TODAY'S AGENDA

Index Locks vs. Latches
Latch Implementations
Index Latching (Logical)
Index Locking (Physical)
DATABASE INDEX

A data structure that improves the speed of data retrieval operations on a table at the cost of additional writes and storage space.

Indexes are used to quickly locate data without having to search every row in a table every time a table is accessed.
DATA STRUCTURES

Order Preserving Indexes
→ A tree-like structure that maintains keys in some sorted order.
→ Supports all possible predicates with $O(\log n)$ searches.

Hashing Indexes
→ An associative array that maps a hash of the key to a particular record.
→ Only supports equality predicates with $O(1)$ searches.
**B-TREE VS. B+TREE**

The original **B-tree** from 1972 stored keys + values in all nodes in the tree.  
→ More memory efficient since each key only appears once in the tree.

A **B+tree** only stores values in leaf nodes. Inner nodes only guide the search process.  
→ Easier to manage concurrent index access when the values are only in the leaf nodes.
OBSERVATION

We already know how to use locks to protect objects in the database.

But we have to treat indexes differently because the physical structure can change as long as the logical contents are consistent.
SIMPLE EXAMPLE

| K0 | K2 | A |

Txn #1: READ(K2)
SIMPLE EXAMPLE

<table>
<thead>
<tr>
<th>K0</th>
<th>K2</th>
</tr>
</thead>
</table>

**Txn #1:** READ(K2)

**Txn #2:** INSERT(K1)
SIMPLE EXAMPLE

Txn #1: READ(K2)

Txn #2: INSERT(K1)
SIMPLE EXAMPLE

Txn #1: READ(K2)

Txn #2: INSERT(K1)
SIMPLE EXAMPLE

Txn #1: READ(K2)

Txn #2: INSERT(K1)

Txn #1: READ(K2)
LOCKS VS. LATCHES

Locks
→ Protects the index’s logical contents from other txns.
→ Held for txn duration.
→ Need to be able to rollback changes.

Latches
→ Protects the critical sections of the index’s internal data structure from other 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>User transactions</td>
</tr>
<tr>
<td><strong>Protect...</strong></td>
<td>Database Contents</td>
</tr>
<tr>
<td><strong>During...</strong></td>
<td>Entire Transactions</td>
</tr>
<tr>
<td><strong>Modes...</strong></td>
<td>Shared, Exclusive, Update, Intention</td>
</tr>
<tr>
<td><strong>Deadlock</strong></td>
<td>Detection &amp; Resolution</td>
</tr>
<tr>
<td><strong>...by...</strong></td>
<td>Waits-for, Timeout, Aborts</td>
</tr>
<tr>
<td><strong>Kept in...</strong></td>
<td>Lock Manager</td>
</tr>
</tbody>
</table>

Source: Goetz Graefe
LOCK-FREE INDEXES

Possibility #1: No Locks
→ Txns don’t acquire locks to access/modify database.
→ Still have to use latches to install updates.

Possibility #2: No Latches
→ Swap pointers using atomic updates to install changes.
→ Still have to use locks to validate txns.
LATCH IMPLEMENTATIONS

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

Source: Anastasia Ailamaki
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.

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

__sync_bool_compare_and_swap(&M, 20, 30)
LATCH IMPLEMENTATIONS

Choice #1: Blocking OS Mutex

→ Simple to use
→ Non-scalable (about 25ns per lock/unlock invocation)
→ Example: `std::mutex`

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

Choice #2: Test-and-Set Spinlock (TAS)
→ Very efficient (single instruction to lock/unlock)
→ Non-scalable, not cache friendly
→ Example: `std::atomic<T>`

```cpp
std::atomic_flag latch;
:
while (latch.test_and_set(…)) {
    // Yield? Abort? Retry?
}
```
LATCH IMPLEMENTATIONS

Choice #2: Test-and-Set Spinlock (TAS)
→ Very efficient (single instruction to lock/unlock)
→ Non-scalable, not cache friendly
→ Example: `std::atomic<T>`

```
std::atomic_flag latch;
:
while (latch.test_and_set(...)) {
    // Yield? Abort? Retry?
}
```
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`

