Index
Concurrency Control
**ADMINISTRIVIA**

**Project #1** is due October 1\(^{st}\) @ 11:59pm
→ Special Office Hours: **Saturday Sept 30\(^{th}\) @ 3pm-5pm**

**Homework #2** is due Wed Oct 4\(^{th}\) @ 11:59pm

**Homework #3** is due Sun Oct 8\(^{th}\) @ 11:59pm

**Mid-Term Exam** is Wednesday Oct 11\(^{th}\)
→ During regular class time from 2:00-3:20pm
→ Please contact us if you need accommodations.
→ More details next week…
OBSERVATION

We (mostly) assumed all the data structures that we have discussed so far are single-threaded.

But a DBMS needs to allow multiple threads to safely access data structures to take advantage of additional CPU cores and hide disk I/O stalls.

They Don't Do This!
A **concurrency control** protocol is the method that the DBMS uses to ensure “correct” results for concurrent operations on a shared object.

A protocol's correctness criteria can vary:
→ **Logical Correctness**: Can a thread see the data that it is supposed to see?
→ **Physical Correctness**: Is the internal representation of the object sound?
TODAY'S AGENDA

Latches Overview
Hash Table Latching
B+Tree Latching
Leaf Node Scans
LOCKS VS. LATCHES

Locks (Transactions)
→ Protect the database's logical contents from other transactions.
→ Held for transaction's duration.
→ Need to be able to rollback changes.

Latches (Workers)
→ Protect the critical sections of the DBMS's internal data structure from other workers (e.g., threads).
→ Held for operation duration.
→ Do not need to be able to rollback changes.
# LOCKS VS. LATCHES

<table>
<thead>
<tr>
<th>Locks</th>
<th>Latches</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Separate...</strong></td>
<td>Transactions</td>
</tr>
<tr>
<td><strong>Protect...</strong></td>
<td>Workers (threads, processes)</td>
</tr>
<tr>
<td><strong>During...</strong></td>
<td>Database Contents</td>
</tr>
<tr>
<td><strong>Modes...</strong></td>
<td>In-Memory Data Structures</td>
</tr>
<tr>
<td><strong>Shared, Exclusive, Update, Intention</strong></td>
<td>Critical Sections</td>
</tr>
<tr>
<td><strong>Deadlock</strong></td>
<td>Detection &amp; Resolution</td>
</tr>
<tr>
<td><strong>...by...</strong></td>
<td>Avoidance</td>
</tr>
<tr>
<td><strong>Waits-for, Timeout, Aborts</strong></td>
<td>Coding Discipline</td>
</tr>
<tr>
<td><strong>Kept in...</strong></td>
<td>Protected Data Structure</td>
</tr>
<tr>
<td><strong>Lock Manager</strong></td>
<td></td>
</tr>
</tbody>
</table>

Source: [Goetz Graefe](#)
LATCH MODES

Read Mode
→ Multiple threads can read the same object at the same time.
→ A thread can acquire the read latch if another thread has it in read mode.

Write Mode
→ Only one thread can access the object.
→ A thread cannot acquire a write latch if another thread has it in any mode.

Compatibility Matrix

<table>
<thead>
<tr>
<th></th>
<th>Read</th>
<th>Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read</td>
<td>✔</td>
<td>X</td>
</tr>
<tr>
<td>Write</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
**LATCH IMPLEMENTATION GOALS**

Small memory footprint.

Fast execution path when no contention.
Deschedule thread when it has been waiting for too long to avoid burning cycles.

Each latch should not have to implement their own queue to track waiting threads.

Source: Filip Pizlo
LATCHE IMPLEMENTATION GOALS

Small memory footprint.

Fast execution path when no contention.

Deschedule thread when it has been waiting for too long to avoid burning cycles.

Each latch should not have to implement their own queue to track waiting threads.

Source: Filip Pizlo
LATCH IMPLEMENTATION GOALS

Small memory footprint.

Fast execution path when no contention.

Deschedule thread when it has been waiting for too long to avoid burning cycles.

Each latch should not have to implement its own queue to track waiting threads.

