Ack: Prof. Jignesh Patel @ CMU
Prof. Andy Pavlo @CMU

CSC3170
9: Index Concurrency Control

Chenhao Ma
School of Data Science
The Chinese University of Hong Kong, Shenzhen

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.

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!

VOLTDB

redis

H-Store

SCHOOL OF DATA SCIENCE

Concurrency Control

- 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?

Concurrency Control

- 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?

This Lecture

- Latches Overview
- Hash Table Latching
- B+Tree Latching
- Leaf Node Scans

SCHOOL OF
DATA SCIENCE

Latches

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

	Locks	***Latches***
Separate…	Transactions	Workers (threads, processes)
Protect…	Database Contents	In-Memory Data Structures
During…	Entire Transactions	Critical Sections
Modes…	Shared, Exclusive, Update, Intention	Read, Write
Deadlock	Detection & Resolution	Avoidance
…by…	Waits-for, Timeout, Aborts	Coding Discipline
Kept in…	Lock Manager	Protected Data Structure

Source: Goetz Graefe

SCHOOL OF DATA SCIENCE

Locks vs. Latches

	Locks	*Latches*
Separate…	Transactions	Workers (threads, processes)
Protect…	Database Contents	In-Memory Data Structures
During…	Entire Transactions	Critical Sections
Modes…	Shared, Exclusive, Update, Intention	Read, Write
Deadlock	Detection & Resolution	Avoidance
…by…	Waits-for, Timeout, Aborts	Coding Discipline
Kept in…	Lock Manager	Protected Data Structure

Source: Goetz Graefe

SCHOOL OF DATA SCIENCE

Locks vs. Latches

Future Lecture

	Locks	*Latches*
Separate…	Transactions	Workers (threads, processes)
Protect…	Database Contents	In-Memory Data Structures
During…	Entire Transactions	Critical Sections
Modes…	Shared, Exclusive, Update, Intention	Read, Write
Deadlock	Detection & Resolution	Avoidance
…by…	Waits-for, Timeout, Aborts	Coding Discipline
Kept in…	Lock Manager	Protected Data Structure

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.

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

	Read	Write
Read	✓	X
Write	X	X

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

Latch Implementations

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

- Advanced approaches:
  - Adaptive Spinlock (Apple ParkingLot)
  - Queue-based Spinlock (MCS Locks)

Latch Implementations

- **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>`

Latch Implementations

- **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>`

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

SCHOOL OF DATA SCIENCE

Latch Implementations

- **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>`

std::atomic<bool>

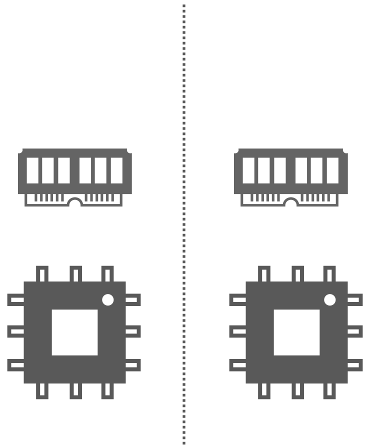
```
std::atomic_flag latch;
 ⋮
while (latch.test_and_set(…)) {
   // Retry? Yield? Abort?
}
```

Latch Implementations

- **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>`

std::atomic<bool>

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

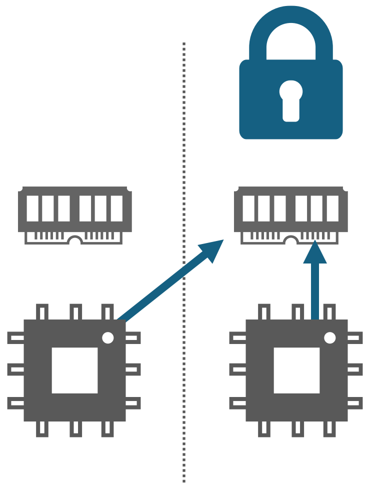


Latch Implementations

- **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>`

std::atomic<bool>

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



Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex`

Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex`

```
std::mutex m;
 ⋮
m.lock();
// Do something special...
m.unlock();
```

SCHOOL OF DATA SCIENCE

Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex` ⟶ `pthread_mutex`

```
std::mutex m;
 ⋮
m.lock();
// Do something special...
m.unlock();
```

Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex` ⟶ `pthread_mutex` ⟶ `futex`

```
std::mutex m;
 ⋮
m.lock();
// Do something special...
m.unlock();
```

Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex` ⟶ `pthread_mutex` ⟶ `futex`

```
std::mutex m;
 ⋮
m.lock();
// Do something special...
m.unlock();
```

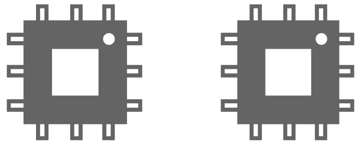
🔒 *OS Latch*
🔒 *Userspace Latch*

Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex` ⟶ `pthread_mutex` ⟶ `futex`

```
std::mutex m;
 ⋮
m.lock();
// Do something special...
m.unlock();
```

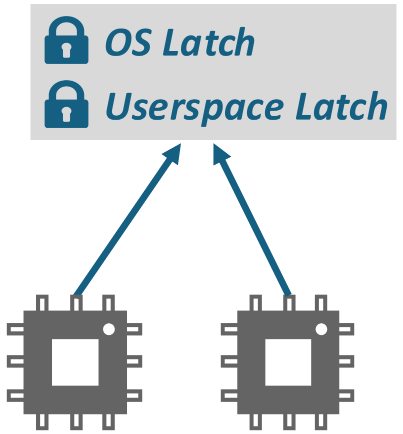
🔒 *OS Latch*
🔒 *Userspace Latch*



Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex` ⟶ `pthread_mutex` ⟶ `futex`

```
std::mutex m;
 ⋮
m.lock();
// Do something special...
m.unlock();
```

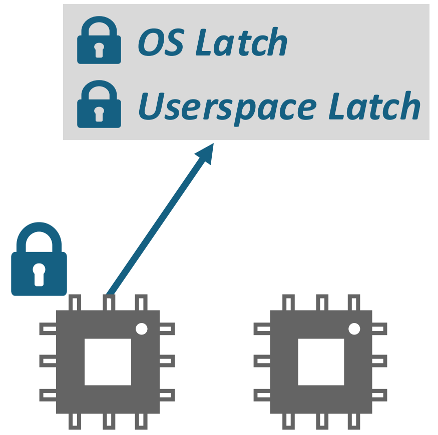


SCHOOL OF DATA SCIENCE

Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: **`std::mutex`** ⟶ `pthread_mutex` ⟶ `futex`

```
std::mutex m;
 ⋮
m.lock();
// Do something special...
m.unlock();
```