**Base Latch**

![Base Latch Diagram]
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: std::atomic<Latch>*

Mellor-Crummey and Scott

Base Latch

next

CPU1 Latch

next
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`

Mellor-Crummey and Scott
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`

Mellor-Crummey and Scott

Base Latch

CPU1 Latch

next

CPU1
**LATCH IMPLEMENTATIONS**

**Choice #3: Queue-based Spinlock (MCS)**

- More efficient than mutex, better cache locality
- Non-trivial memory management
- Example: `std::atomic<Latch*>`

---

**Mellor-Crummey and Scott**

![Diagram showing Base Latch and CPU1 Latch connections]
Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)

→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`

Mellor-Crummey and Scott

Base Latch

CPU1 Latch

CPU2 Latch

CPU1

CPU2
Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`

Base Latch → CPU1 Latch → CPU2 Latch

Mellor-Crummey and Scott
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`

Mellor-Crummey and Scott
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`

Mellor-Crummey and Scott

Base Latch

CPU1 Latch

CPU2 Latch

CPU3 Latch

(next)
LATCH IMPLEMENTATIONS

Choice #3: Queue-based Spinlock (MCS)
→ More efficient than mutex, better cache locality
→ Non-trivial memory management
→ Example: `std::atomic<Latch*>`
LATCH IMPLEMENTATIONS

Choice #4: Reader-Writer Locks

→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks
Latch Implementations

Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks

Latch

read
▫️=0
▫️=0

write
▫️=0
▫️=0
LATCH IMPLEMENTATIONS

Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks
LATCH IMPLEMENTATIONS

Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks

Latch
read
write
\[
\begin{align*}
\text{read: } & L = 1 \\
\text{write: } & W = 0
\end{align*}
\]
LATCH IMPLEMENTATIONS

Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks
LATCH IMPLEMENTATIONS

Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks
LATCH IMPLEMENTATIONS

Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks
Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks
LATCH IMPLEMENTATIONS

Choice #4: Reader-Writer Locks
→ Allows for concurrent readers
→ Have to manage read/write queues to avoid starvation
→ Can be implemented on top of spinlocks

Latch

read
\[\text{read} = 2\]
\[\text{read} = 1\]

write
\[\text{write} = 0\]
\[\text{write} = 1\]
LATCH CRABBING / COUPLING

Acquire and release latches on B+Tree nodes when traversing the data structure.

A thread can release latch on a parent node if its child node considered **safe**.

- Any node that won’t split or merge when updated.
- Not full (on insertion)
- More than half-full (on deletion)
LATCH CRABBING

**Search:** Start at root and go down; repeatedly,
→ Acquire read (R) latch on child
→ Then unlock the parent node.

**Insert/Delete:** Start at root and go down,
obtaining write (W) latches as needed.
Once child is locked, check if it is safe:
→ If child is safe, release all locks on ancestors.
EXAMPLE #1: SEARCH 23
EXAMPLE #1: SEARCH 23

Diagram showing a binary search tree with nodes labeled A, B, C, D, E, F, G. The search process involves comparing the target value 23 with the values in the tree nodes.
EXAMPLE #1: SEARCH 23

We can release the latch on A as soon as we acquire the latch for C.
EXAMPLE #1: SEARCH 23

We can release the latch on $A$ as soon as we acquire the latch for $C$. 
We can release the latch on A as soon as we acquire the latch for C.
EXAMPLE #1: SEARCH 23

We can release the latch on A as soon as we acquire the latch for C.
EXAMPLE #2: DELETE 44
EXAMPLE #2: DELETE 44
EXAMPLE #2: DELETE 44

We may need to coalesce C, so we can’t release the latch on A.
EXAMPLE #2: DELETE 44

We may need to coalesce C, so we can’t release the latch on A.

G will not merge with F, so we can release latches on A and C.
EXAMPLE #2: DELETE 44

We may need to coalesce C, so we can’t release the latch on A.

