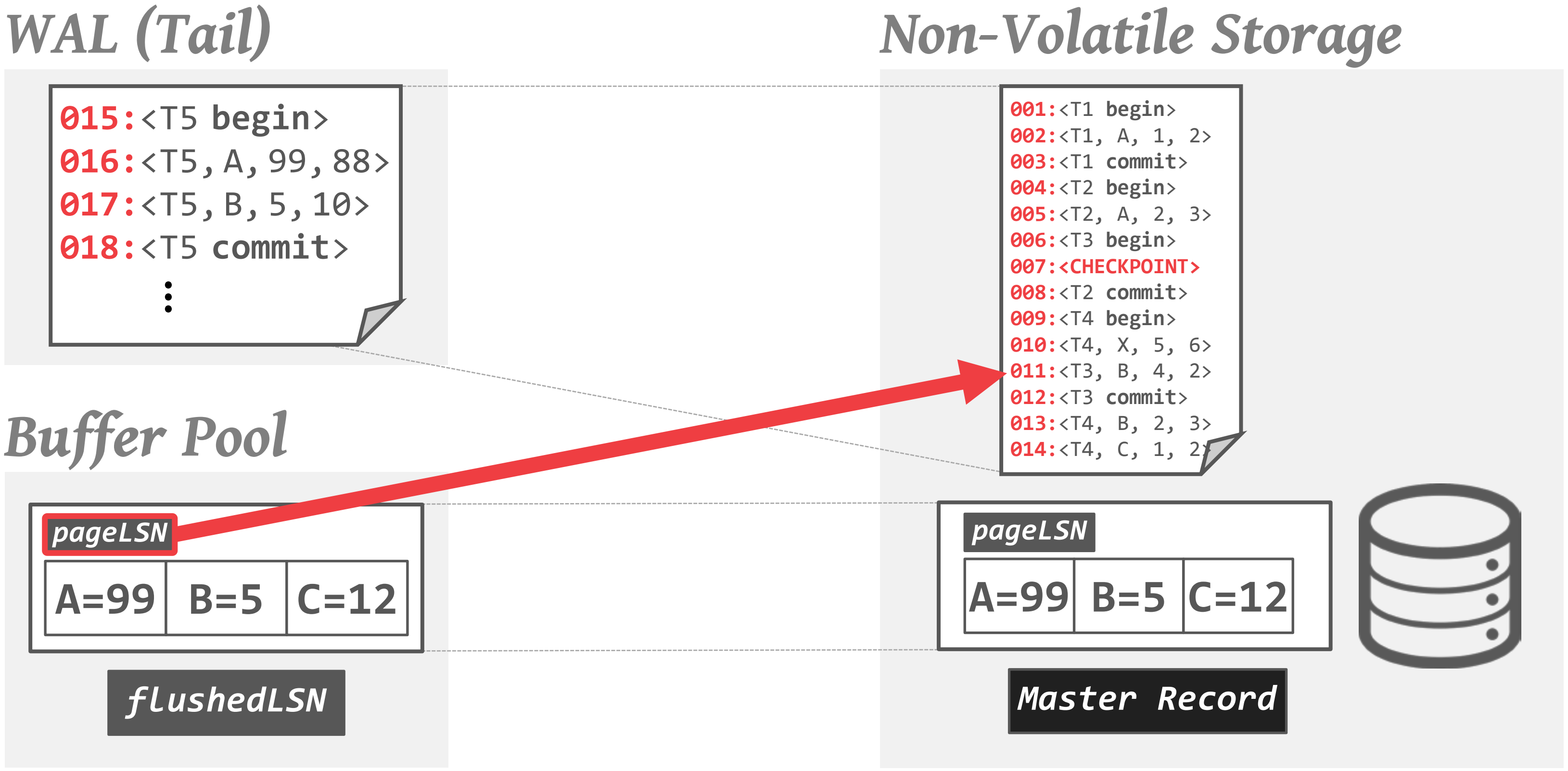
Buffer Pool

pageLSN

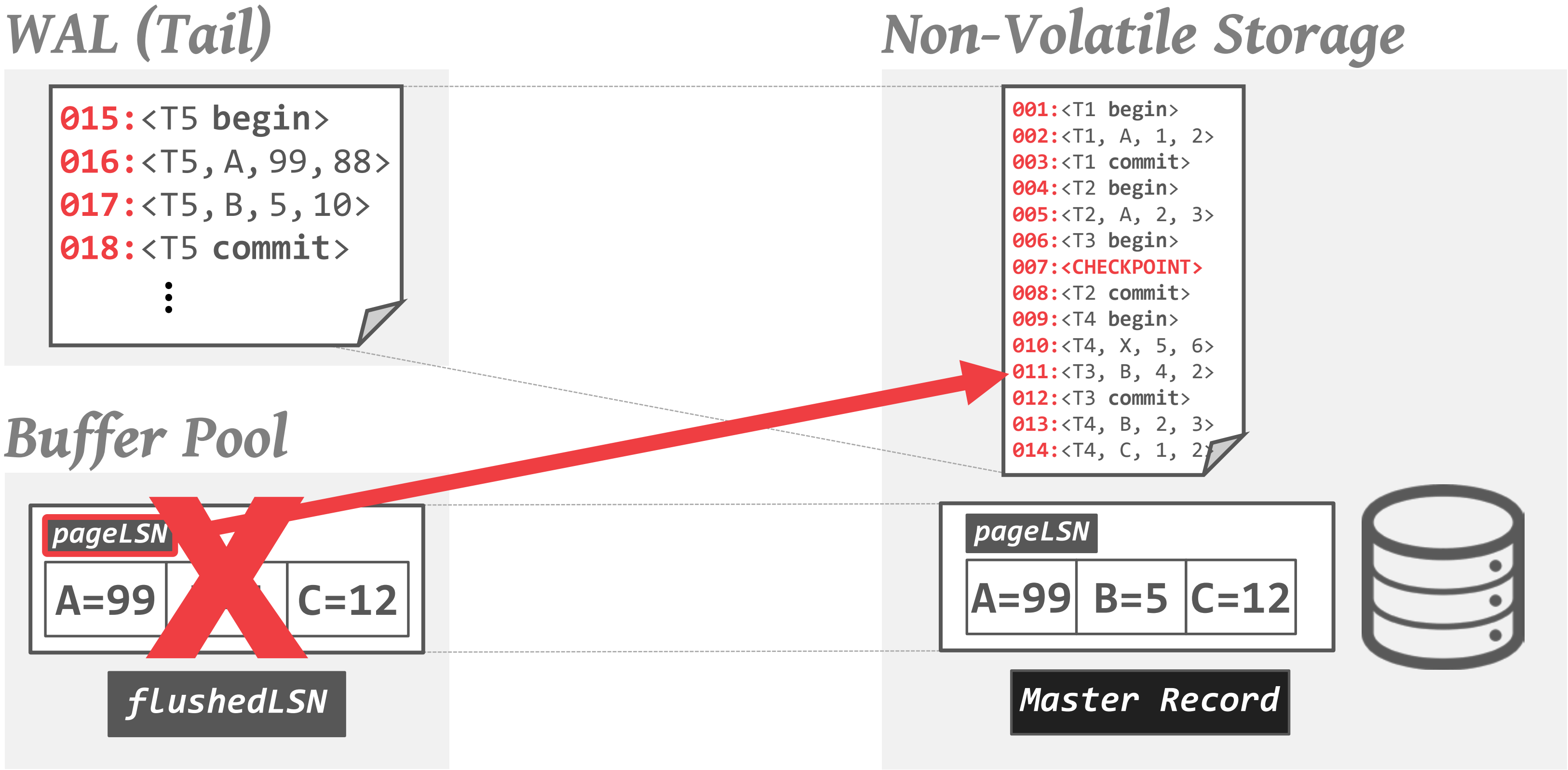
A=99	B=5	C=12

flushedLSN

LOG SEQUENCE NUMBERS

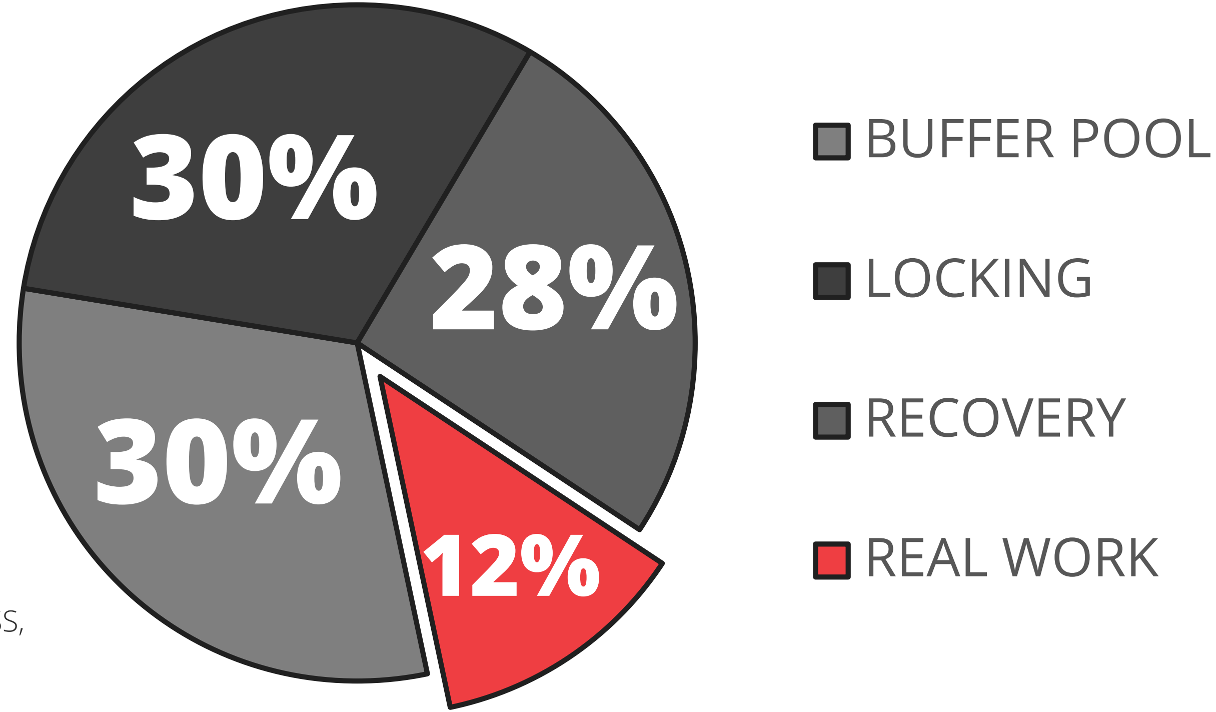


LOG SEQUENCE NUMBERS



DISK-ORIENTED DBMS OVERHEAD

Measured CPU Cycles



OLTP THROUGH THE LOOKING GLASS, AND WHAT WE FOUND THERE
SIGMOD, pp. 981-992, 2008.

ARIES – RECOVERY PHASES

Phase #1: Analysis

→ Read the WAL 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.

→ Log redo steps in case of crash during recovery.

Phase #3: Undo

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

OBSERVATION

Often the slowest part of the txn is waiting for the DBMS to flush the log records to disk.

Have to wait until the records are safely written before the DBMS can return the acknowledgement to the client.

GROUP COMMIT

Batch together log records from multiple txns and flush them together with a single **`fsync`**.

→ Logs are flushed either after a timeout or when the buffer gets full.

→ Originally developed in IBM IMS FastPath in the 1980s

This amortizes the cost of I/O over several txns.

EARLY LOCK RELEASE

A txn's locks can be released before its commit record is written to disk as long as it does not return results to the client before becoming durable.

Other txns that read data updated by a **pre-committed** txn become dependent on it and also have to wait for their predecessor's log records to reach disk.

IN-MEMORY DATABASE RECOVERY

Recovery is slightly easier because the DBMS does not have to worry about tracking dirty pages in case of a crash during recovery.

An in-memory DBMS also does not need to store undo records.

But the DBMS is still stymied by the slow sync time of non-volatile storage

OBSERVATION

The early papers (1980s) on recovery for in-memory DBMSs assume that there is non-volatile memory.

This hardware is still not widely available so we want to use existing SSD/HDDs.

A RECOVERY ALGORITHM FOR A HIGH-PERFORMANCE MEMORY-RESIDENT DATABASE SYSTEM
SIGMOD 1987

SILO – LOGGING AND RECOVERY

SiloR uses the epoch-based OCC that we discussed previously.

It achieves high performance by parallelizing all aspects of logging, checkpointing, and recovery.

Again, Eddie Kohler is unstoppable.

FAST DATABASES WITH FAST DURABILITY AND RECOVERY THROUGH MULTICORE PARALLELISM
OSDI 2014

SILOR – LOGGING PROTOCOL

The DBMS assumes that there is one storage device per CPU socket.
→ Assigns one logger thread per device.
→ Worker threads are grouped per CPU socket.

As the worker executes a txn, it creates new log records that contain the values that were written to the database (i.e., REDO).

SILOR – LOGGING PROTOCOL

Each logger thread maintains a pool of log buffers that are given to its worker threads.

When a worker's buffer is full, it gives it back to the logger thread to flush to disk and attempts to acquire a new one.

→ If there are no available buffers, then it stalls.

SILOR – LOG FILES

The logger threads write buffers out to files

→ After 100 epochs, it creates a new file.
→ The old file is renamed with a marker indicating the max epoch of records that it contains.

Log record format:

→ Id of the txn that modified the record (TID).
→ A set of value log triplets (Table, Key, Value).
→ The value can be a list of attribute + value pairs.

SILOR – LOG FILES

```
root@magneto:/var/lib/mysql# ls -lah
total 5.5G
drwxr-x---  5 mysql mysql 4.0K Dec 22 07:56 .
drwxr-xr-x 69 root  root  4.0K Dec 16 20:22 ..
-rw-rw----  1 mysql mysql   56 Aug 16  2015 auto.cnf
-rw-------  1 mysql mysql 1.7K Dec 16 20:22 ca-key.pem
-rw-r--r--  1 mysql mysql 1.1K Dec 16 20:22 ca.pem
-rw-r--r--  1 mysql mysql 1.1K Dec 16 20:22 client-cert.pem
-rw-------  1 mysql mysql 1.7K Dec 16 20:22 client-key.pem
-rw-r-----  1 mysql mysql 1.1K Dec 16 20:29 ib_buffer_pool
-rw-rw----  1 mysql mysql  76M Dec 21 08:38 ibdata1
-rw-r-----  1 mysql mysql 500M Dec 22 07:00 ib_logfile0
-rw-r-----  1 mysql mysql 500M Dec 21 08:39 ib_logfile1
-rw-rw----  1 mysql mysql 4.4G Dec 21 08:38 magneto.log
-rw-rw----  1 mysql mysql  55M Dec 21 08:38 magneto-slow.log
drwxr-x---  2 mysql mysql 4.0K Dec 16 20:27 mysql
-rw-r--r--  1 root  root     6 Dec 16 20:27 mysql_upgrade_info
```

SILOR – LOG FILES

```
root@magneto:/var/lib/mysql# ls -lah
total 5.5G
drwxr-x---  5 mysql mysql 4.0K Dec 22 07:56 .
drwxr-xr-x 69 root  root  4.0K Dec 16 20:22 ..
-rw-rw----  1 mysql mysql   56 Aug 16  2015 auto.cnf
-rw-------  1 mysql mysql 1.7K Dec 16 20:22 ca-key.pem
-rw-r--r--  1 mysql mysql 1.1K Dec 16 20:22 ca.pem
-rw-r--r--  1 mysql mysql 1.1K Dec 16 20:22 client-cert.pem
-rw-------  1 mysql mysql 1.7K Dec 16 20:22 client-key.pem
-rw-r-----  1 mysql mysql 1.1K Dec 16 20:29 ib_buffer_pool
-rw-rw----  1 mysql mysql  76M Dec 21 08:38 ibdata1
-rw-r-----  1 mysql mysql 500M Dec 22 07:00 ib_logfile0
-rw-r-----  1 mysql mysql 500M Dec 21 08:39 ib_logfile1
-rw-rw----  1 mysql mysql 4.4G Dec 21 08:38 magneto.log
-rw-rw----  1 mysql mysql  55M Dec 21 08:38 magneto-slow.log
drwxr-x---  2 mysql mysql 4.0K Dec 16 20:27 mysql
-rw-r--r--  1 root  root     6 Dec 16 20:27 mysql_upgrade_info
```

SILOR – LOG FILES

The logger threads write buffers out to files

→ After 100 epochs, it creates a new file.

→ The old file is renamed with a marker indicating the max epoch of records that it contains.

Log record format:

→ Id of the txn that modified the record (TID).

→ A set of value log triplets (Table, Key, Value).

→ The value can be a list of attribute + value pairs.

SILOR – LOG FILES

The logger threads write buffers out to files
→ After 100 epochs, it creates a new file.
→ The old file is renamed with a marker indicating the max epoch of records that it contains.