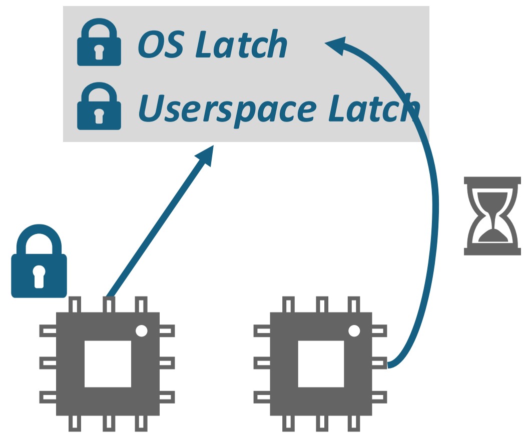


SCHOOL OF DATA SCIENCE

Latch Implementations

- **Approach #2: Blocking OS Mutex**
  - Simple to use
  - Non-scalable (about 25ns per lock/unlock invocation)
  - Example: `std::mutex` ⟶ `pthread_mutex` ⟶ `futex`

```
std::mutex m;
  ⋮
m.lock();
// Do something special...
m.unlock();
```



Latch Implementations

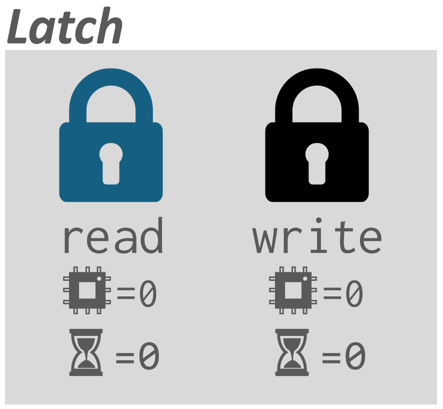
- **Approach #3: Reader-Writer Latches**
  - Allows for concurrent readers. Must manage read/write queues to avoid starvation.
  - Can be implemented on top of spinlocks.
  - Example: `std::shared_mutex`

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` $\longrightarrow$ `pthread_rwlock`

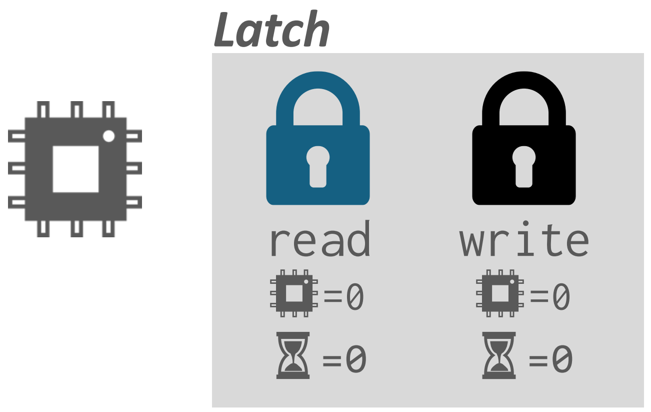
Latch Implementations

- **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` $\longrightarrow$ `pthread_rwlock`



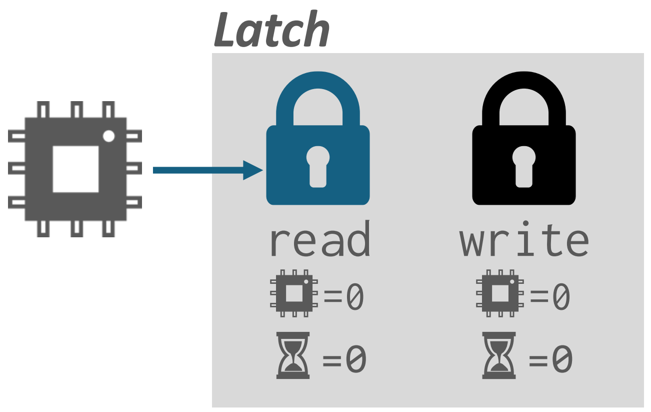
Latch Implementations

- **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` $\longrightarrow$ `pthread_rwlock`