I repeat: do not use spinlocks in user space, unless you actually know what you're doing. And be aware that the likelihood that you know what you are doing is basically nil.
LATCH IMPLEMENTATIONS

Test-and-Set Spinlock
Blocking OS Mutex
Reader-Writer Locks

Advanced approaches:
→ Adaptive Spinlock (Apple ParkingLot)
→ Queue-based Spinlock (MCS Locks)
Approach #1: Test-and-Set Spin Latch (TAS)
→ Very efficient (single instruction to latch/unlatch)
→ Non-scalable, not cache friendly, not OS friendly.
→ Example: `std::atomic<T>`

```cpp
std::atomic_flag latch;
:
while (latch.test_and_set(...)) {
    // Retry? Yield? Abort?
}
```
Approach #2: Blocking OS Mutex
→ Simple to use
→ Non-scalable (about 25ns per lock/unlock invocation)
→ Example: `std::mutex` → `pthread_mutex` → `futex`

```cpp
std::mutex m;
:
  m.lock();
// Do something special...
  m.unlock();
```
LATCH IMPLEMENTATIONS

Approach #3: Reader-Writer Latches
→ Allows for concurrent readers. Must manage read/write queues to avoid starvation.
→ Can be implemented on top of spinlocks.
→ Example: `std::shared_mutex` → `pthread_rwlock_t`
HASH TABLE LATCHING

Easy to support concurrent access due to the limited ways threads access the data structure.
→ All threads move in the same direction and only access a single page/slot at a time.
→ Deadlocks are not possible.

To resize the table, take a global write latch on the entire table (e.g., in the header page).
HASH TABLE LATCHING

Approach #1: Page Latches
→ Each page has its own reader-writer latch that protects its entire contents.
→ Threads acquire either a read or write latch before they access a page.

Approach #2: Slot Latches
→ Each slot has its own latch.
→ Can use a single-mode latch to reduce meta-data and computational overhead.
T₁: Find D

hash(D)
**HASH TABLE – PAGE LATCHES**

**T₁:** Find D

\[ \text{hash}(D) \]

**T₂:** Insert E

\[ \text{hash}(E) \]
**HASH TABLE - PAGE LATCHES**

**T₁:** Find D  
\[ \text{hash}(D) \]

**T₂:** Insert E  
\[ \text{hash}(E) \]
HASH TABLE - PAGE LATCHES

- **T₁ (Find D)**: Find the value D in the hash table. 
  
  - hash(D)

- **T₂ (Insert E)**: Insert the value E into the hash table. 
  
  - hash(E)

**It’s safe to release the latch on Page #1.**

```
B | val
A | val
C | val
D | val
```
**HASH TABLE – PAGE LATCHES**

**T_1:** Find D  
\[ \text{hash}(D) \]

**T_2:** Insert E  
\[ \text{hash}(E) \]
**HASH TABLE – PAGE LATCHES**

### T₁: Find D

\[\text{hash}(D)\]

### T₂: Insert E

\[\text{hash}(E)\]
**HASH TABLE - PAGE LATCHES**

**T₁: Find D**  
$h(D)$

**T₂: Insert E**  
$h(E)$

- $D|val$  
- $C|val$  
- $A|val$  
- $B|val$
**T₁: Find D**

hash(D)

**T₂: Insert E**

hash(E)

- **hash(D)**: Find D
- **hash(E)**: Insert E

**Diagram:**

- **Page 0:** B
- **Page 1:** A
- **Page 2:** C, D

Page 2 is locked due to an operation being performed.

---

**CMU-DB 15-445/645 (Fall 2023)**
$T_1$: Find $D$

$\text{hash}(D)$

$T_2$: Insert $E$

$\text{hash}(E)$
**T₁: Find D**

\[\text{hash}(D)\]

**T₂: Insert E**

\[\text{hash}(E)\]
**HASH TABLE – SLOT LATCHES**

**T_1:** Find D

\[ \text{hash}(D) \]

**T_2:** Insert E

\[ \text{hash}(E) \]
**T₁: Find D**

$hash(D)$

**T₂: Insert E**

$hash(E)$
**T₁: Find D**

\[ hash(D) \]

**T₂: Insert E**

\[ hash(E) \]
\textbf{HASH TABLE – SLOT LATCHES}

$T_1$: Find D

$hash(D)$

$T_2$: Insert E

$hash(E)$

It's safe to release the latch on A
**HASH TABLE – SLOT LATCHES**

T₁: Find D

\[
\text{hash}(D)
\]

T₂: Insert E

\[
\text{hash}(E)
\]
**HASH TABLE - SLOT LATCHES**

T₁: Find D

\[ \text{hash}(D) \]

T₂: Insert E

\[ \text{hash}(E) \]

<table>
<thead>
<tr>
<th>Slot 0</th>
<th>Slot 1</th>
<th>Slot 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>val</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>val</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>D</td>
</tr>
</tbody>
</table>
**T₁: Find D**

\( \text{hash}(D) \)

**T₂: Insert E**

\( \text{hash}(E) \)
**T_1:** Find D 
\( hash(D) \)

**T_2:** Insert E 
\( hash(E) \)
**HASH TABLE - SLOT LATCHES**

\[ T_1: \text{Find D} \]
\[ \text{hash(D)} \]

\[ T_2: \text{Insert E} \]
\[ \text{hash(E)} \]
\( T_1: \) Find D  
\[
\text{hash}(D)
\]

\( T_2: \) Insert E  
\[
\text{hash}(E)
\]
We want to allow multiple threads to read and update a B+Tree at the same time.