G will not merge with F, so we can release latches on A and C.
EXAMPLE #2: DELETE 44

We may need to coalesce C, so we can’t release the latch on A. G will not merge with F, so we can release latches on A and C.
EXAMPLE #3: INSERT 40
EXAMPLE #3: INSERT 40
EXAMPLE #3: INSERT 40

C has room if its child has to split, so we can release the latch on A.
EXAMPLE #3: INSERT 40

C has room if its child has to split, so we can release the latch on A.
C has room if its child has to split, so we can release the latch on A.

G has to split, so we can’t release the latch on C.
EXEMPLARY #3: INSERT 40

C has room if its child has to split, so we can release the latch on A.

G has to split, so we can’t release the latch on C.
C has room if its child has to split, so we can release the latch on A.

G has to split, so we can’t release the latch on C.
EXAMPLE #3: INSERT 40

C has room if its child has to split, so we can release the latch on A.

G has to split, so we can’t release the latch on C.
OBSERVATION

What was the first step that the DBMS took in the two examples that updated the index?

**Delete 44**

**Insert 40**
BETTER LATCH CRABBING

Optimistically assume that the leaf is safe.
→ Take R latches as you traverse the tree to reach it and verify.
→ If leaf is not safe, then do previous algorithm.
EXAMPLE #4: DELETE 44
EXAMPLE #4: DELETE 44
EXAMPLE #4: DELETE 44

We assume that C is safe, so we can release the latch on A.
EXAMPLE #4: DELETE 44

We assume that C is safe, so we can release the latch on A.
EXAMPLE #4: DELETE 44

We assume that C is safe, so we can release the latch on A.

Acquire an exclusive latch on G.
We assume that C is safe, so we can release the latch on A.
Acquire an exclusive latch on G.
We assume that C is safe, so we can release the latch on A.

Acquire an exclusive latch on G.
OBSERVATION

Crabbing ensures that txns do not corrupt the internal data structure during modifications.

But because txns release latches on each node as soon as they are finished their operations, we cannot guarantee that phantoms do not occur...
PROBLEM SCENARIO #1
PROBLEM SCENARIO #1

Txn #1: READ(25)
PROBLEM SCENARIO #1

Txn #1: READ(25)
PROBLEM SCENARIO #1

Txn #1: READ(25)
PROBLEM SCENARIO #1

Txn #1:
READ(25)
PROBLEM SCENARIO #1

Txn #1: READ(25)

Txn #2: INSERT(25)
PROBLEM SCENARIO #1

Txn #1: READ(25)

Txn #2: INSERT(25)
PROBLEM SCENARIO #1

Txn #1: READ(25)

Txn #2: INSERT(25)
PROBLEM SCENARIO #1

Txn #1:
- READ(25)

Txn #2:
- INSERT(25)

Txn #1:
- INSERT(25)
PROBLEM SCENARIO #1

Txn #1:
- INSERT(25)

Txn #2:
- INSERT(25)
- READ(25)

Txn #1:
- INSERT(25)
PROBLEM SCENARIO #2

Txn #1: [12, 23]
PROBLEM SCENARIO #2

Txn #1: [12, 23]
PROBLEM SCENARIO #2

Txn #1: [12, 23]
PROBLEM SCENARIO #2

Txn #1:  
INSERT(21)  
[12, 23]

Txn #2:  

Diagram:
- Node A (20)
- Node B (10)
- Node C (35)
- Node D (6)
- Node E (23)
- Node F (38)
- Node G (44)

Edges and transactions:
- Node A to B: Write 21
- Node A to C: Insert 21
- Node B to D: Write 21
- Node B to E: Write 21
- Node C to F: Insert 21
PROBLEM SCENARIO #2

Txn #1:
INSERT(21)
[12, 23]

Txn #2:
23
21
PROBLEM SCENARIO #2

Txn #1: 
INSERT(21)
[12, 23]

Txn #2: 
INSERT(21)

Diagram:

- A
- B
- C
- D
- E
- F
- G

Nodes:
- 10
- 20
- 35
- 6
- 12
- 21
- 23
- 38
- 44

Transactions:
- Txn #1: R
- Txn #2: R
- Txn #1: R
INDEX LOCKS

Need a way to protect the index’s logical contents from other txns to avoid phantoms.