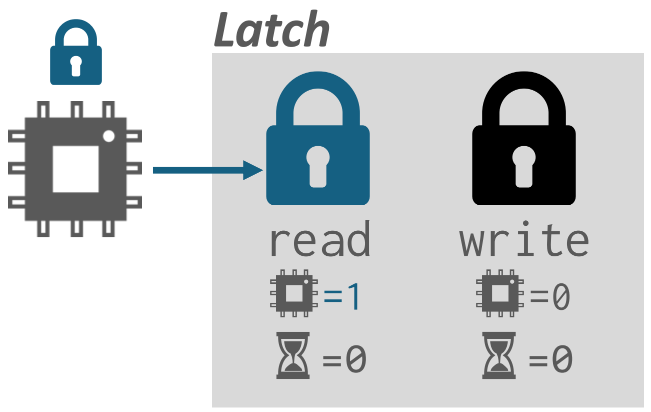
Latch Implementations

- **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`



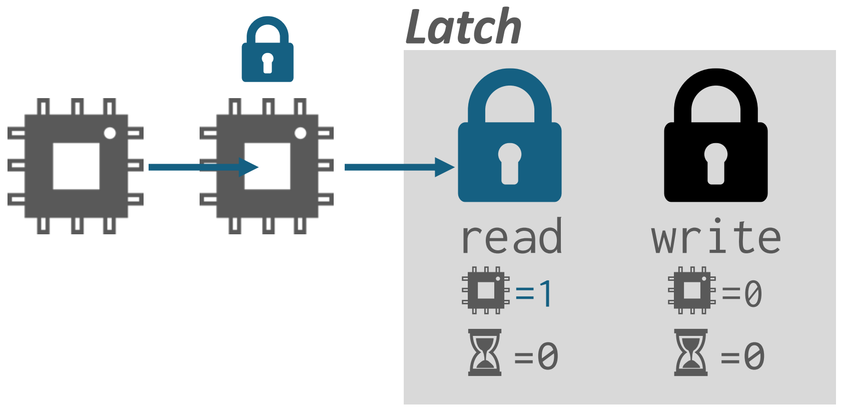
Latch Implementations

- **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`



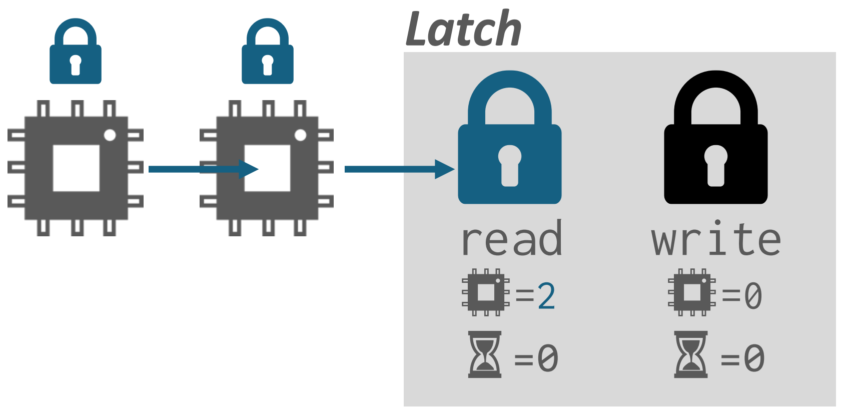
Latch Implementations

- **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`



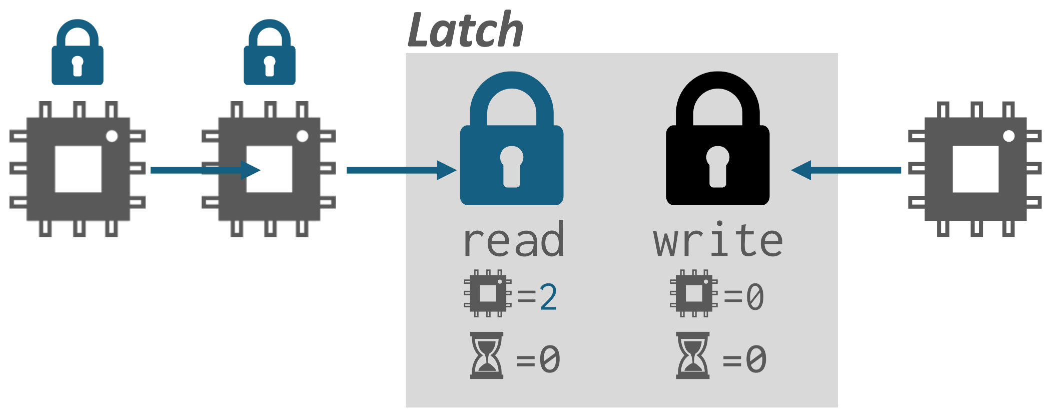
Latch Implementations

- **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`



Latch Implementations

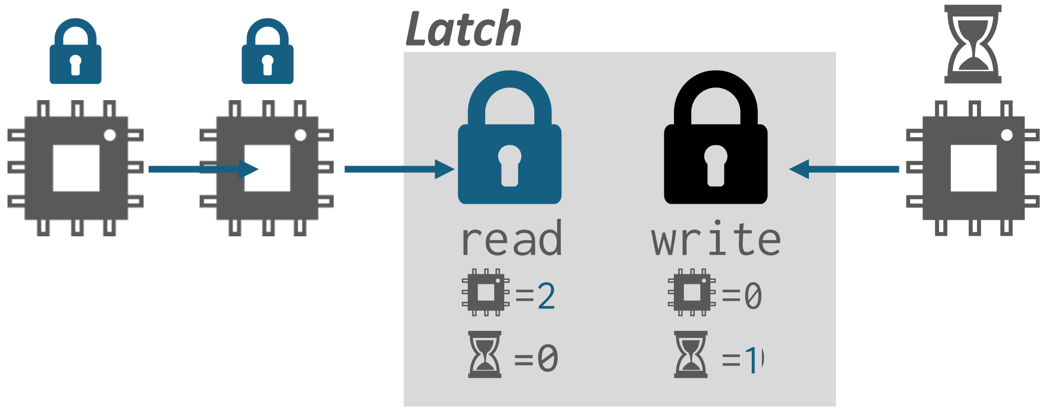
- **Approach #3: Reader-Writer Latches**
  - Allows for concurrent readers. Must manage read/write queues to avoid starvation.
  - Can be implemented on top of spinlocks.
  - Example: `std::shared_mutex` ⟶ `pthread_rwlock`