Log record format:
→ Id of the txn that modified the record (TID).
→ A set of value log triplets (Table, Key, Value).
→ The value can be a list of attribute + value pairs.

```
UPDATE people
   SET isLame = true
 WHERE name IN (‘Joy’,‘Andy’)
```

SILOR – LOG FILES

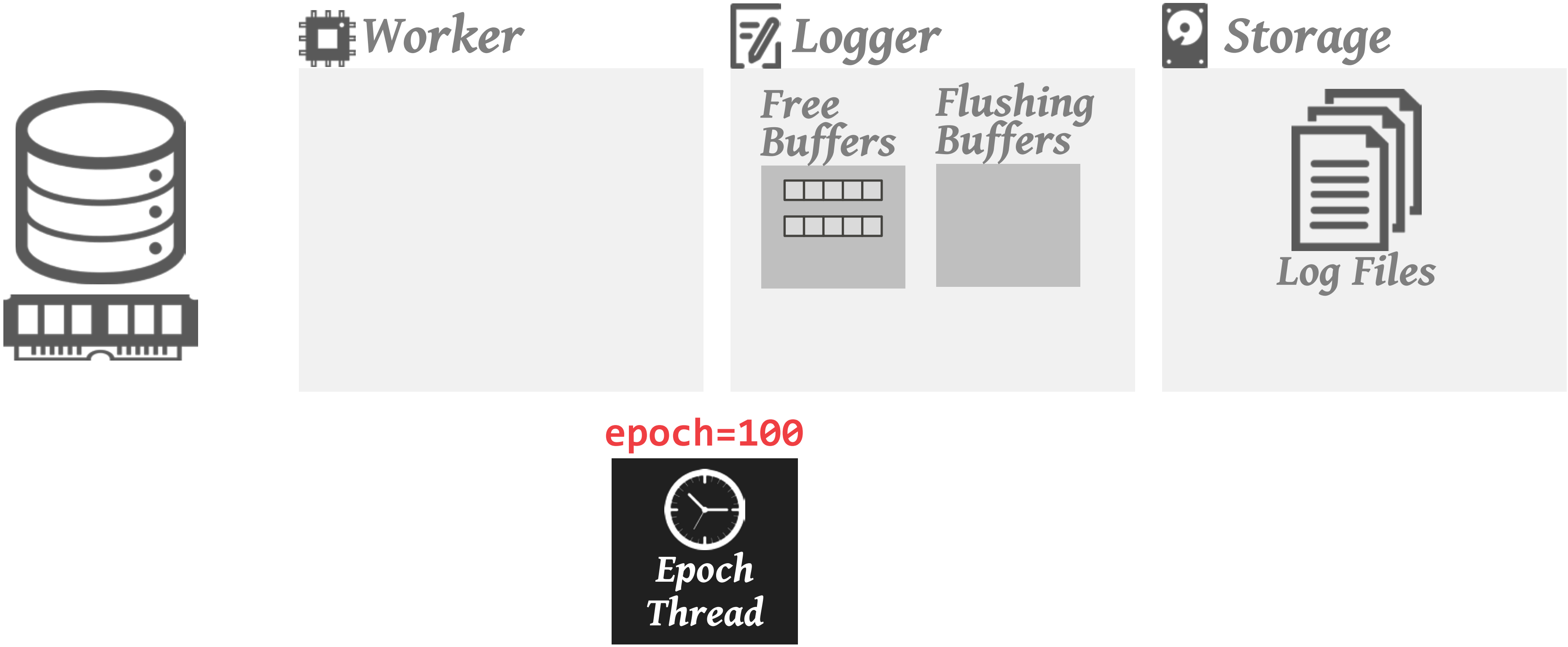
The logger threads write buffers out to files
- → After 100 epochs, it creates a new file.
- → The old file is renamed with a marker indicating the max epoch of records that it contains.

Log record format:
- → Id of the txn that modified the record (TID).
- → A set of value log triplets (Table, Key, Value).
- → The value can be a list of attribute + value pairs.

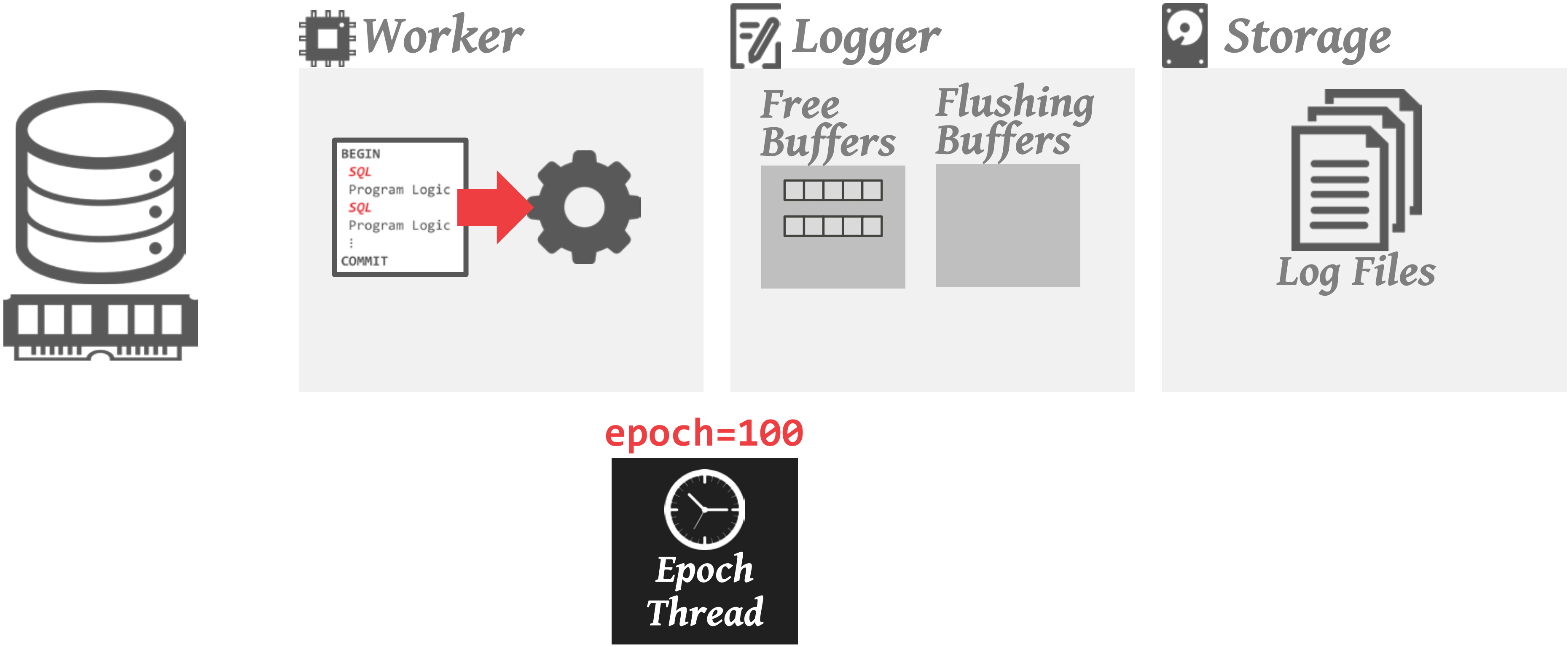
```
UPDATE people
   SET isLame = true
 WHERE name IN ('Joy','Andy')
```

```
Txn#1001
[people, 888, (isLame→true)]
[people, 999, (isLame→true)]
```