Difference with index latches:
→ Locks are held for the entire duration of a txn.
→ Only acquired at the leaf nodes.
→ Not physically stored in index data structure.

Can be used with any order-preserving index.
INDEX LOCKS

Lock Table

- txn1
- txn2
- txn3
- ...
- ...
- ...
- ...

-txn4
-txn6
-txn5

-IX
-X
-S
INDEX LOCKING SCHEMES

Predicate Locks
Key-Value Locks
Gap Locks
Key-Range Locks
Hierarchical Locking
PREDICATE LOCKS

Proposed locking scheme from System R.
→ Shared lock on the predicate in a WHERE clause of a SELECT query.
→ Exclusive lock on the predicate in a WHERE clause of any UPDATE, INSERT, or DELETE query.

Never implemented in any system.
### PREDICATE LOCKS

**SQL Query**

- **SELECT** SUM(balance) 
  FROM account 
  WHERE name = 'Biggie'

- **INSERT INTO** account 
  (name, balance) 
  VALUES ('Biggie', 100);

**Records in Table "account"**

- name='Biggie'
- name='Biggie' ∧ balance=100
KEY-VALUE LOCKS

Locks that cover a single key value.
Need “virtual keys” for non-existent values.

B+Tree Leaf Node

10  12  14  16

Key [14, 14]
GAP LOCKS

Each txn acquires a key-value lock on the single key that it wants to access. Then get a gap lock on the next key gap.

B+Tree Leaf Node

<table>
<thead>
<tr>
<th>10</th>
<th>{Gap}</th>
<th>12</th>
<th>{Gap}</th>
<th>14</th>
<th>{Gap}</th>
<th>16</th>
</tr>
</thead>
</table>

Gap (14, 16)
KEY-RANGE LOCKS

A txn takes locks on ranges in the key space.
→ Each range is from one key that appears in the relation, to the next that appears.
→ Define lock modes so conflict table will capture commutativity of the operations available.
KEY-RANGE LOCKS

Locks that cover a key value and the gap to the next key value in a single index.
→ Need “virtual keys” for artificial values (infinity)

B+Tree Leaf Node

| 10 | {Gap} | 12 | {Gap} | 14 | {Gap} | 16 |

Next Key [14, 16]
KEY-RANGE LOCKS

Locks that cover a key value and the gap to the next key value in a single index.
→ Need “virtual keys” for artificial values (infinity)

B+Tree Leaf Node

10 {Gap} 12 {Gap} 14 {Gap} 16

Prior Key (12, 14)
HIERARCHICAL LOCKING

Allow for a txn to hold wider key-range locks with different locking modes.
→ Reduces the number of visits to lock manager.
HIERARCHICAL LOCKING

Allow for a txn to hold wider key-range locks with different locking modes.
→ Reduces the number of visits to lock manager.

[B+Tree Leaf Node]

10 {Gap} 12 {Gap} 14 {Gap} 16
[10, 16]
HIERARCHICAL LOCKING

Allow for a txn to hold wider key-range locks with different locking modes.
→ Reduces the number of visits to lock manager.
HIERARCHICAL LOCKING

Allow for a txn to hold wider key-range locks with different locking modes.
→ Reduces the number of visits to lock manager.
PARTING THOUGHTS

Hierarchical locking essentially provides predicate locking without complications.
→ Index locking occurs only in the leaf nodes.
→ Latching is to ensure consistent data structure.

Peloton currently does not support serializable isolation with range scans.
ANDY’S TIPS FOR PROFILING
MOTIVATION

Consider a program with functions \texttt{foo} and \texttt{bar}.