SCHOOL OF DATA SCIENCE

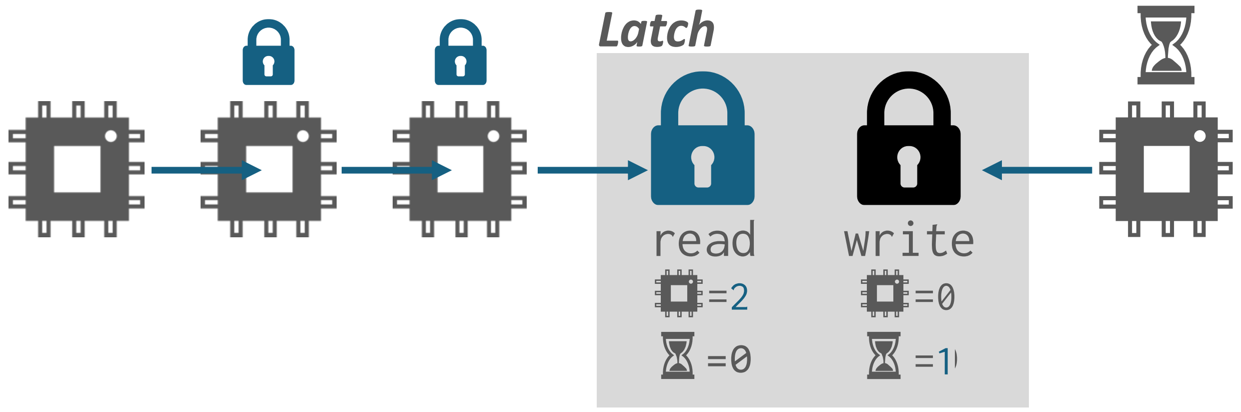
Latch Implementations

- **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`



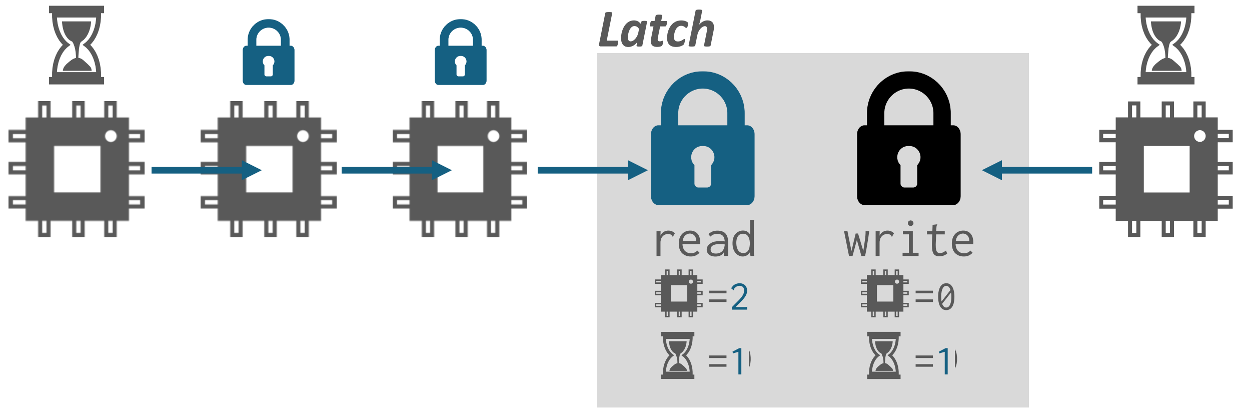
Latch Implementations

- **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`



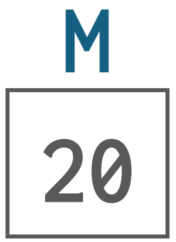
Latch Implementations

- **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`



Compare-and-Swap

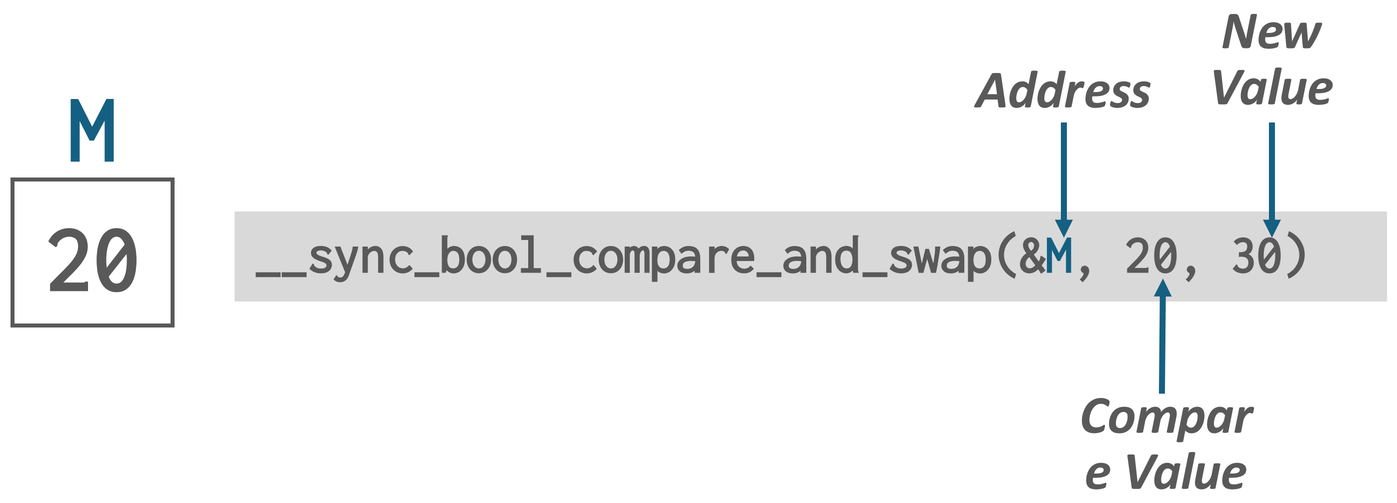
- Atomic instruction that compares contents of a memory location `M` to a given value `V`
  - If values are equal, installs new given value `V’` in `M`
  - Otherwise, operation fails



```
__sync_bool_compare_and_swap(&M, 20, 30)
```

Compare-and-Swap

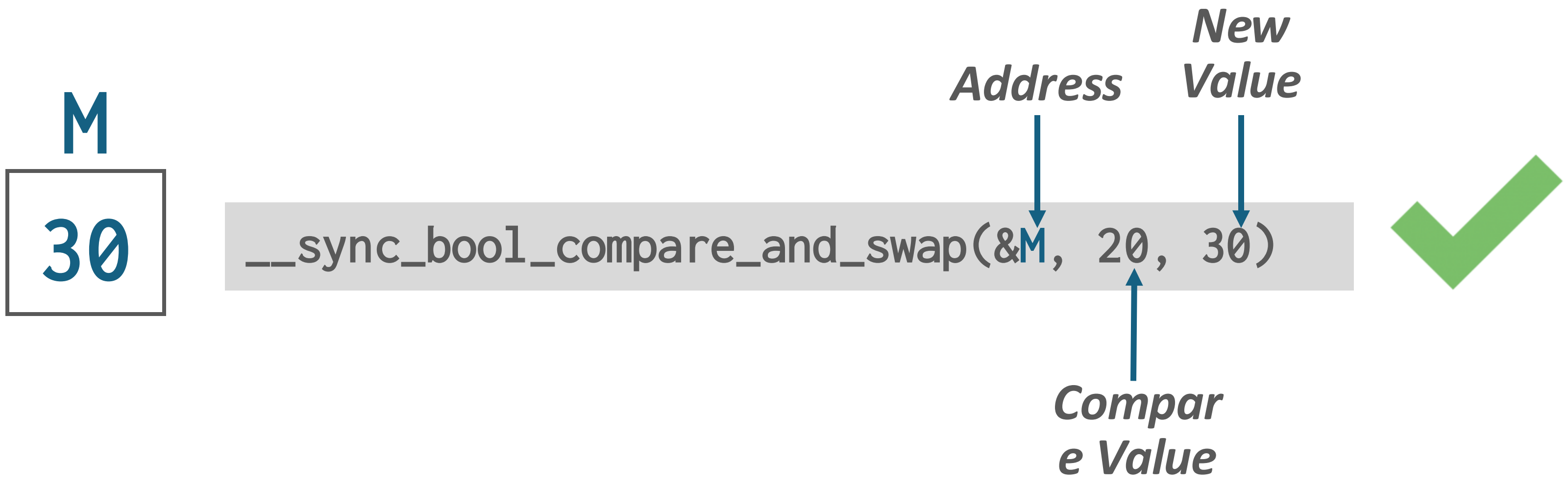
- Atomic instruction that compares contents of a memory location `M` to a given value `V`
  - If values are equal, installs new given value `V'` in `M`
  - Otherwise, operation fails



```
__sync_bool_compare_and_swap(&M, 20, 30)
```