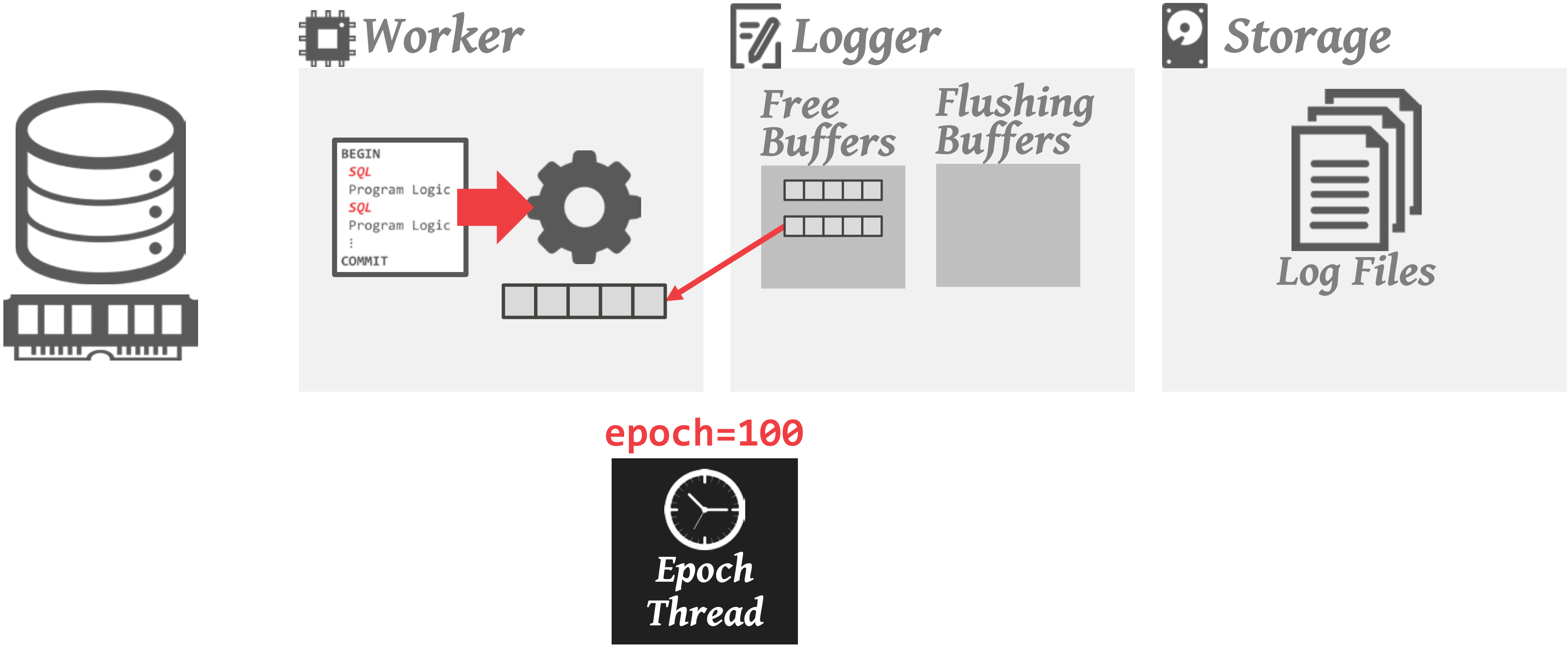
SILOR – ARCHITECTURE



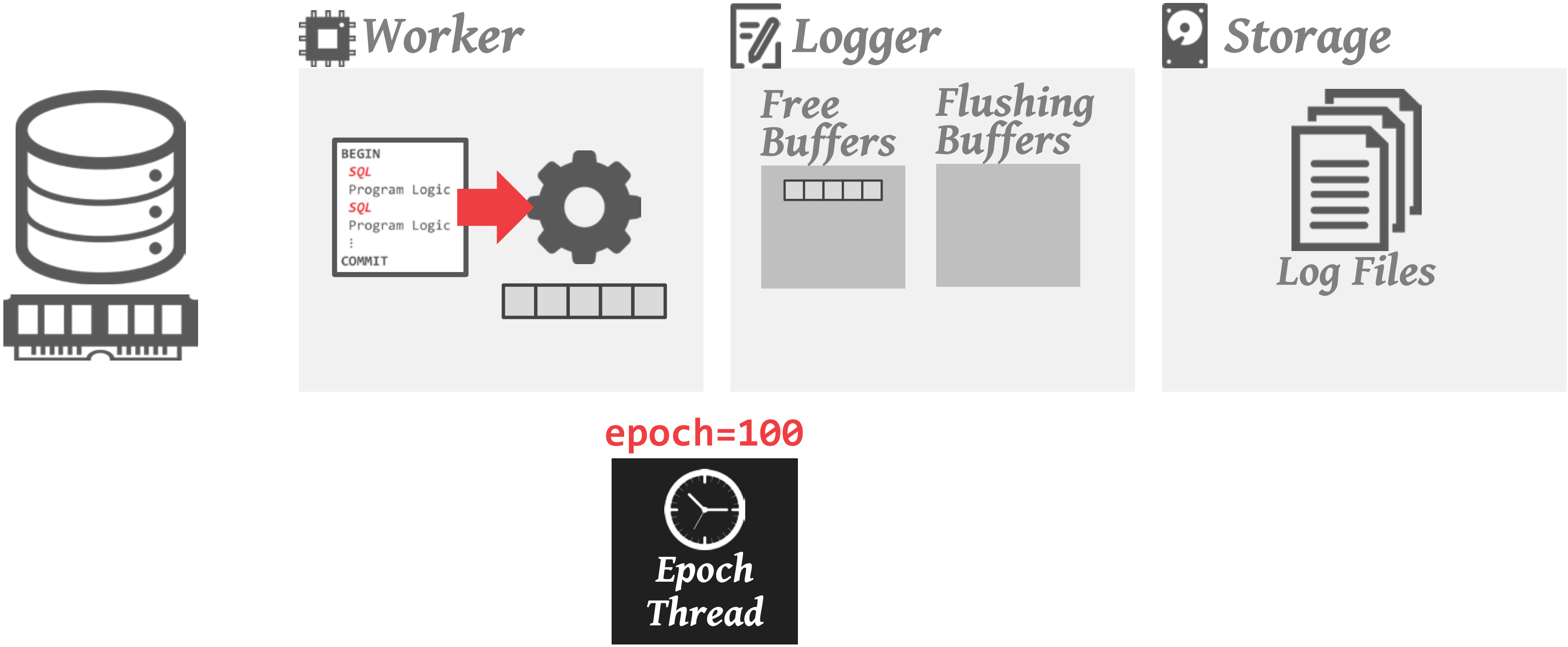
SILOR – ARCHITECTURE



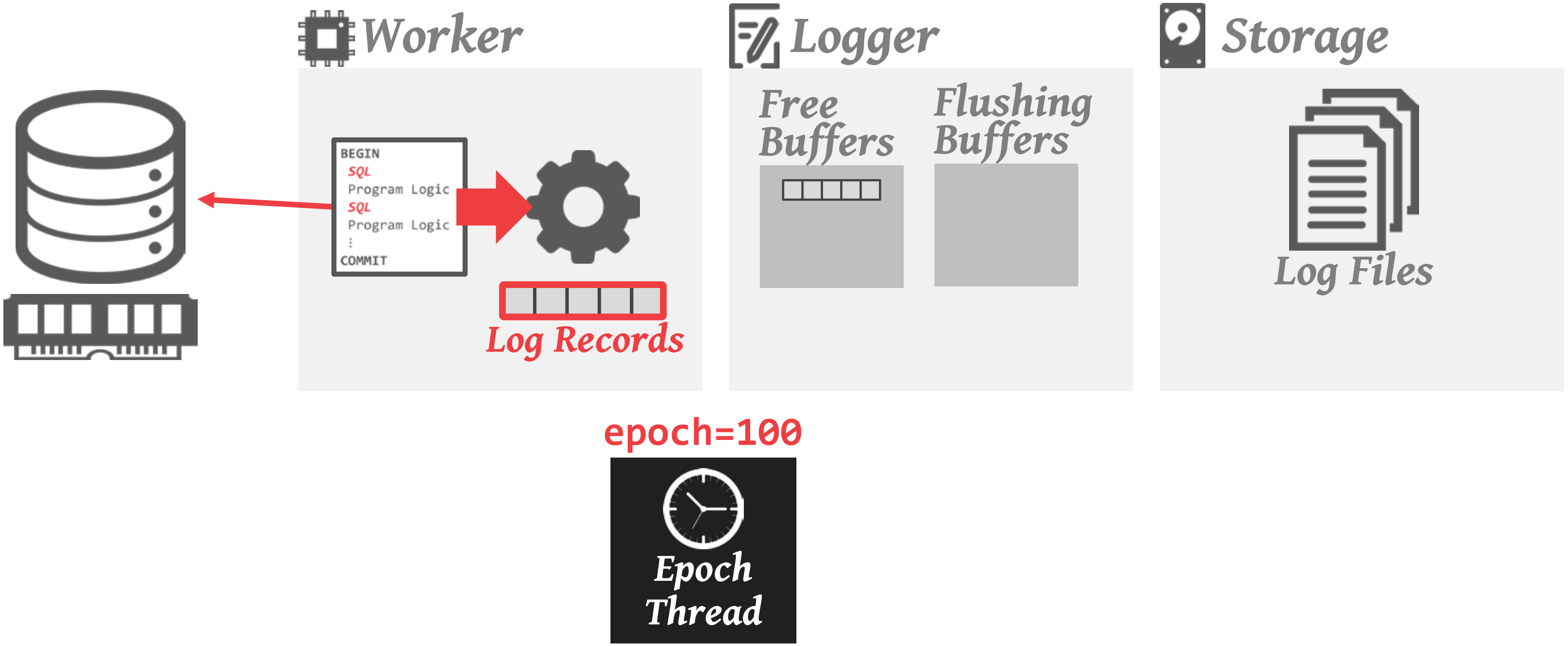
SILOR – ARCHITECTURE



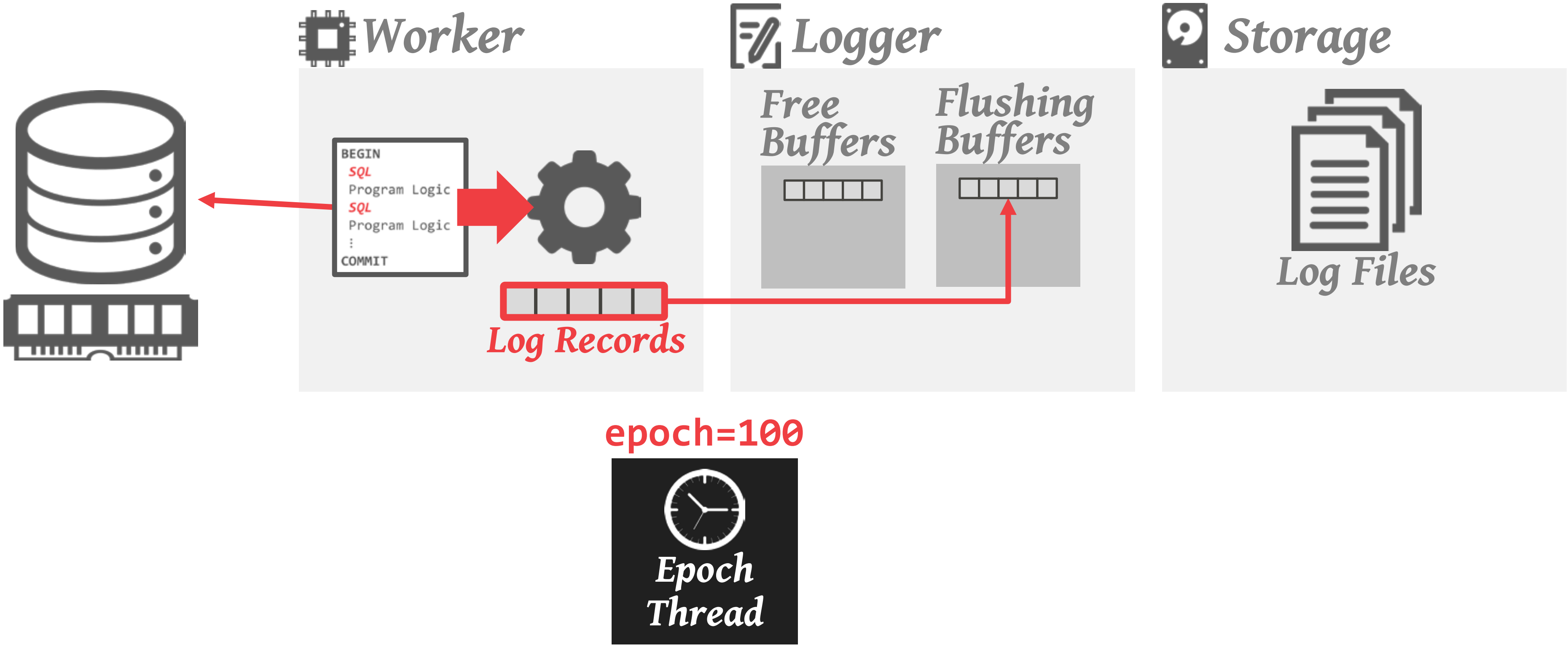
SILOR – ARCHITECTURE



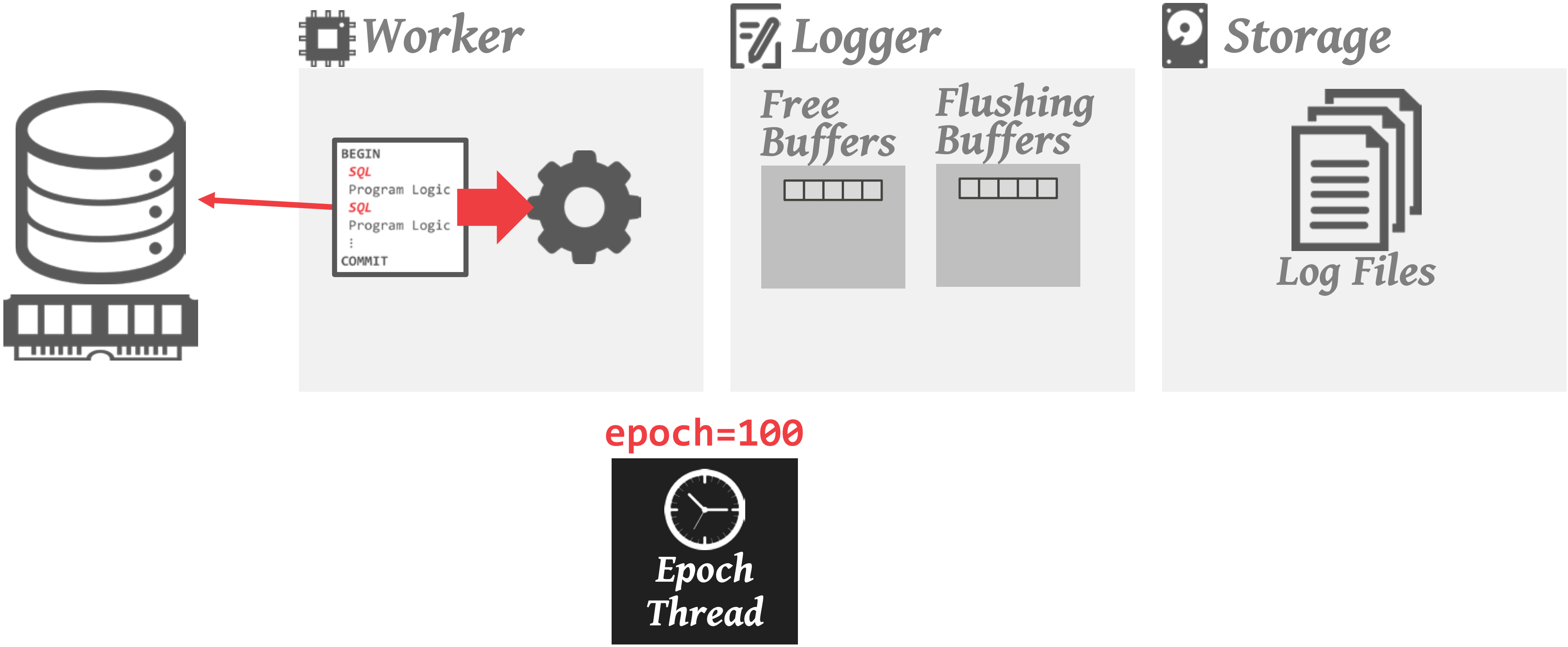
SILOR – ARCHITECTURE



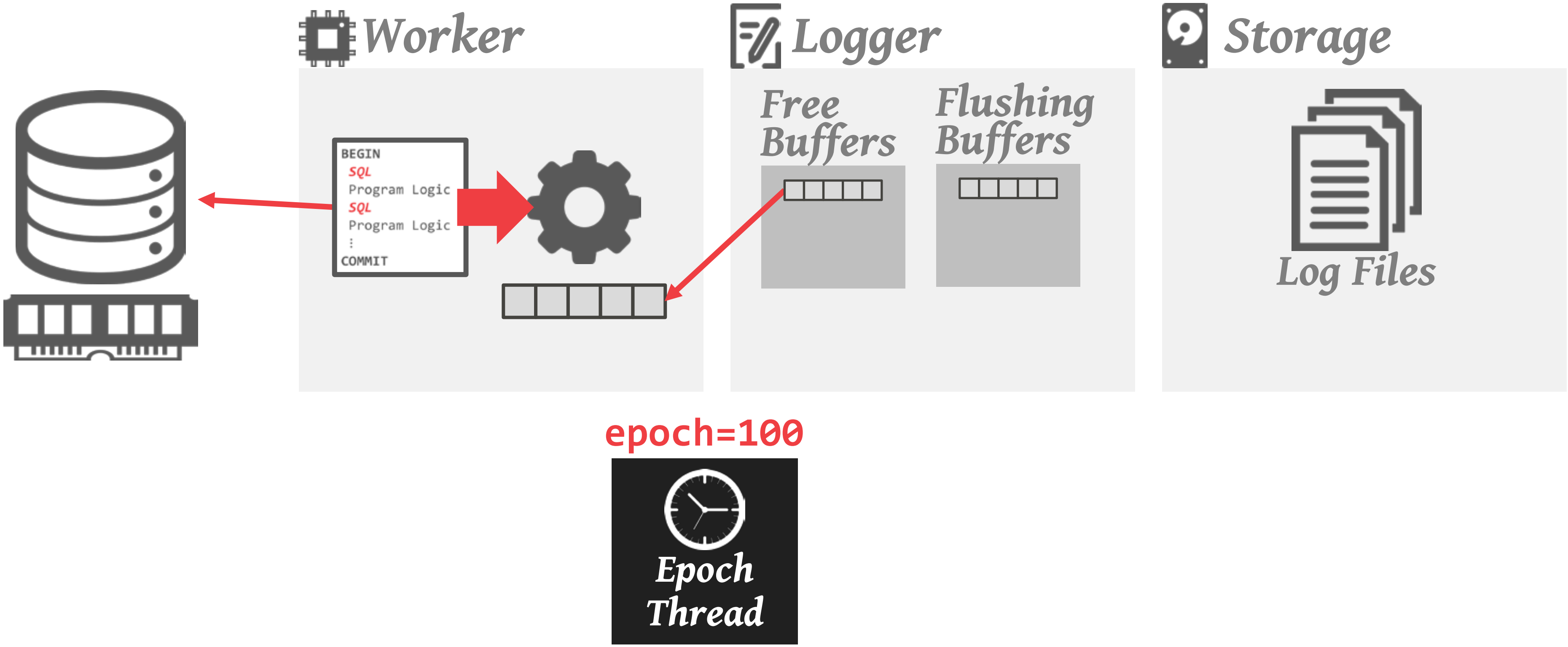
SILOR – ARCHITECTURE



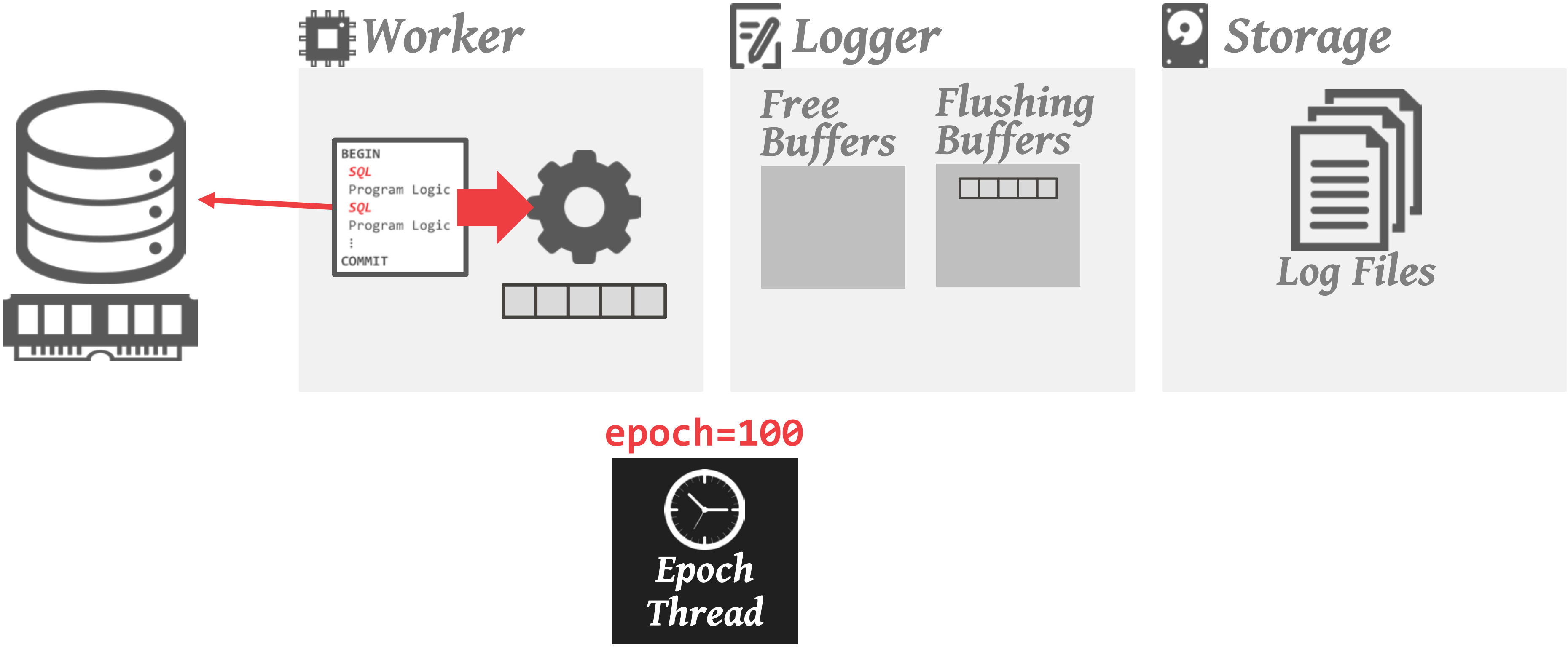
SILOR – ARCHITECTURE