How can we speed it up with only a debugger?
\rightarrow Randomly pause it during execution
\rightarrow Collect the function call stack
RANDOM PAUSE METHOD

Consider this scenario
→ Collected 10 call stack samples
→ Say 6 out of the 10 samples were in foo

What percentage of time was spent in foo?
→ Roughly 60% of the time was spent in foo
→ Accuracy increases with # of samples
Say we optimized `foo` to run two times faster
What’s the expected overall speedup?
→ 60% of time spent in `foo` drops in half
→ 40% of time spent in `bar` unaffected

By Amdahl’s law, overall speedup = $\frac{1}{p + \frac{1}{s}(1-p)}$
→ $p$ = percentage of time spent in optimized task
→ $s$ = speed up for the optimized task
→ Overall speedup = $\frac{1}{\frac{0.6}{2} + 0.4} = 1.4$ times faster
PROFILING TOOLS FOR REAL

Choice #1: Valgrind
→ Heavyweight binary instrumentation framework with different tools to measure different events.

Choice #2: Perf
→ Lightweight tool that uses hardware counters to capture events during execution.
CHOICE #1: VALGRIND

Instrumentation framework for building dynamic analysis tools.

→ **memcheck**: a memory error detector
→ **callgrind**: a call-graph generating profiler
→ **massif**: memory usage tracking.
Using callgrind to profile the index test and Peloton in general:

```
$ valgrind --tool=callgrind --trace-children=yes ./relwithdebinfo/concurrent_read_benchmark
```

Profile data visualization tool:

```
$ kcachegrind callgrind.out.12345
```
Using callgrind to profile the index test and Peloton in general:

Profile data visualization tool:

```
$ kcachegrind callgrind.out.12345
```

```
$ valgrind --tool=callgrind --trace=children=yes ./
relwithdebinfo/concurrent_read_benchmark
```

Cumulative Time Distribution

Callgraph View
CHOICE #2: PERF

Tool for using the performance counters subsystem in Linux.

- `-e` = sample the event cycles at the user level only
- `-c` = collect a sample every 2000 occurrences of event

```
$ perf record -e cycles:u -c 2000
./relwithdebinfo/concurrent_read_benchmark
```

Uses counters for tracking events

- On counter overflow, the kernel records a sample
- Sample contains info about program execution
PERF VISUALIZATION

We can also use `perf` to visualize the generated profile for our application.

```
$ perf report
```

There are also third-party visualization tools:
→ `Hotspot`
We can also use `perf` to visualize the generated profile for our application. There are also third-party visualization tools:

---

<table>
<thead>
<tr>
<th>Command</th>
<th>Shared Object</th>
<th>Symbol</th>
</tr>
</thead>
</table>
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | libb-2.27.so | .....
| concurrent_read | libb-2.27.so | .....
| concurrent_read | libb-2.27.so | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | libb.so.2 | .....
| concurrent_read | libb.so.2 | .....
| concurrent_read | libb.so.2 | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | concurrent_read_benchmark | .....
| concurrent_read | Concurrent_read_benchmark | .....

---

Cumulative Event Distribution
PERF VISUALIZATION

We can also use `perf` to visualize the generated profile for our application.

$ perf report

There are also third-party visualization tools:
→ Hotspot
We can also use `perf` to visualize the generated profile for our application. There are also third-party visualization tools:

$ perf report$
We can also use `perf` to visualize the generated profile for our application. There are also third-party visualization tools: `Hotspot`.
PERF EVENTS

Supports several other events like:
→ L1-dcache-load-misses
→ branch-misses

To see a list of events:

$ perf list

Another usage example:

$ perf record -e cycles,LLC-load-misses -c 2000
./relwithdebinfo/concurrent_read_benchmark
REFERENCES

Valgrind
→ The Valgrind Quick Start Guide
→ Callgrind
→ Kcache
gind
→ Tips for the Profiling/Optimization process

Perf
→ Perf Tutorial
→ Perf Examples
→ Perf Analysis Tools
Next Class

- Index Key Representation
- Memory Allocation & Garbage Collection
- T-Trees (1980s / TimesTen)
- Bw-Tree (Hekaton)
- Concurrent Skip Lists (MemSQL)