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)
```

Hash Table Latching

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.

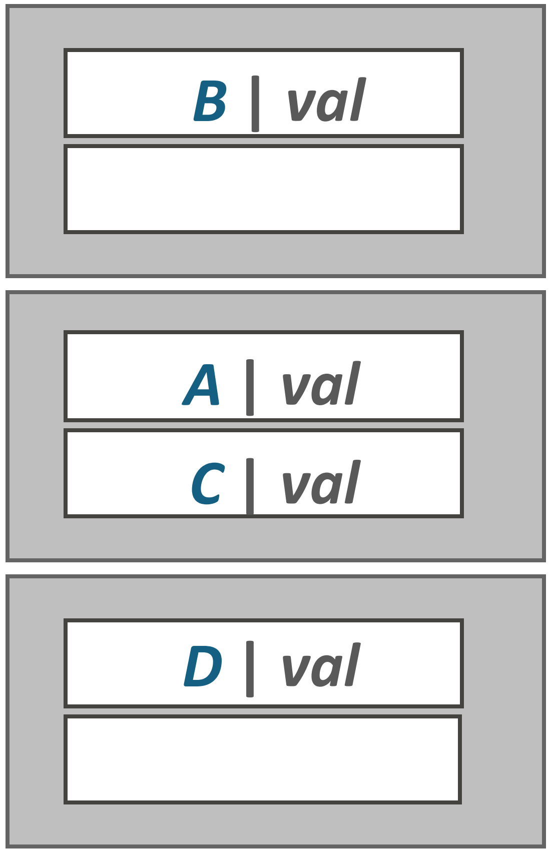
- To resize the table, take a global write latch on the entire table (e.g., in the header page).

Hash Table Latching