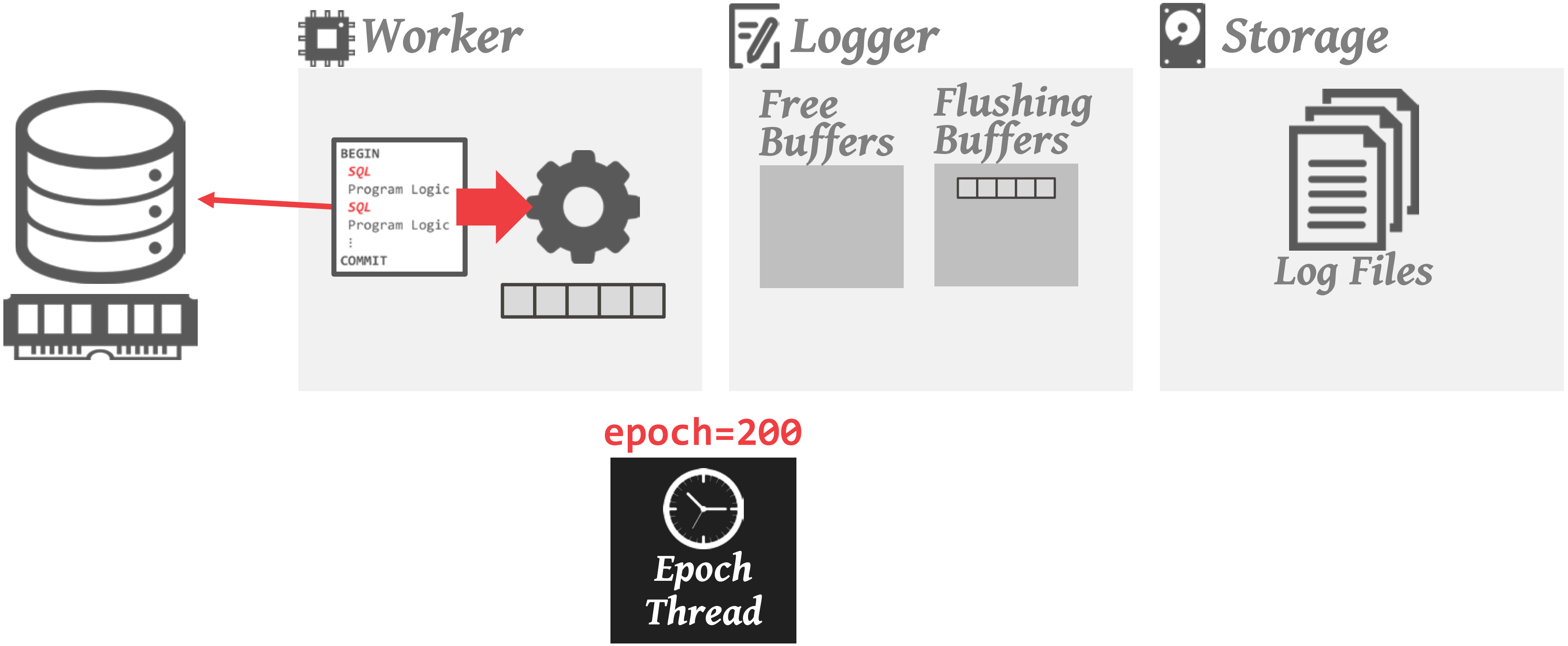
SILOR – ARCHITECTURE



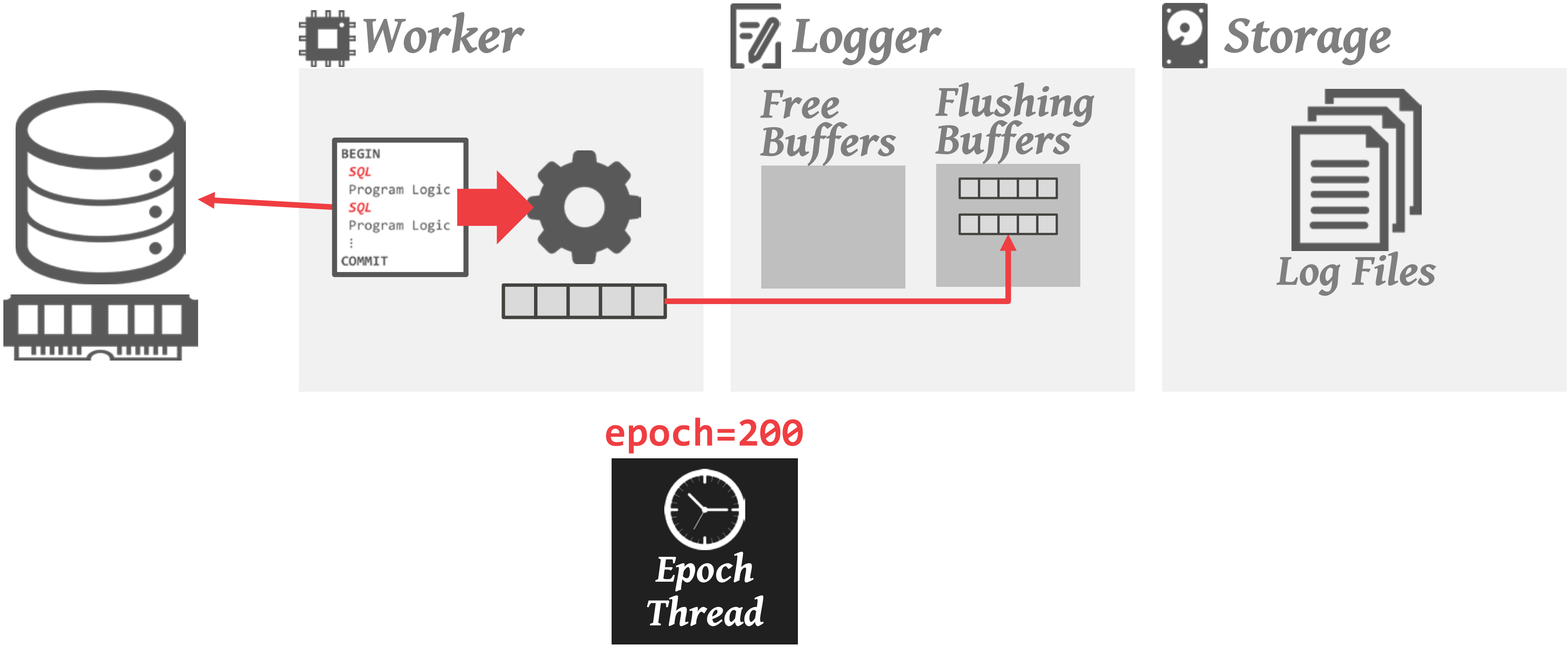
SILOR – ARCHITECTURE



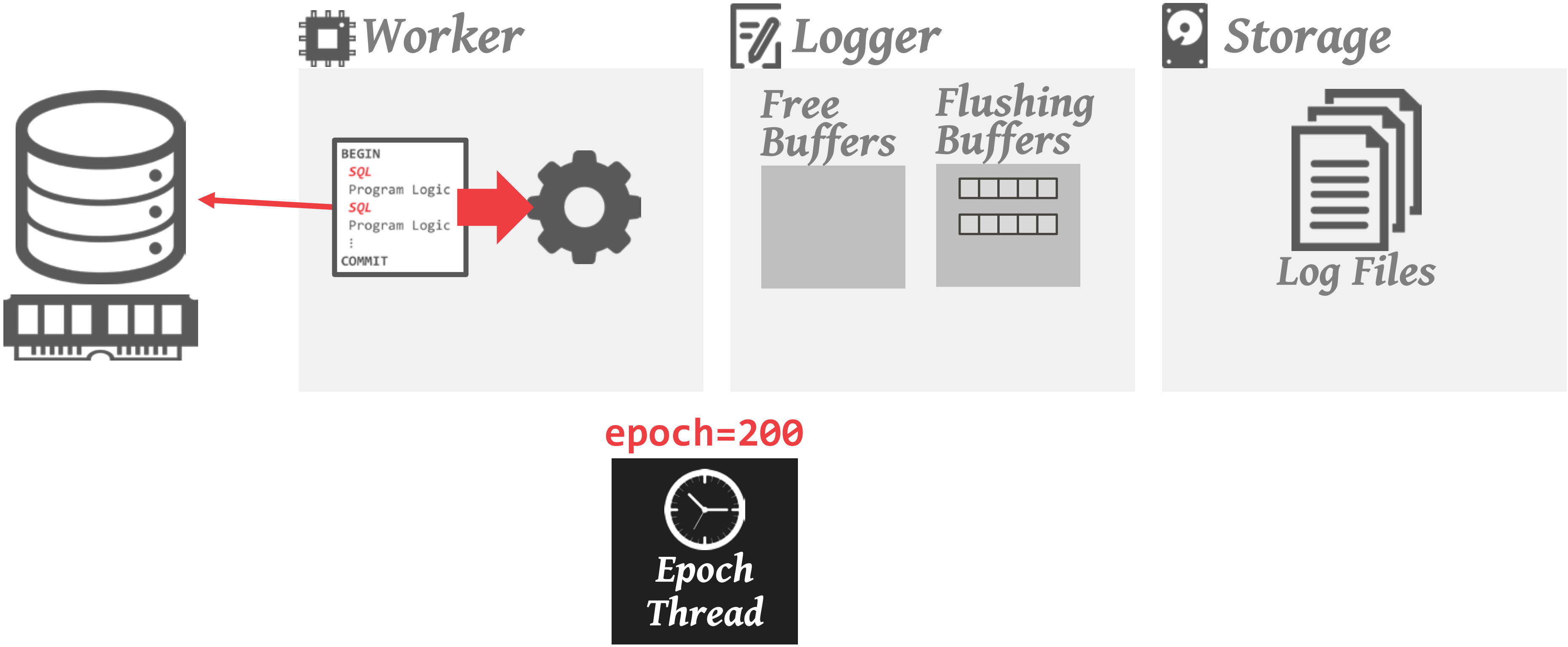
SILOR – ARCHITECTURE



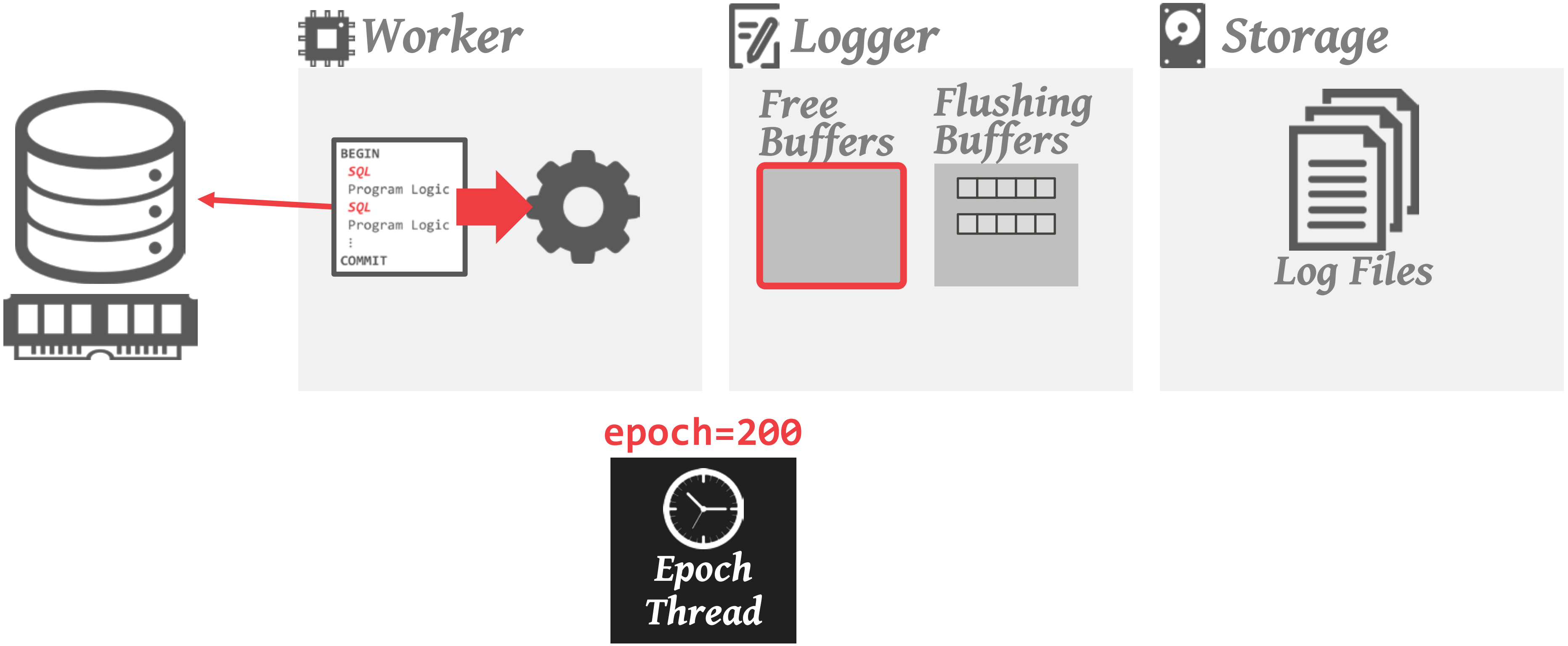
SILOR – ARCHITECTURE



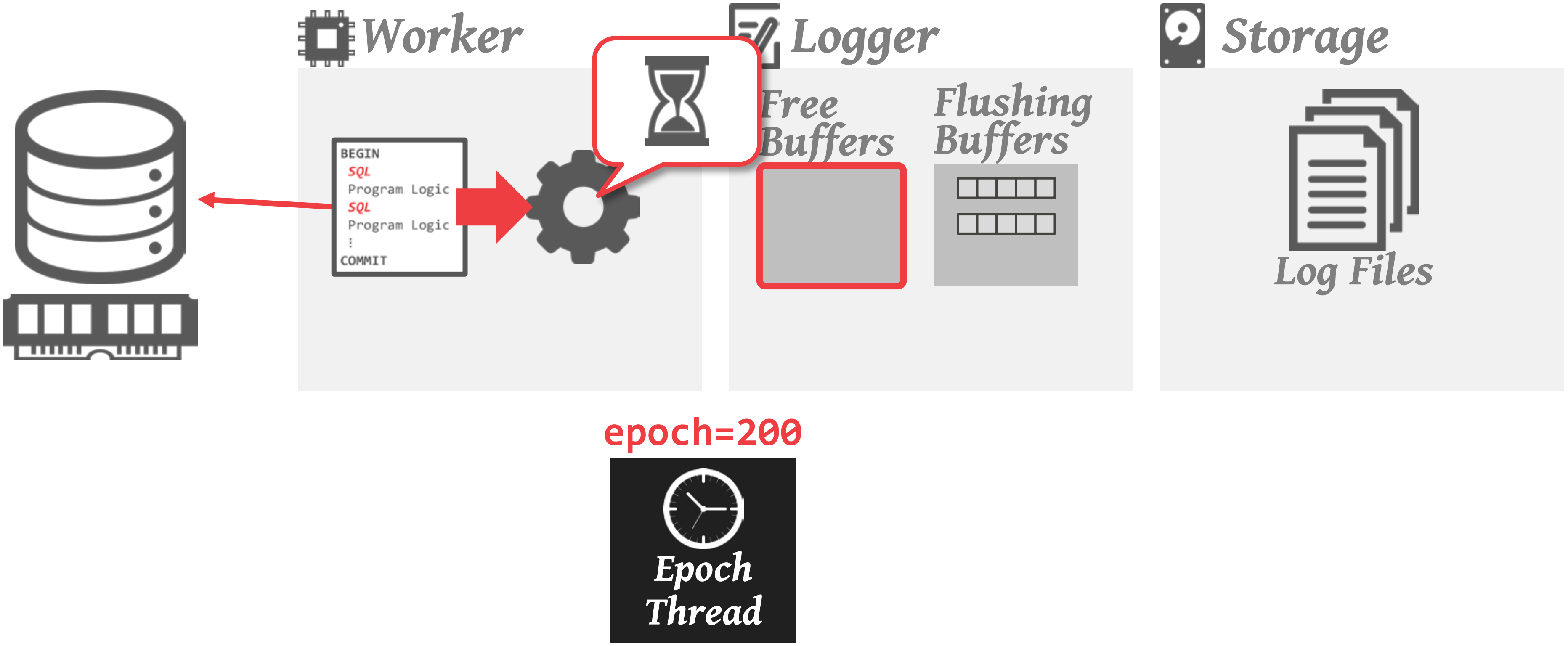
SILOR – ARCHITECTURE



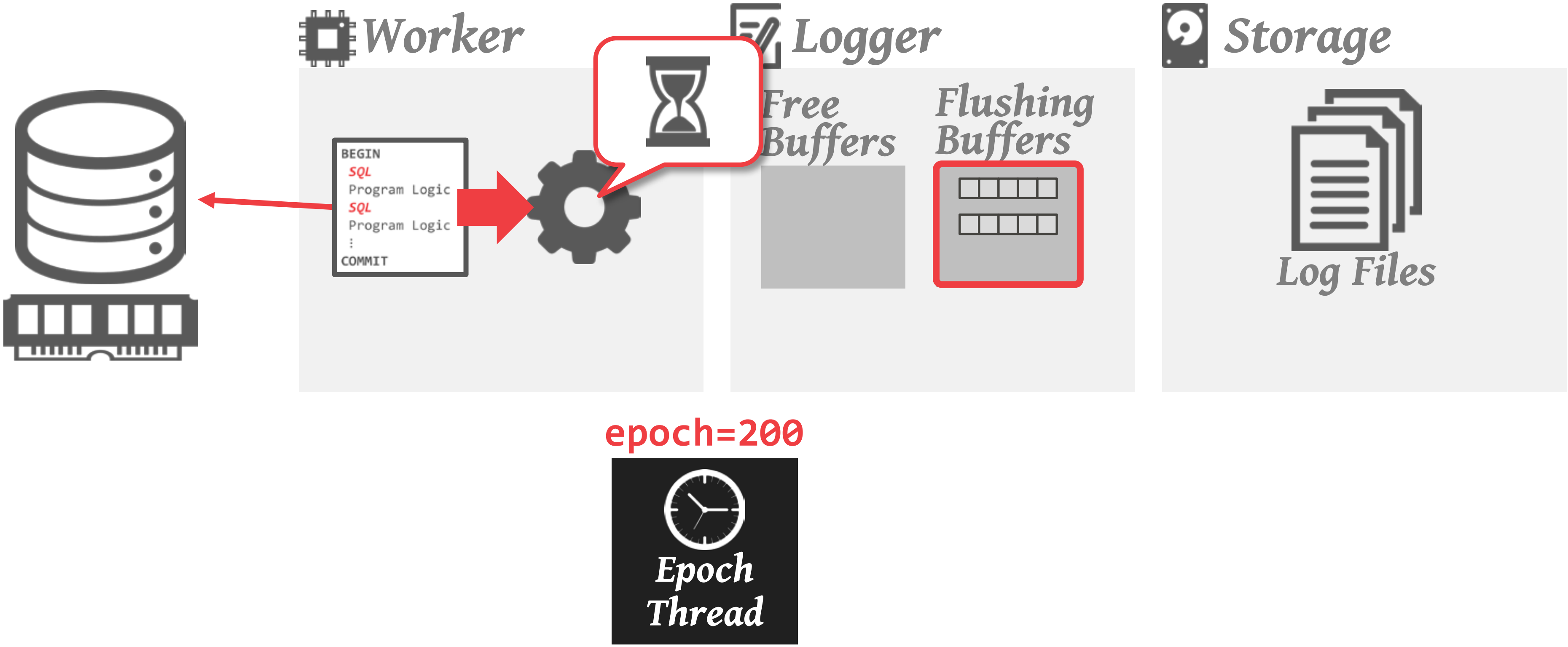
SILOR – ARCHITECTURE



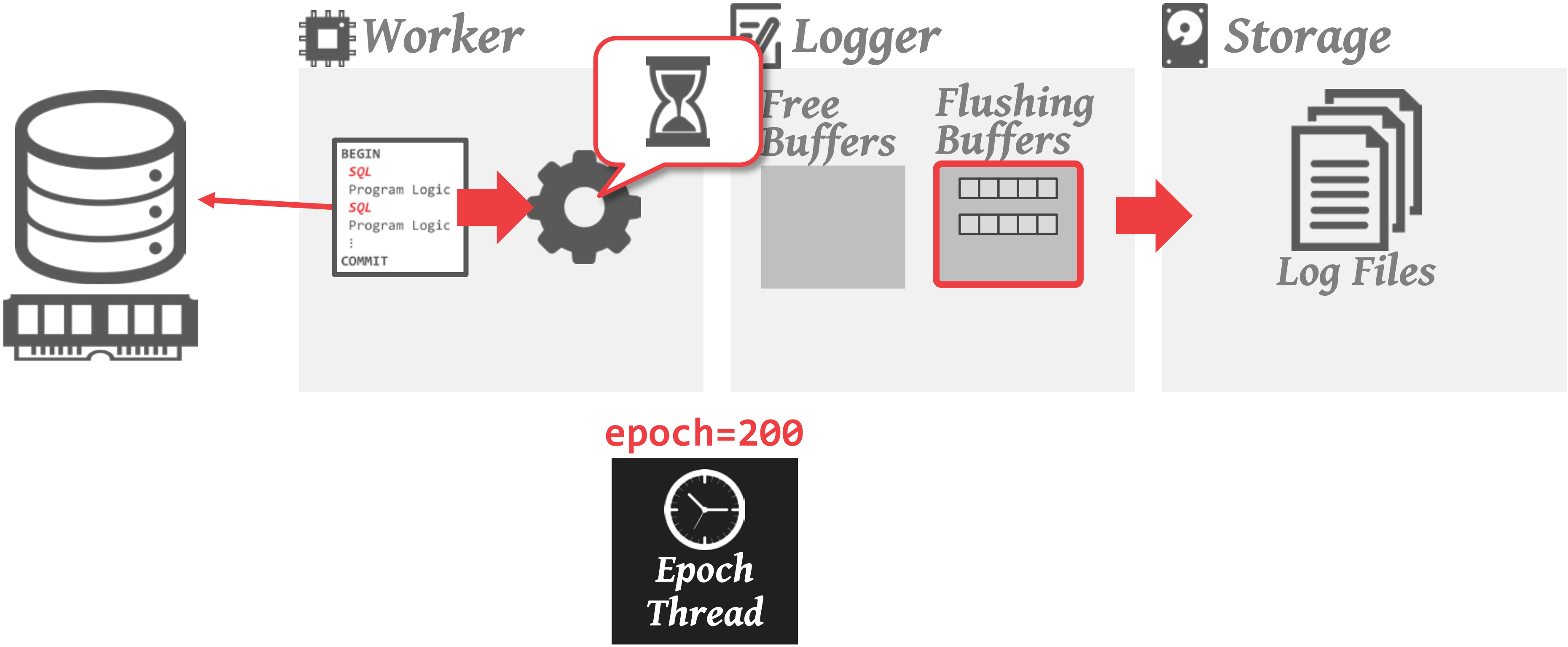
SILOR – ARCHITECTURE