We need to protect against two types of problems:

→ Threads trying to modify the contents of a node at the same time.
→ One thread traversing the tree while another thread splits/merges nodes.
**B+TREE MULTI-THREADED EXAMPLE**

T₁: Delete 44
\[ T_1: \text{Delete 44} \]
B+TREE MULTI-THREADED EXAMPLE

$T_1$: Delete 44

Rebalance!
**B+TREE MULTI-THREADED EXAMPLE**

\[ T_1: \text{Delete 44} \]

---

Rebalance!
**B+TREE MULTI-THREADED EXAMPLE**

T<sub>1</sub>: Delete 44
T<sub>2</sub>: Find 41

Rebalance!
**B+TREE MULTI-THREADED EXAMPLE**

\[ \text{T}_1: \text{Delete 44} \]
\[ \text{T}_2: \text{Find 41} \]

Rebalance!
B+TREE MULTI-THREADED EXAMPLE

T₁: Delete 44
T₂: Find 41

Rebalance!
**B+TREE MULTI-THREADED EXAMPLE**

**T₁:** Delete 44
**T₂:** Find 41

Rebalance!
B+TREE MULTI-THREADED EXAMPLE

T₁: Delete 44
T₂: Find 41
B+TREE MULTI-THREADED EXAMPLE

T₁: Delete 44
T₂: Find 41
LATCH CRABBING/COUPLING

Protocol to allow multiple threads to access/modify B+Tree at the same time.
→ Get latch for parent
→ Get latch for child
→ Release latch for parent if “safe”

A safe node is one that will not split or merge when updated.
→ Not full (on insertion)
→ More than half-full (on deletion)
**LATCH CRABBING/COUPLING**

**Find:** Start at root and traverse down the tree:
→ Acquire **R** latch on child,
→ Then unlatch parent.
→ Repeat until we reach the leaf node.

**Insert/Delete:** Start at root and go down, obtaining **W** latches as needed. Once child is latched, check if it is safe:
→ If child is safe, release all latches on ancestors
EXAMPLE #1 – FIND 38
EXAMPLE #1 – FIND 38

It is now safe to release the latch on A.
EXAMPLE #1 – FIND 38
EXAMPLE #1 - FIND 38
EXAMPLE #1 – FIND 38
EXAMPLE #2 – DELETE 38
We may need to coalesce B, so we can’t release the latch on A.
We know that D will not merge with C, so it is safe to release latches on A and B.
We know that D will not merge with C, so it is safe to release latches on A and B.
EXAMPLE #2 – DELETE 38
EXAMPLE #2 – DELETE 38
EXAMPLE #2 – DELETE 38
We know that if D needs to split, B has room so it is safe to release the latch on A.
EXAMPLE #3 – INSERT 45
Node I will not split, so we can release B+D.
Example #3 – Insert 45

Node I will not split, so we can release B+D.
EXAMPLE #3 – INSERT 45

Node I will not split, so we can release B+D.
EXAMPLE #3 – INSERT 45
EXAMPLE #4 – INSERT 25

```
3 4 6 9 10 11 12 13 20 22 23 31 35 36 38 41 44
```

```
6 12
```

```
10
```

```
20
```

```
A
```

```
W
```

```
B
```

```
35
```

```
C
```

```
44
```

```
D
```

```
E
```

```
F
```

```
G
```

```
H
```

```
I
```
EXAMPLE #4 – INSERT 25
EXAMPLE #4 – INSERT 25
We need to split F, so we need to hold the latch on its parent node.
We need to split node F, so we need to hold the latch on its parent node.
We need to split F, so we need to hold the latch on its parent node.
What was the first step that all the update examples did on the B+Tree?

**Delete 38**

**Insert 45**

**Insert 25**

Taking a write latch on the root every time becomes a bottleneck with higher concurrency.
BETTER LATCHING ALGORITHM

Most modifications to a B+Tree will not require a split or merge.

Instead of assuming that there will be a split/merge, optimistically traverse the tree using read latches.