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

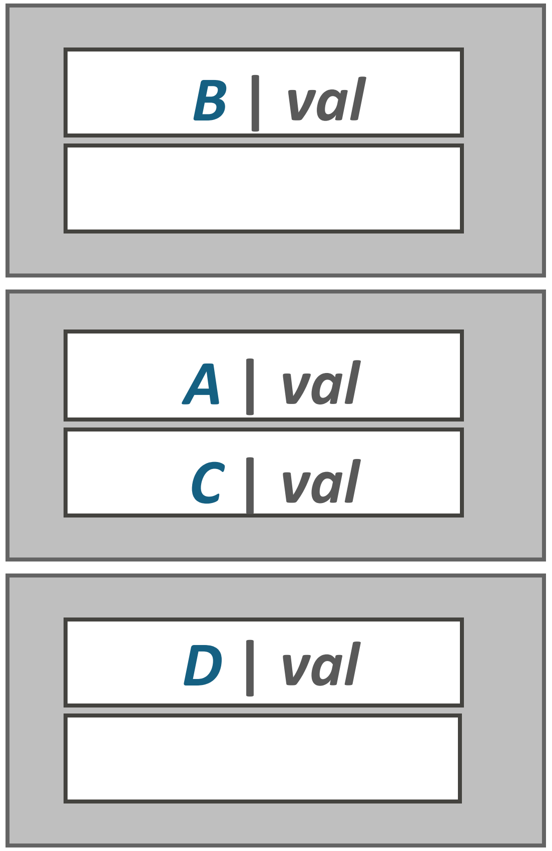
- **Approach #2: Slot Latches**
  - Each slot has its own latch.
  - Can use a single-mode latch to reduce meta-data and computational overhead.

Hash Table - Page/Block Latches