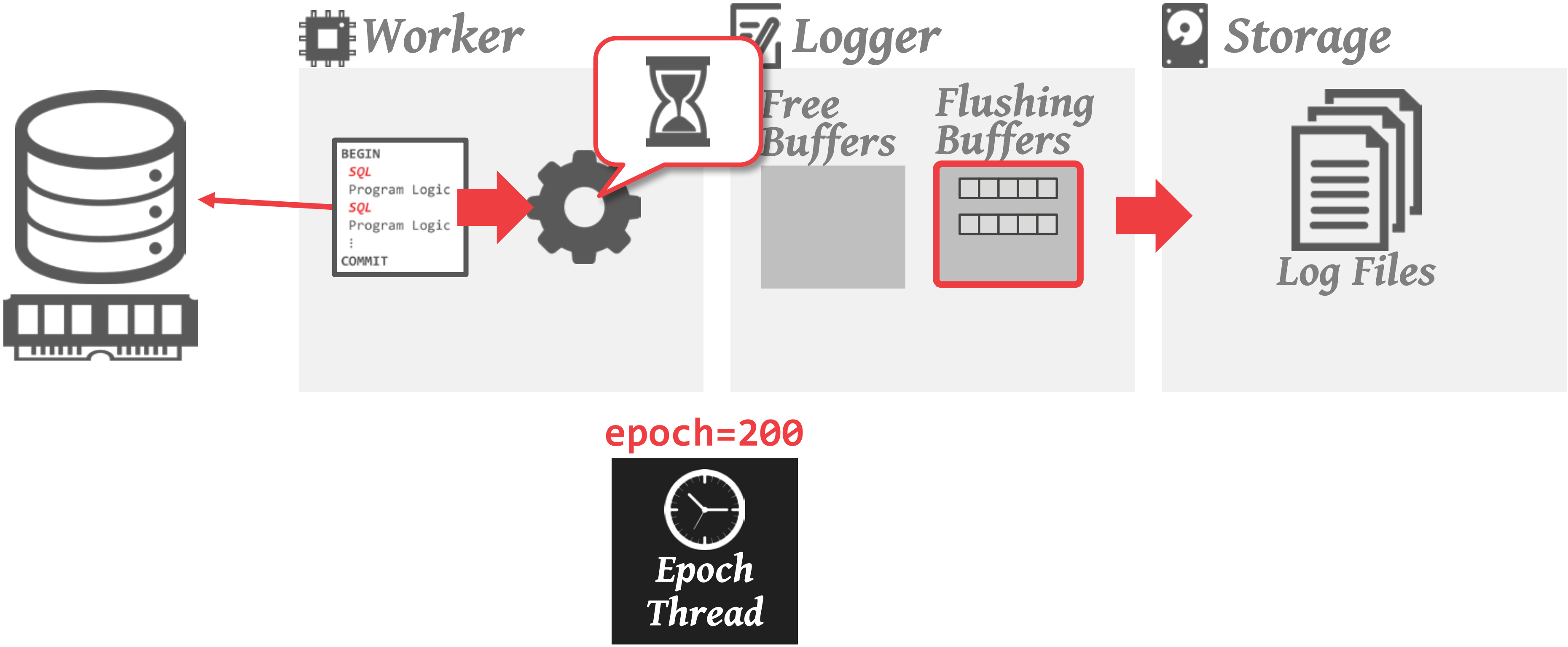
SILOR – ARCHITECTURE



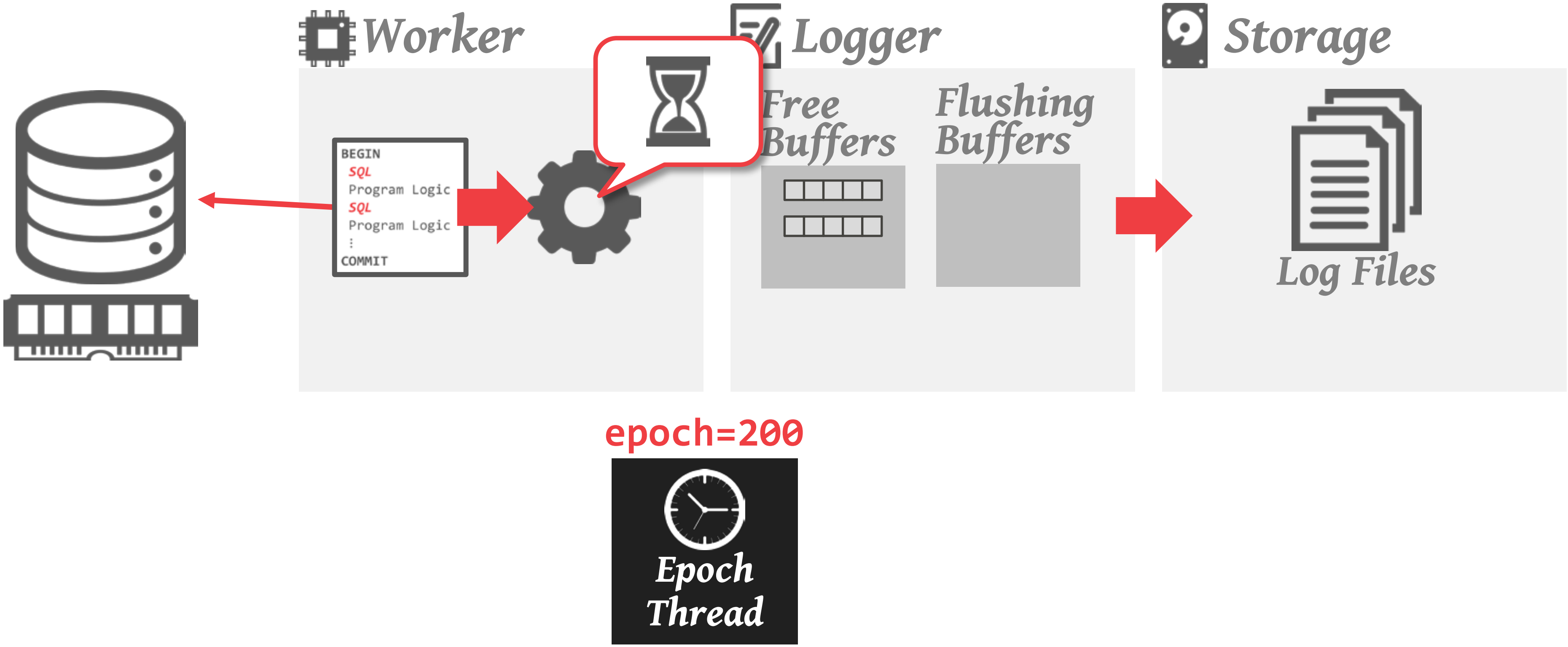
SILOR – ARCHITECTURE



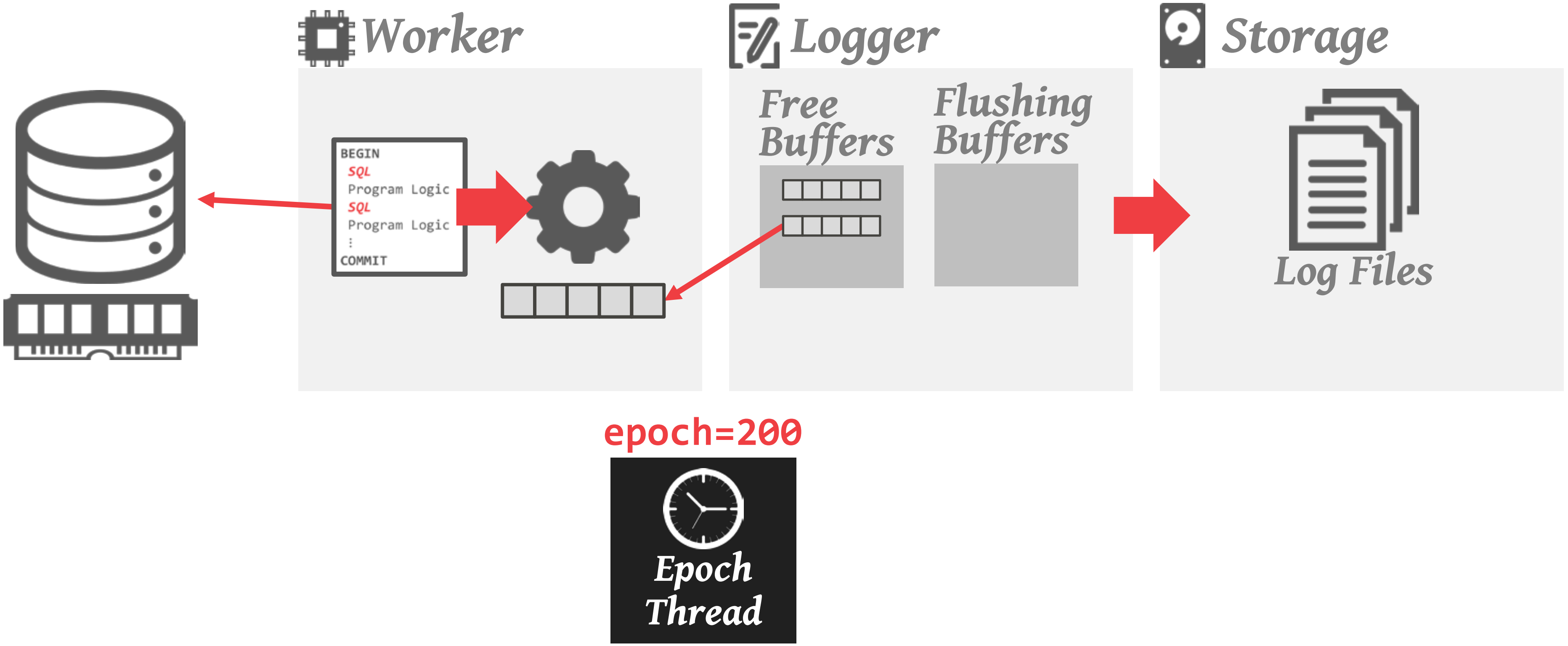
SILOR – ARCHITECTURE



SILOR – ARCHITECTURE



SILOR – ARCHITECTURE

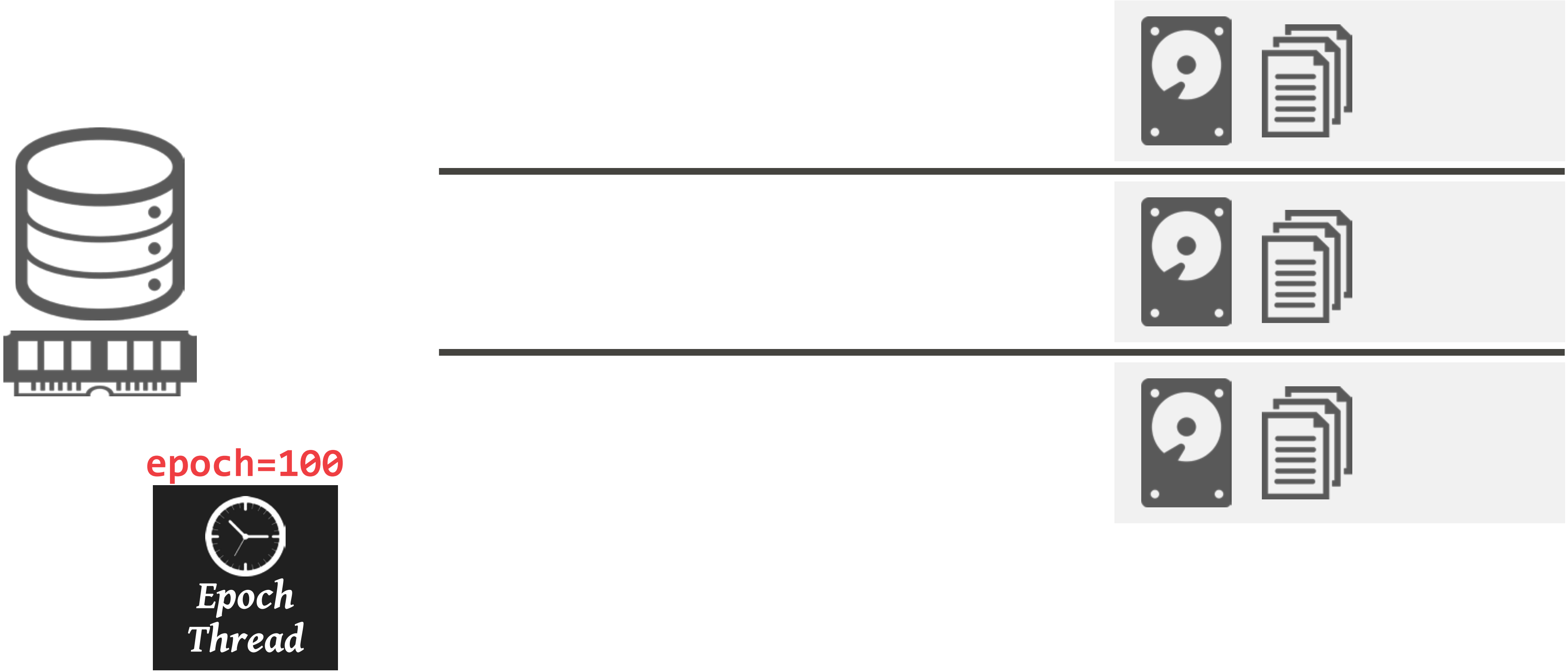


SILOR – PERSISTENT EPOCH

A special logger thread keeps track of the current persistent epoch (***pepoch***)
→ Special log file that maintains the highest epoch that is durable across all loggers.