If you guess wrong, repeat traversal with the pessimistic algorithm.
BETTER LATCHING ALGORITHM

Search: Same as before.

Insert/Delete:
→ Set latches as if for search, get to leaf, and set $W$ latch on leaf.
→ If leaf is not safe, release all latches, and restart thread using previous insert/delete protocol with write latches.

This approach optimistically assumes that only leaf node will be modified; if not, $R$ latches set on the first pass to leaf are wasteful.
EXAMPLE #2 – DELETE 38

R

A

B

C

D

E

F

G

H

I
EXAMPLE #2 – DELETE 38
EXAMPLE #2 – DELETE 38
EXAMPLE #2 – DELETE 38
EXAMPLE #2 – DELETE 38

Node H will not coalesce, so we’re safe!
Example #2 – Delete 38

Node H will not coalesce, so we’re safe!
We need to split F, so we have to restart and re-execute like before.
OBSERVATION

The threads in all the examples so far have acquired latches in a "top-down" manner.
→ A thread can only acquire a latch from a node that is below its current node.
→ If the desired latch is unavailable, the thread must wait until it becomes available.

But what if threads want to move from one leaf node to another leaf node?
LEAF NODE SCAN EXAMPLE #1

$T_1$: Find Keys < 4
LEAF NODE SCAN EXAMPLE #1

$T_1$: Find Keys $< 4$
LEAF NODE SCAN EXAMPLE #1

$T_i$: Find Keys < 4

Do not release latch on C until thread has latch on B
LEAF NODE SCAN EXAMPLE #1

$T_1$: Find Keys < 4

Do not release latch on C until thread has latch on B
LEAF NODE SCAN EXAMPLE #1

$T_1$: Find Keys < 4
LEAF NODE SCAN EXAMPLE #2

T₁: Find Keys < 4
T₂: Find Keys > 1
LEAF NODE SCAN EXAMPLE #2

$T_1$: Find Keys < 4

$T_2$: Find Keys > 1
LEAF NODE SCAN EXAMPLE #2

$T_1$: Find Keys $< 4$

$T_2$: Find Keys $> 1$
LEAF NODE SCAN EXAMPLE #2

Both $T_1$ and $T_2$ now hold this read latch.

$T_1$: Find Keys < 4

$T_2$: Find Keys > 1
LEAF NODE SCAN EXAMPLE #2

$T_1$: Find Keys < 4  
$T_2$: Find Keys > 1

Both $T_1$ and $T_2$ now hold this read latch.
**LEAF NODE SCAN EXAMPLE #2**

Only $T_1$ holds this read latch.

Only $T_2$ holds this read latch.

$T_1$: Find Keys < 4

$T_2$: Find Keys > 1
LEAF NODE SCAN EXAMPLE #3

$T_1$: Delete 4

$T_2$: Find Keys > 1
LEAF NODE SCAN EXAMPLE #3

$T_1$: Delete 4
$T_2$: Find Keys > 1
LEAF NODE SCAN EXAMPLE #3

T₁: Delete 4
T₂: Find Keys > 1
**LEAF NODE SCAN EXAMPLE #3**

**T₁:** Delete 4  
**T₂:** Find Keys > 1

*T₂* cannot acquire the read latch on C

*T₂* does not know what *T₁* is doing…
LEAF NODE SCAN EXAMPLE #3

$T_1$: Delete 4
$T_2$: Find Keys > 1

$T_2$ Choices?
- Wait
- Kill Ourselves
- Kill Other Thread

$T_2$ cannot acquire the read latch on C

$T_2$ does not know what $T_1$ is doing...
T₂ Choices?
- Wait
- Kill Ourself
- Kill Other Thread

T₁: Delete 4
T₂: Find Keys > 1

T₂ cannot acquire the read latch on C

T₂ does not know what T₁ is doing...
Latches do not support deadlock detection or avoidance. The only way we can deal with this problem is through coding discipline.

The leaf node sibling latch acquisition protocol must support a “no-wait” mode.

The DBMS's data structures must cope with failed latch acquisitions.
CONCLUSION

Making a data structure thread-safe is notoriously difficult in practice.

We focused on B+Trees, but the same high-level techniques are applicable to other data structures.
NEXT CLASS

We are finally going to discuss how to execute some queries...
COMPARE-AND-SWAP

Atomic instruction that compares contents of a memory location \( M \) to a given value \( V \)
→ If values are equal, installs new given value \( V' \) in \( M \)
→ Otherwise, operation fails

\[
\text{__sync_bool_compare_and_swap}(\&M, 20, 30)
\]

\( M \)
\[
30
\]