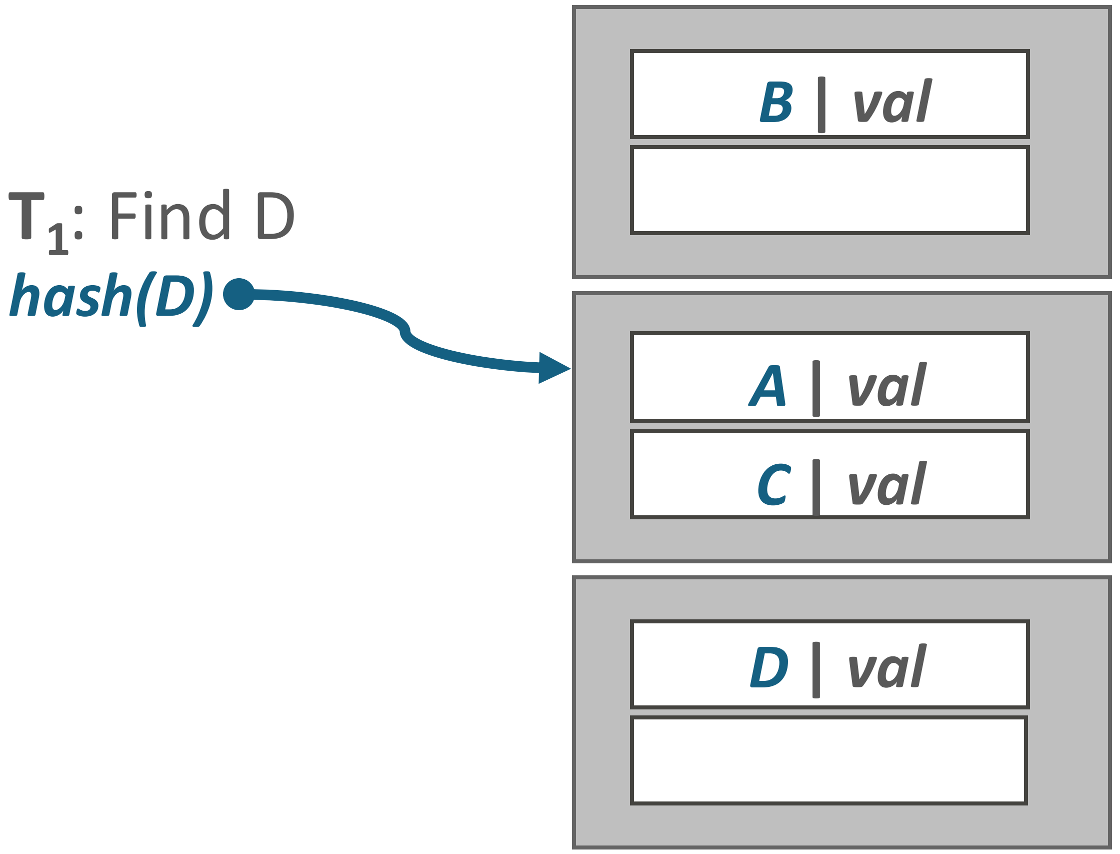


Hash Table - Page/Block Latches

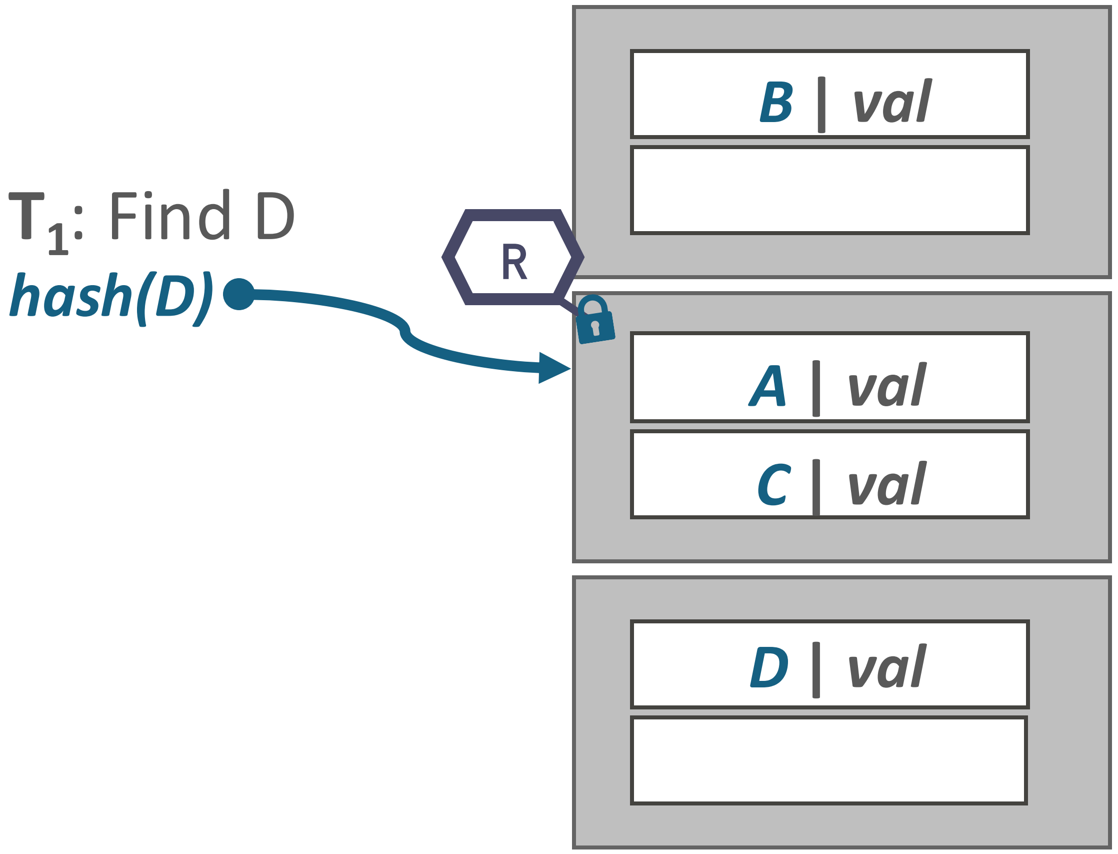
T$_1$: Find D
hash(D)



Hash Table - Page/Block Latches

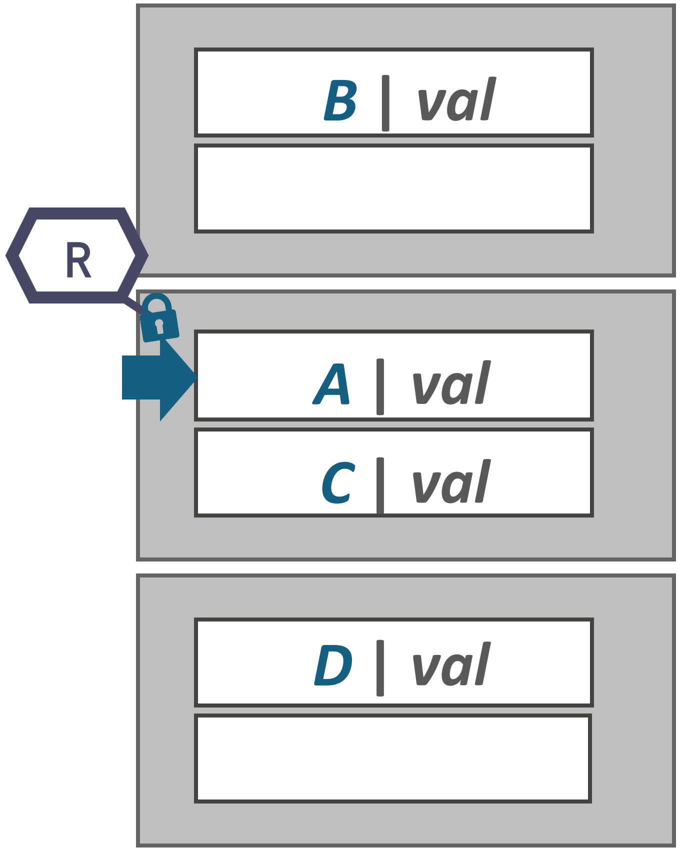


Hash Table - Page/Block Latches



Hash Table - Page/Block Latches

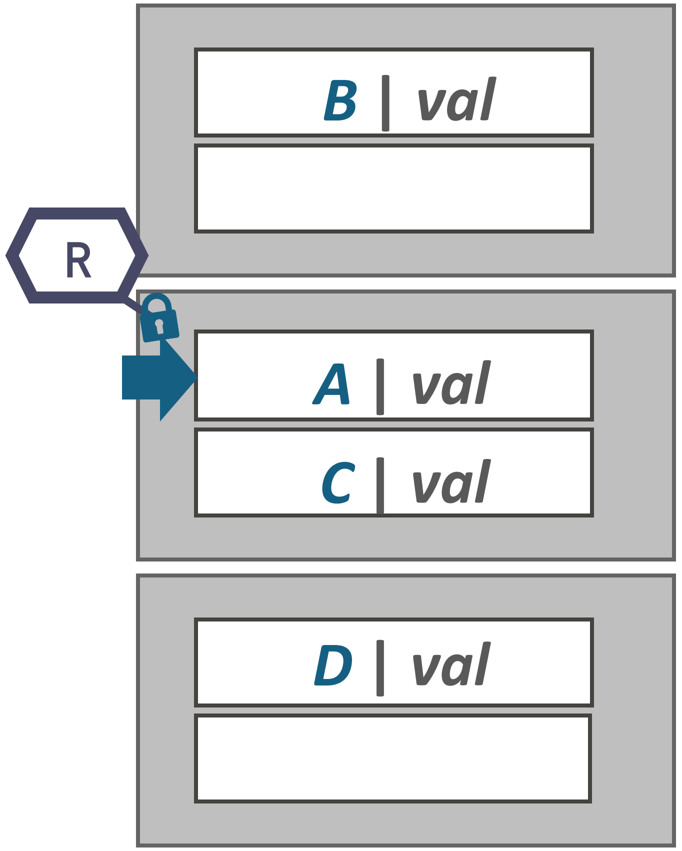