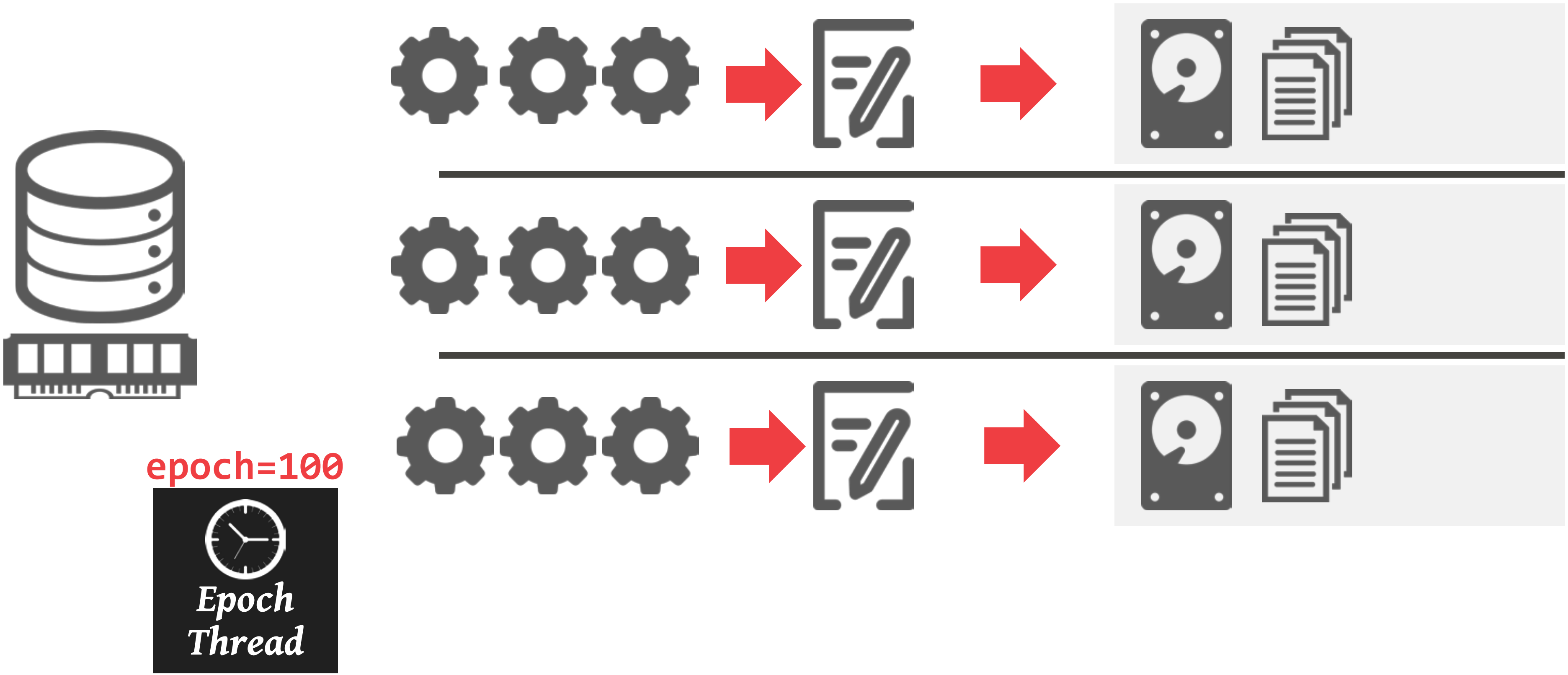
Txns that executed in epoch ***e*** can only release their results when the ***pepoch*** is durable to non-volatile storage.

SILOR – ARCHITECTURE

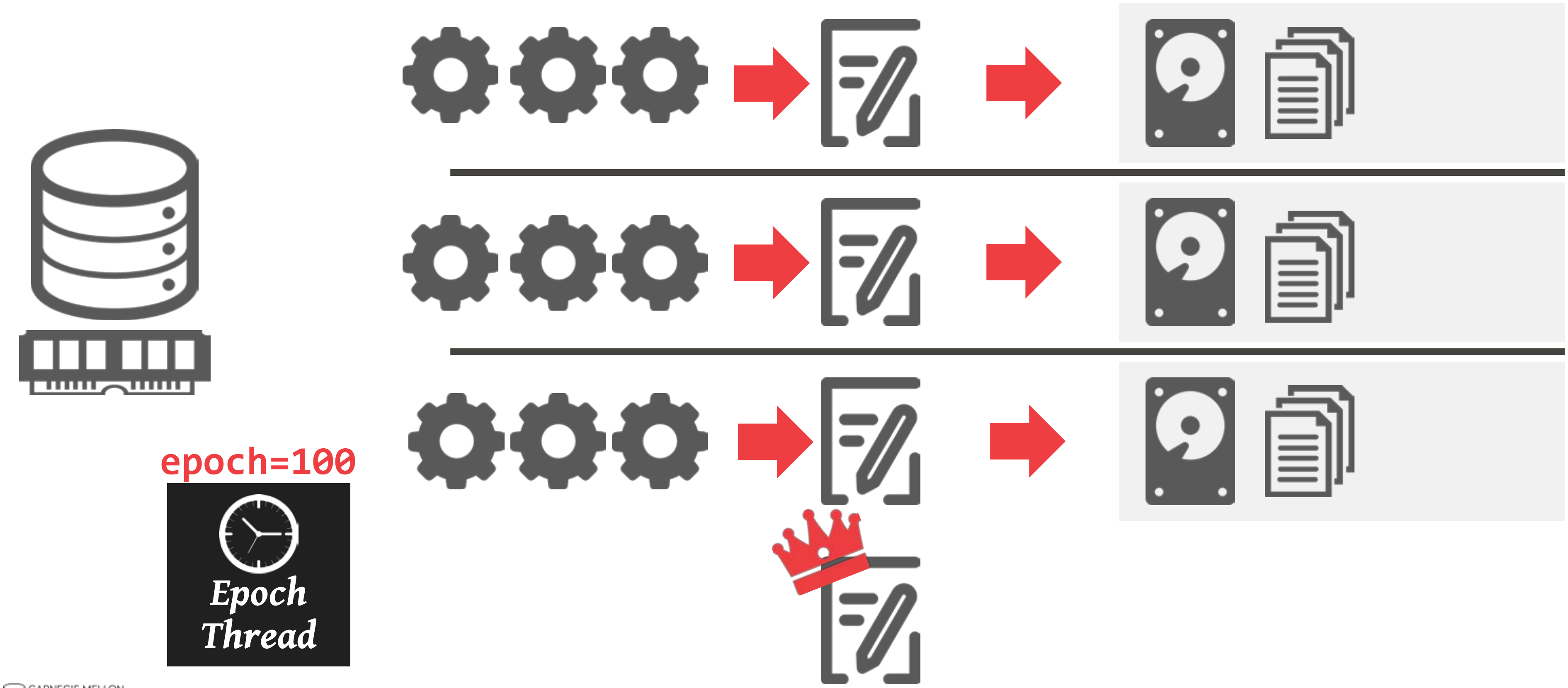


SILOR – ARCHITECTURE

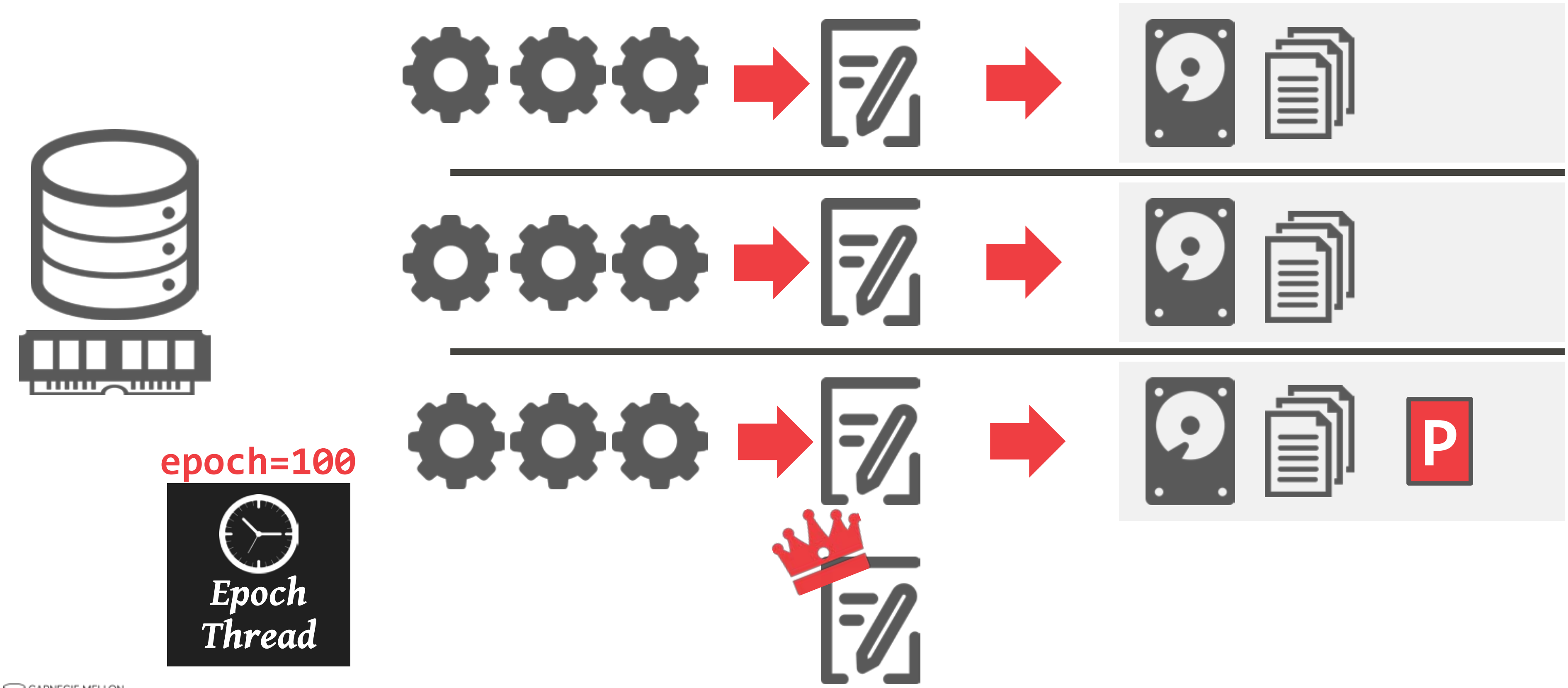
SILOR – ARCHITECTURE



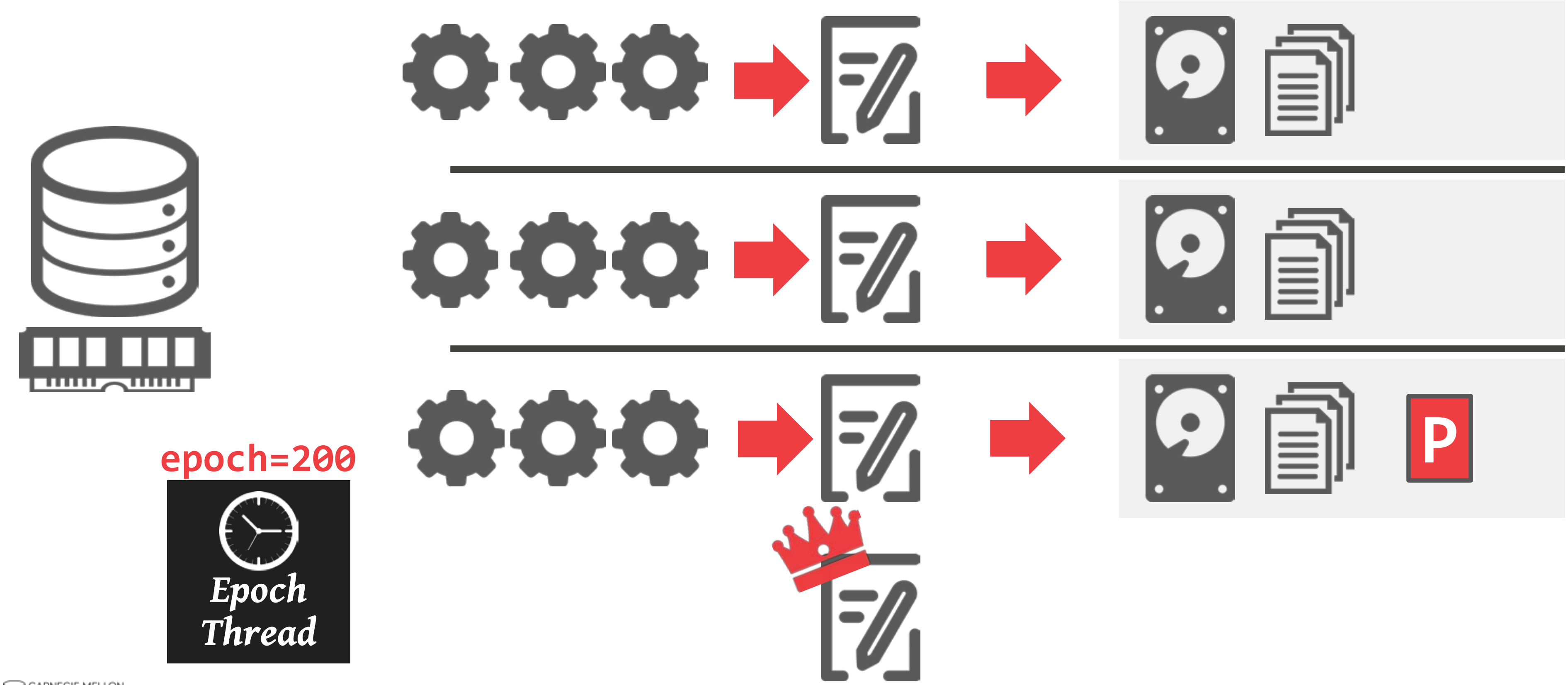
SILOR – ARCHITECTURE



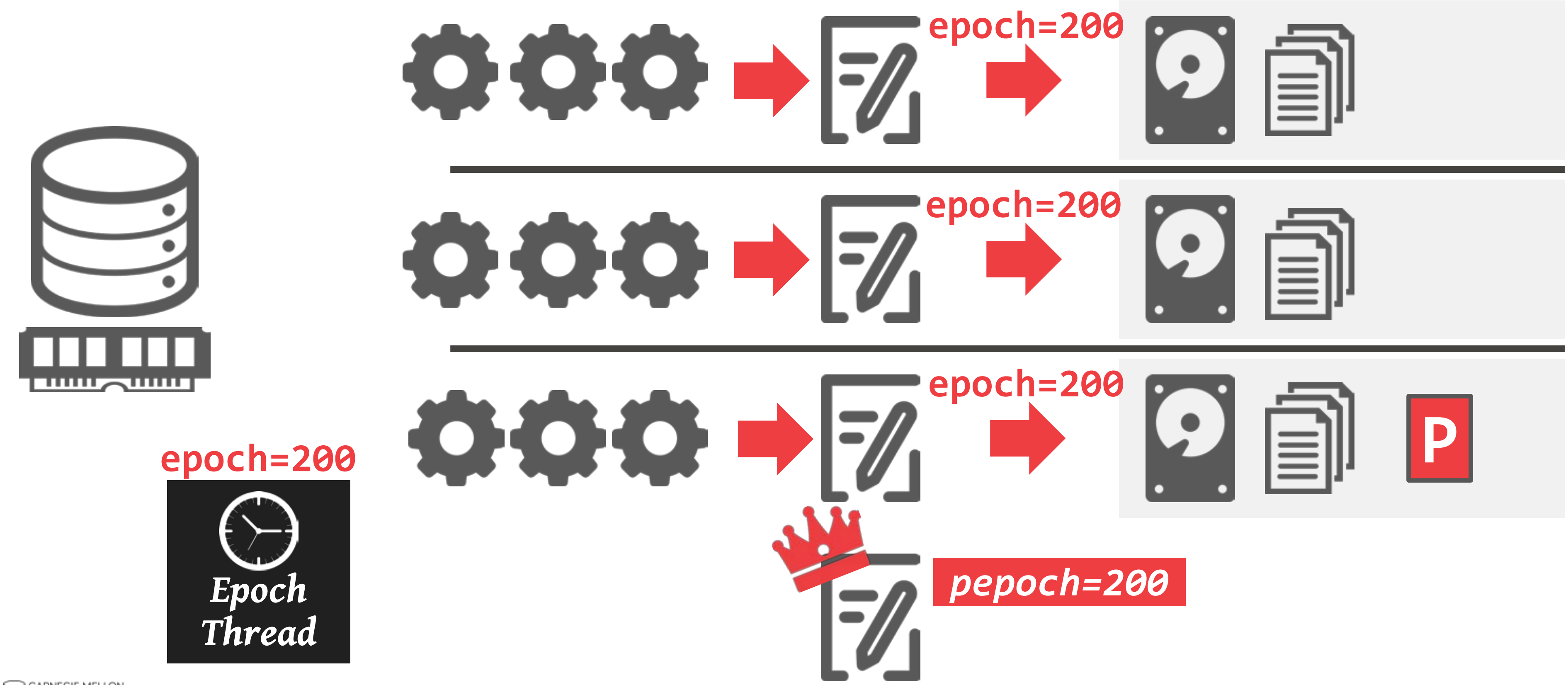
SILOR – ARCHITECTURE



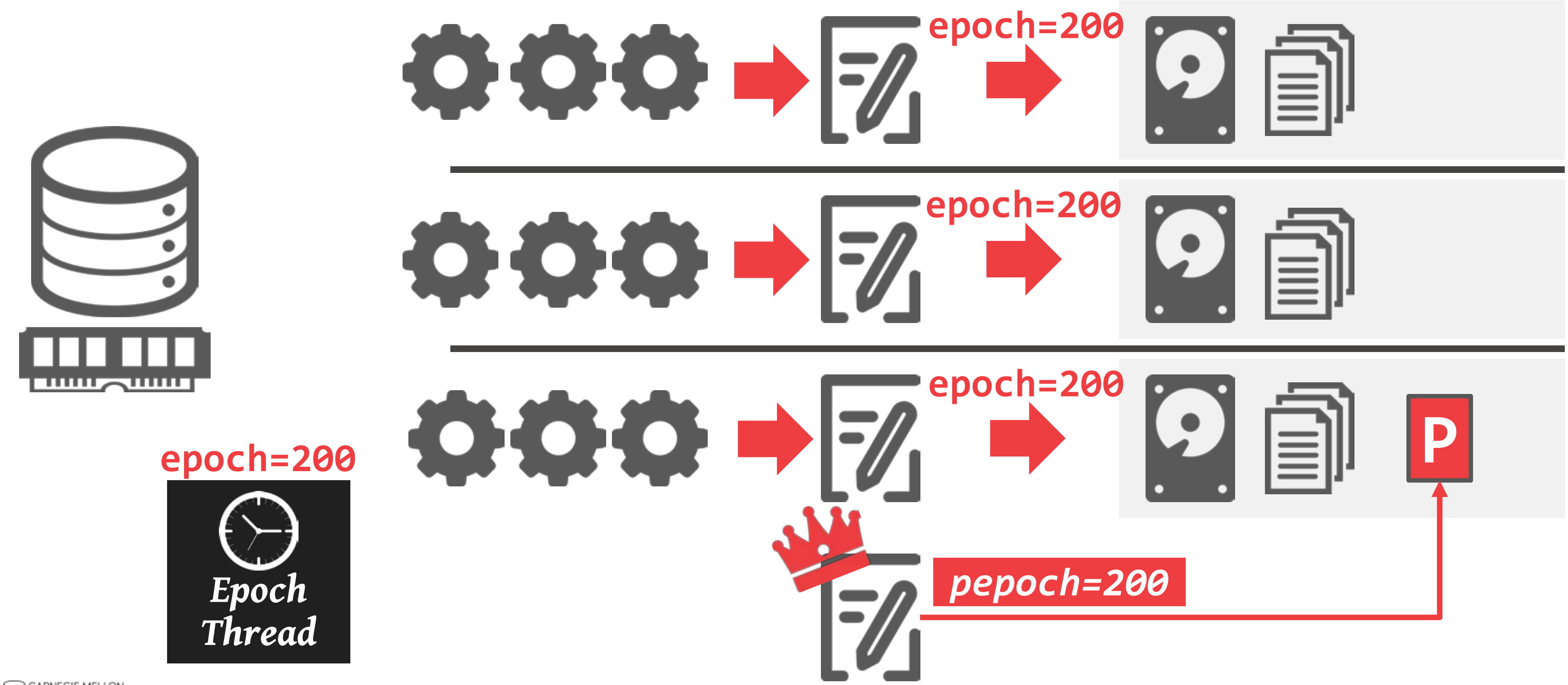
SILOR – ARCHITECTURE



SILOR – ARCHITECTURE



SILOR – ARCHITECTURE



SILOR – CHECKPOINT PROTOCOL

One checkpointer thread per disk.

→ The database is range partitioned and each thread writes to multiple file on a single disk.

Even though the database does not contain changes for uncommitted txns, it still may not see a consistent view of the database.

→ Multiple tuple changes are not atomic.

OBSERVATION

Creating a checkpoint in an MVCC DBMS is easy because older versions are still available.

VoltDB switches into a "multi-version" mode when it takes checkpoints.

CHECKPOINTS – FREQUENCY

Checkpointing too often causes the runtime performance to degrade.

→ The DBMS will spend too much time flushing buffers.

But waiting a long time between checkpoints is just as bad:

→ It will make recovery time much longer because the DBMS will have to replay a large log.

OBSERVATION

Certain segments of the database may not have changed from the last checkpoint.

Why can't the DBMS keep track of what blocks haven't changed since the last checkpoint and store a pointer to them in that checkpoint?

SILOR – RECOVERY PROTOCOL

Phase #1: Load Last Checkpoint

→ Install the contents of the last checkpoint that was saved into the database.

→ All indexes have to be rebuilt.

Phase #2: Replay Log

→ Process logs in reverse order to reconcile the latest version of each tuple.

CHECKPOINT RECOVERY

Multiple threads process the different checkpoint files on each disk.

Sequentially scan the records in each checkpoint file and insert them into database.

LOG RECOVERY

First check the ***pepoch*** file to determine the most recent persistent epoch.

→ Any log record from after the ***pepoch*** is ignored.

Log files are processed from newest to oldest.

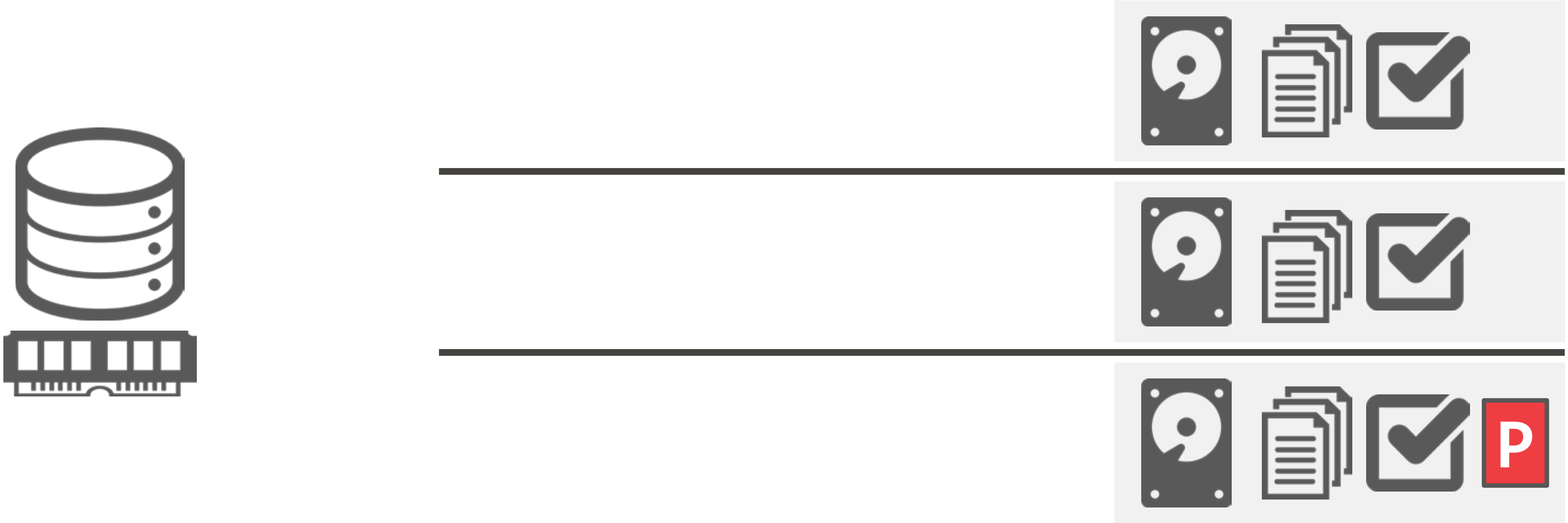
→ Value logging is able to be replayed in any order.

→ For each log record, the thread checks to see whether the tuple already exists.

→ If it does not, then it is created with the value.

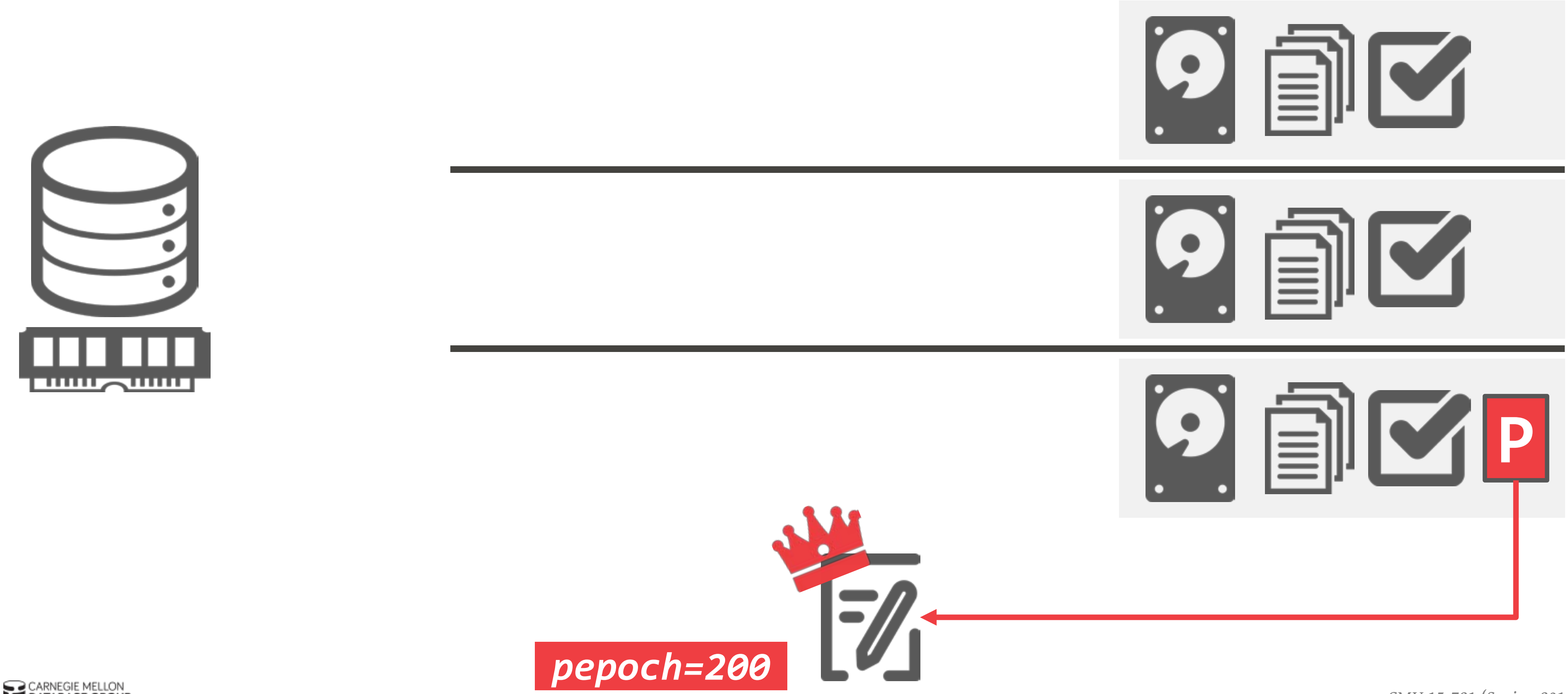
→ If it does, then the tuple's value is overwritten only if the log TID is newer than tuple's TID.

SILOR – RECOVERY PROTOCOL

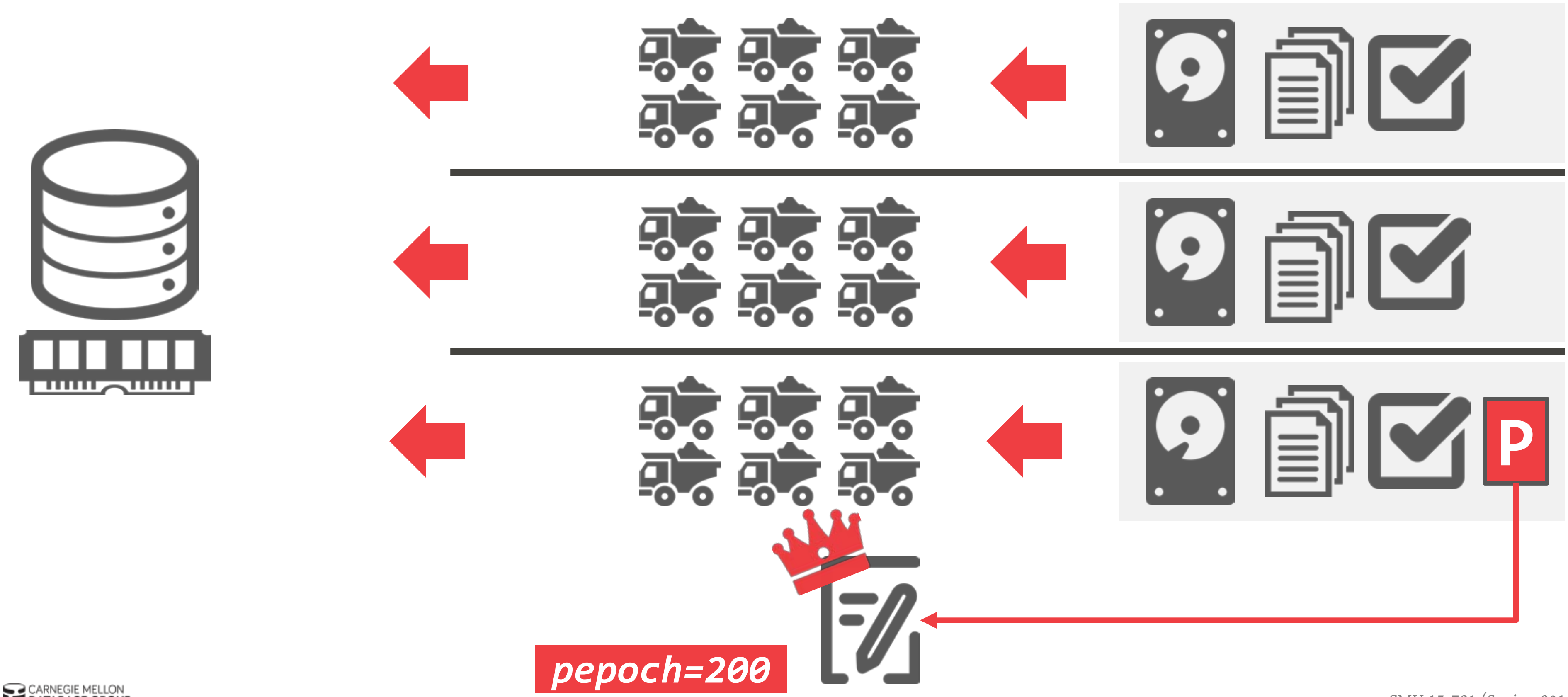


SILOR – RECOVERY PROTOCOL

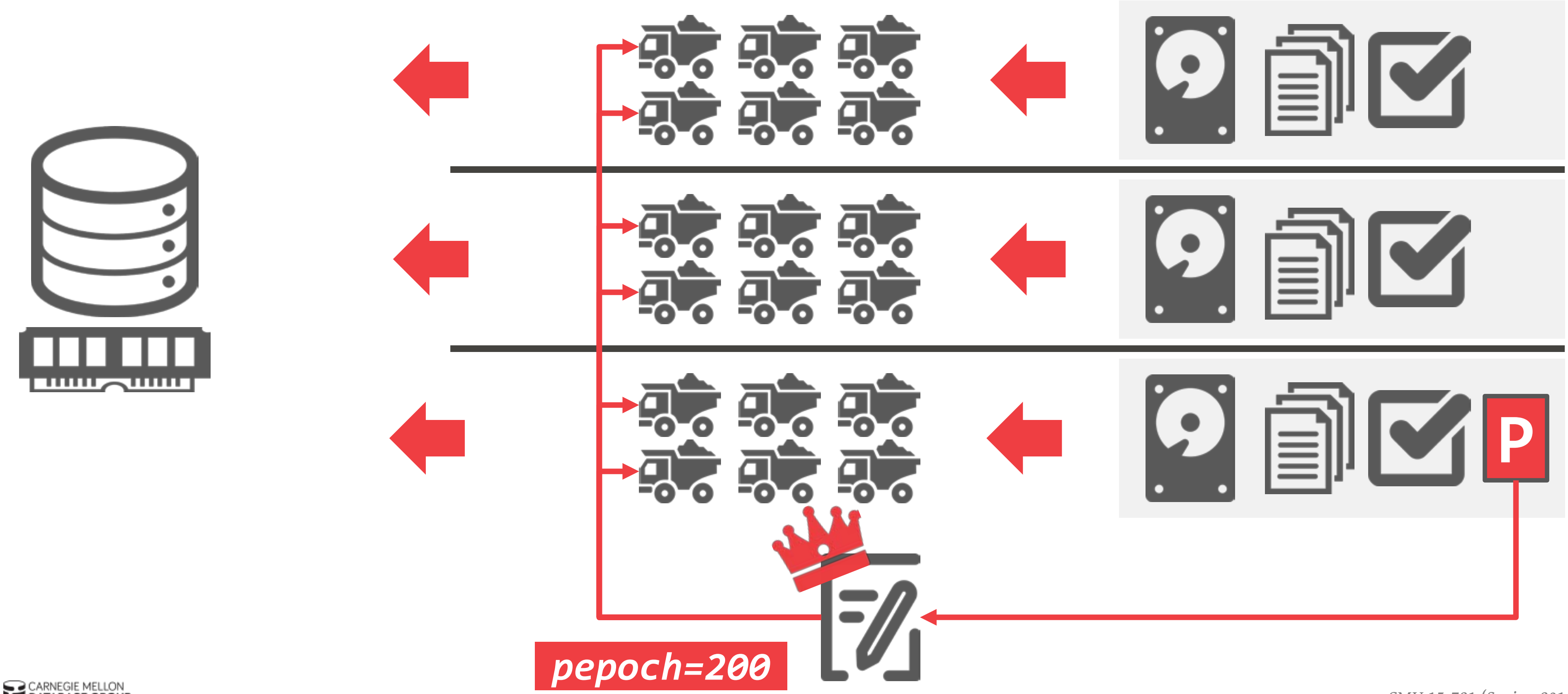
SILOR – RECOVERY PROTOCOL



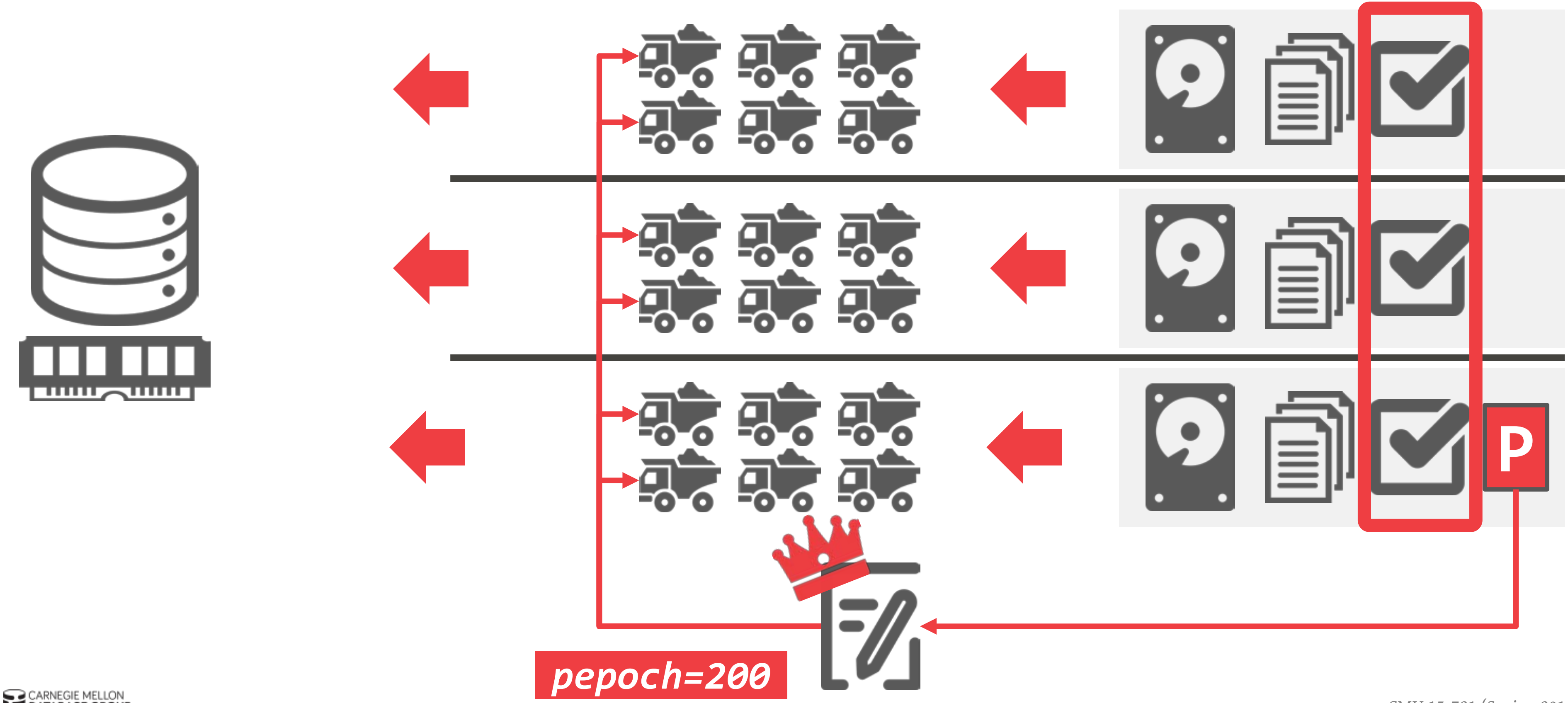
SILOR – RECOVERY PROTOCOL



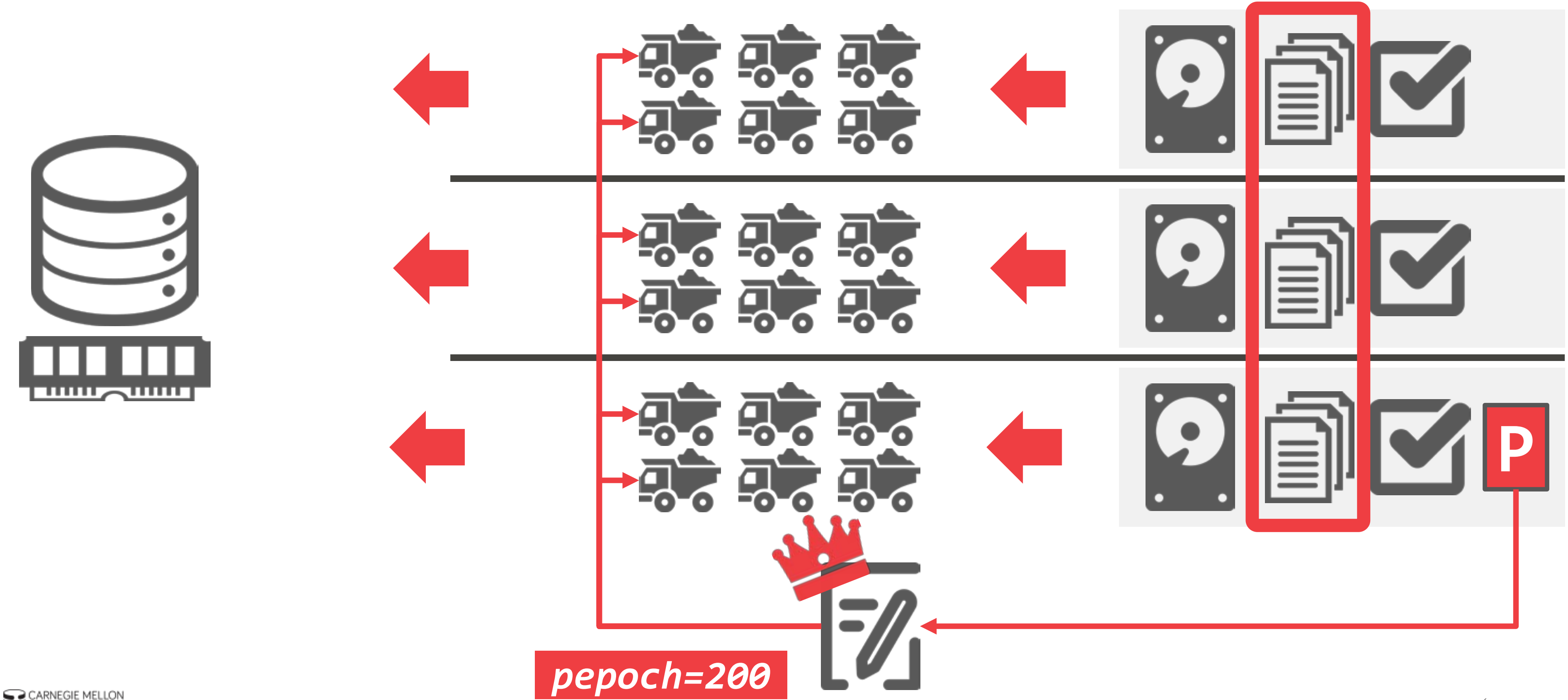
SILOR – RECOVERY PROTOCOL



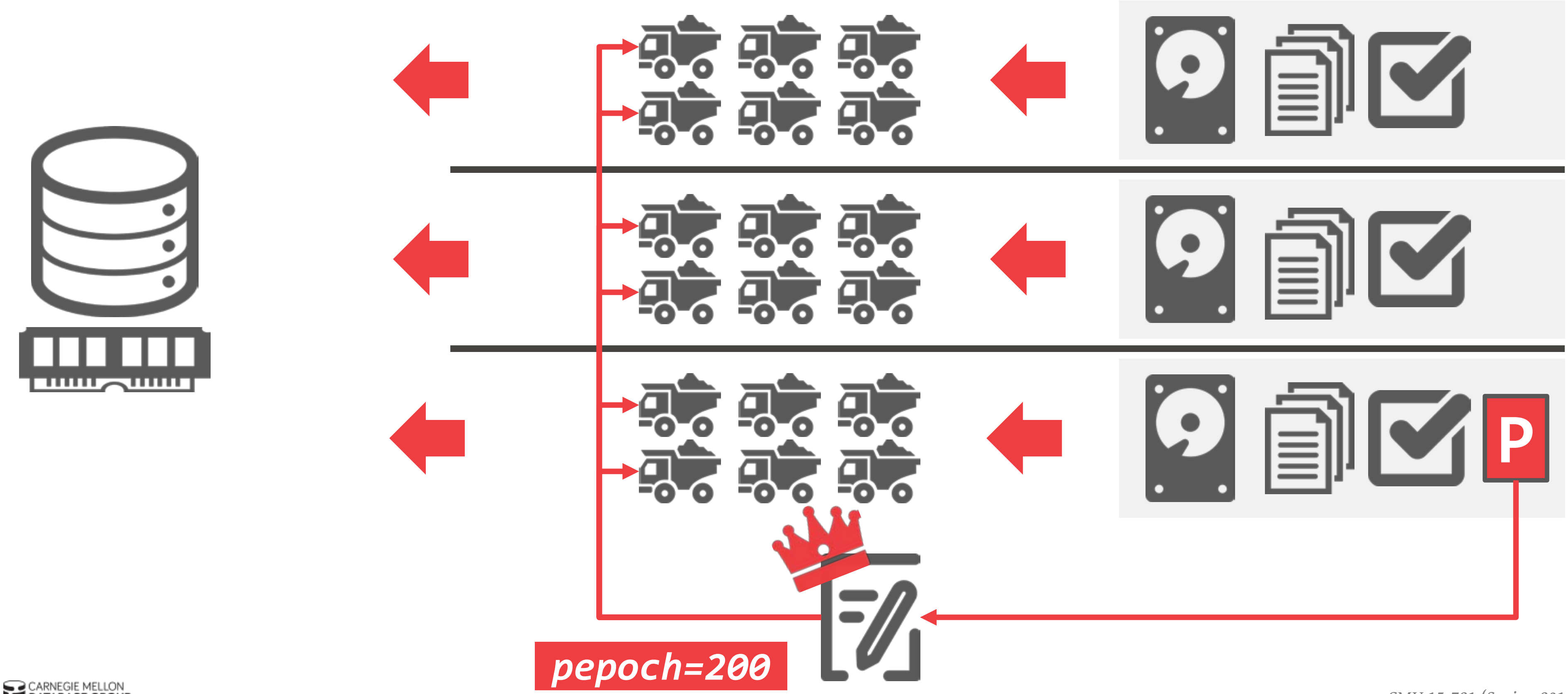
SILOR – RECOVERY PROTOCOL



SILOR – RECOVERY PROTOCOL



SILOR – RECOVERY PROTOCOL



OBSERVATION

The txn ids generated at runtime are enough to determine the serial order on recovery.

This is why SiloR does not need to maintain separate log sequence numbers for each entry.

EVALUATION

Comparing Silo performance with and without logging and checkpoints

YCSB + TPC-C Benchmarks

Hardware:

→ Four Intel Xeon E7-4830 CPUs (8 cores per socket)
→ 256 GB of DRAM
→ Three Fusion ioDrive2
→ RAID-5 Disk Array

EVALUATION

Comparing Silo performance with and without logging and checkpoints

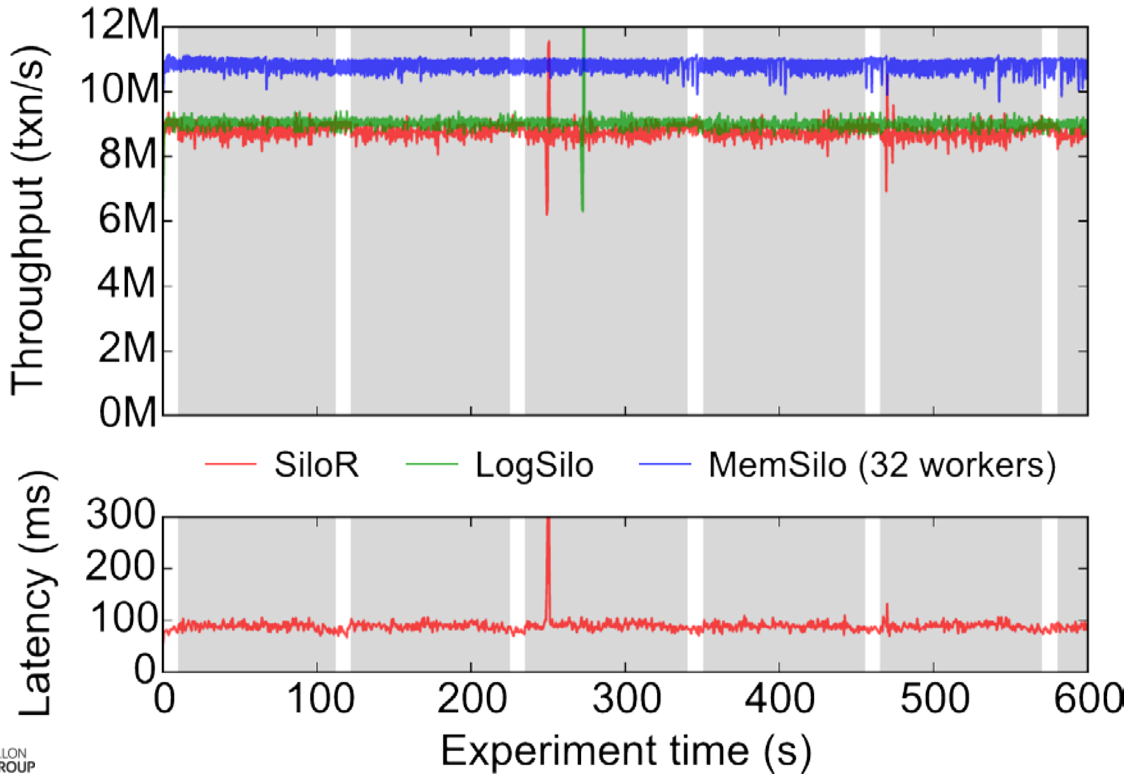
YCSB + TPC-C Benchmarks

Hardware:

→ Four Intel Xeon E7-4830 CPUs (8 cores per socket)
→ 256 GB of DRAM
→ Three Fusion ioDrive2
→ RAID-5 Disk Array

YCSB-A

70% Reads / 30% Writes



Average Throughput

SiloR: 8.76M txns/s

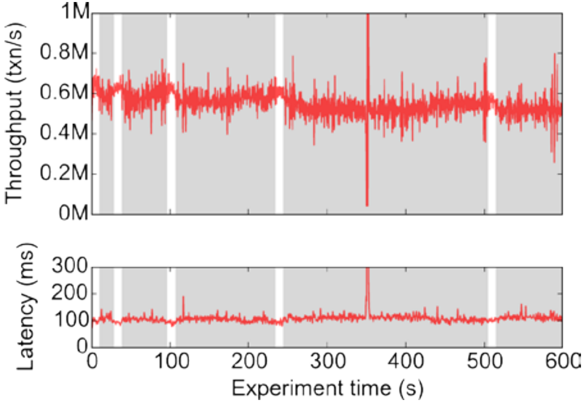
LogSilo: 9.01M txns/s

MemSilo: 10.83M txns/s

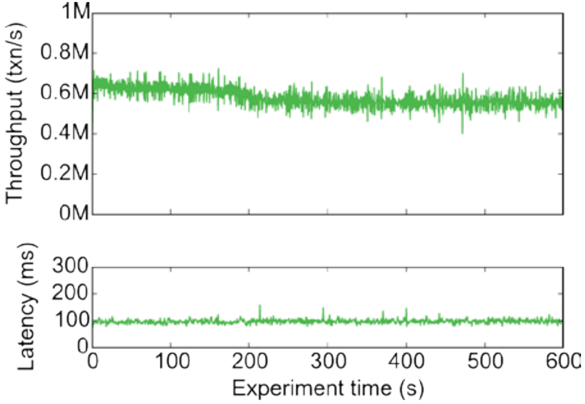
TPC-C

28 workers, 4 loggers, 4 checkpoint threads

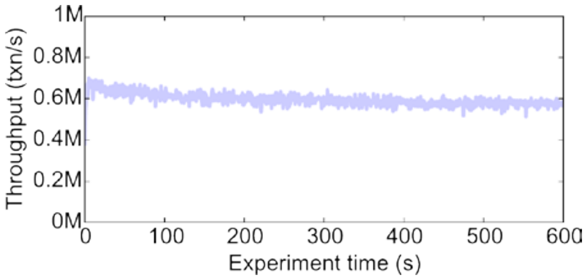
Logging+Checkpoints



Logging Only



No Recovery



Average Throughput

SiloR: 548K txns/s

LogSilo: 575K txns/s

MemSilo: 592 txns/s

RECOVERY TIMES

		Recovered Database	Checkpoint	Log	Total
YCSB	Size	43.2 GB	36 GB	64 GB	100 GB
	Recovery	-	33 sec	73 sec	106 sec
TPC-C	Size	72.2 GB	16.7 GB	180 GB	195.7 GB
	Recovery	-	17 sec	194 sec	211 sec

PARTING THOUGHTS

Physical logging is a general purpose approach that supports all concurrency control schemes.

NEXT CLASS

Checkpoint Schemes
Logical Logging
Facebook's Fast Restarts