T_1: Find D
hash(D)



SCHOOL OF DATA SCIENCE

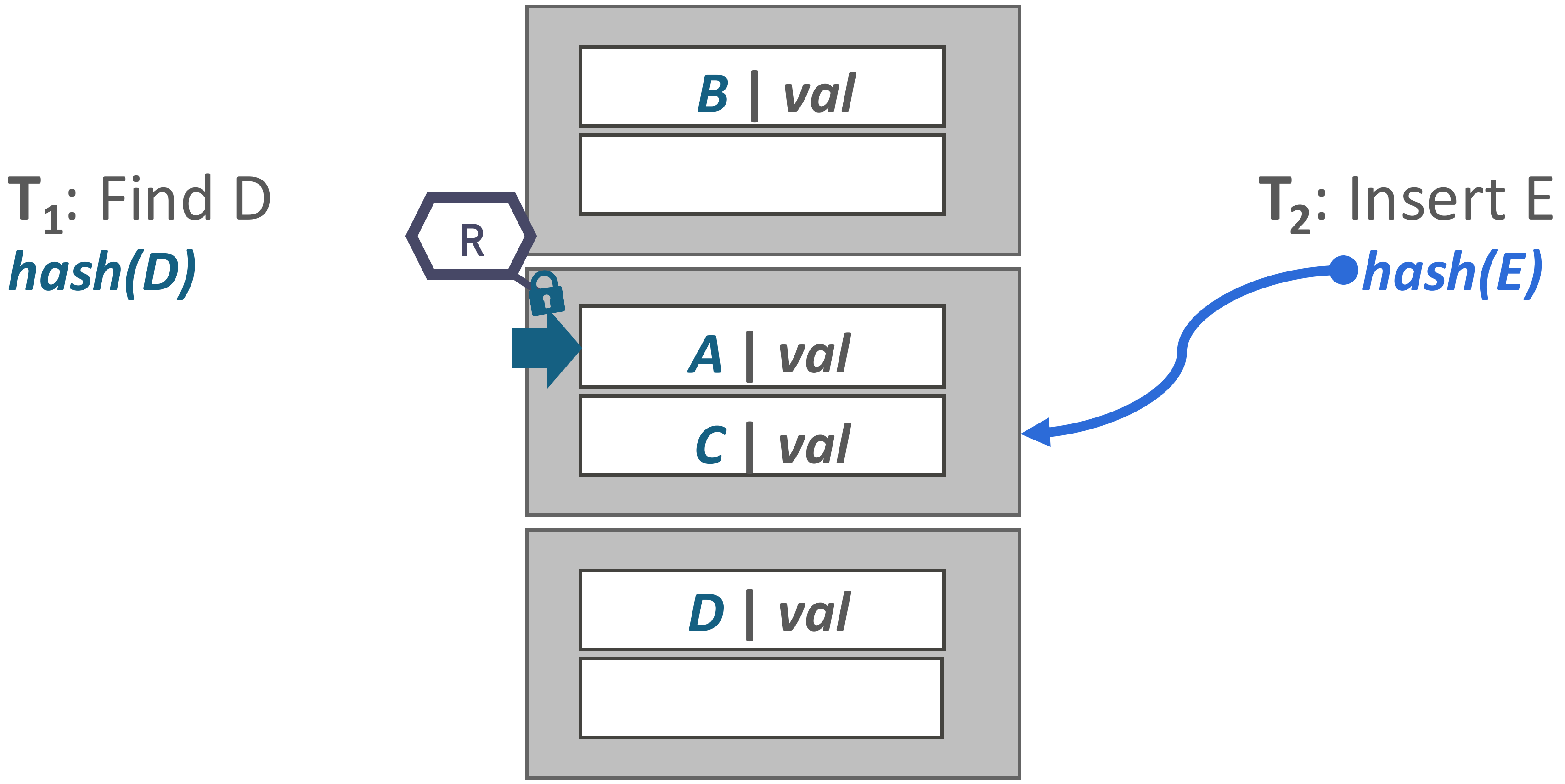
Hash Table - Page/Block Latches

T$_1$: Find D
hash(D)

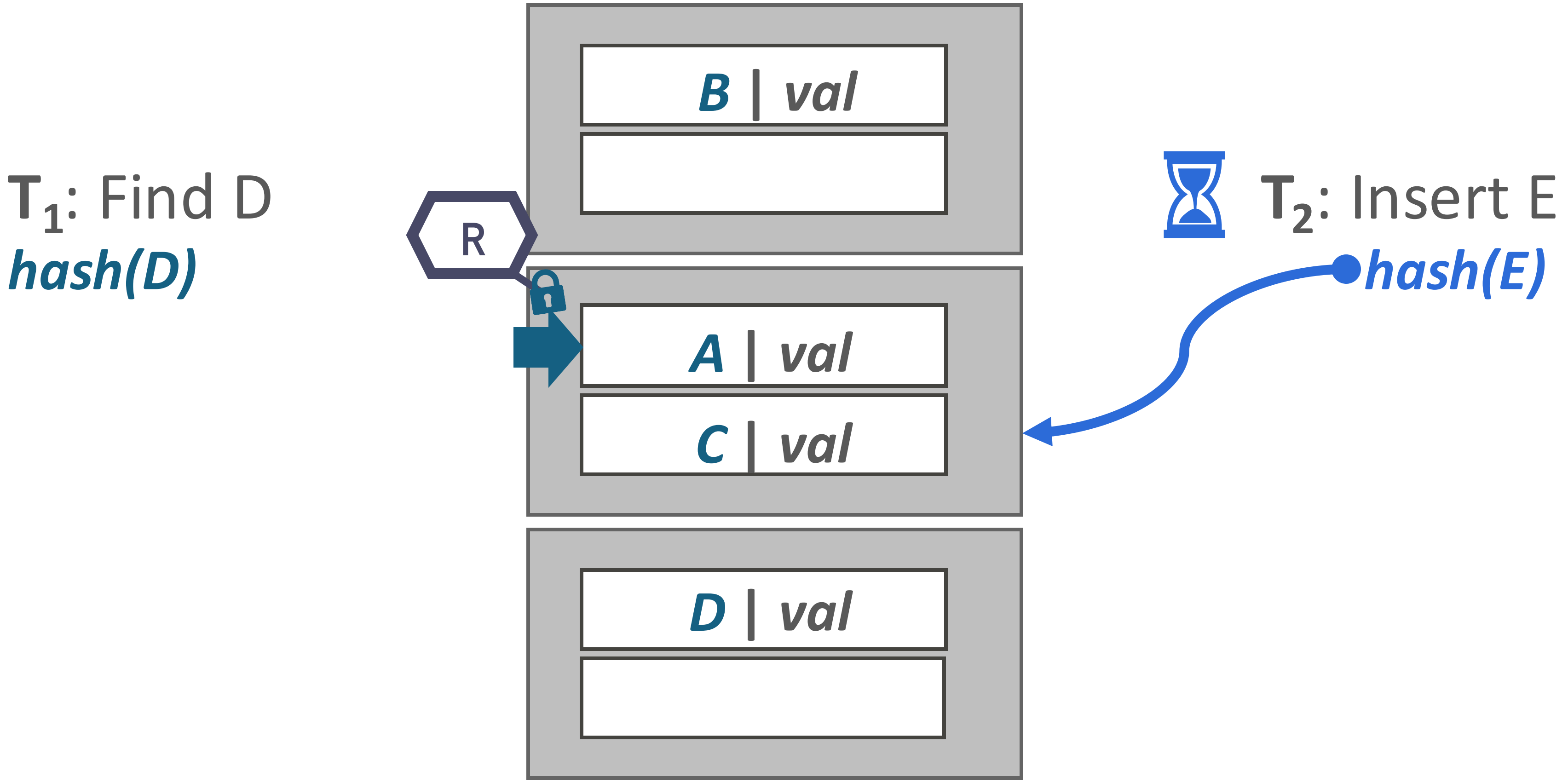


T$_2$: Insert E
hash(E)

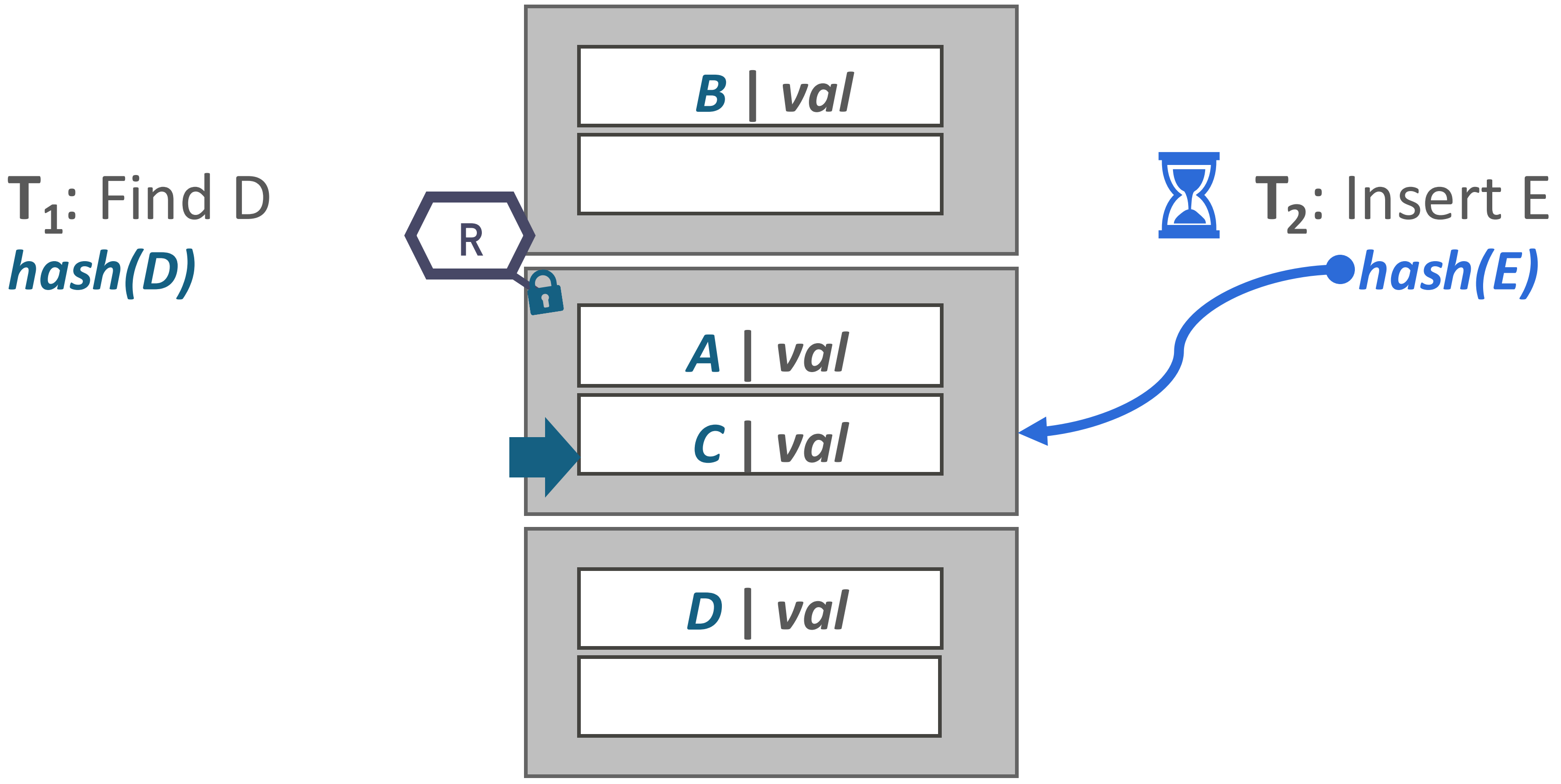
Hash Table - Page/Block Latches



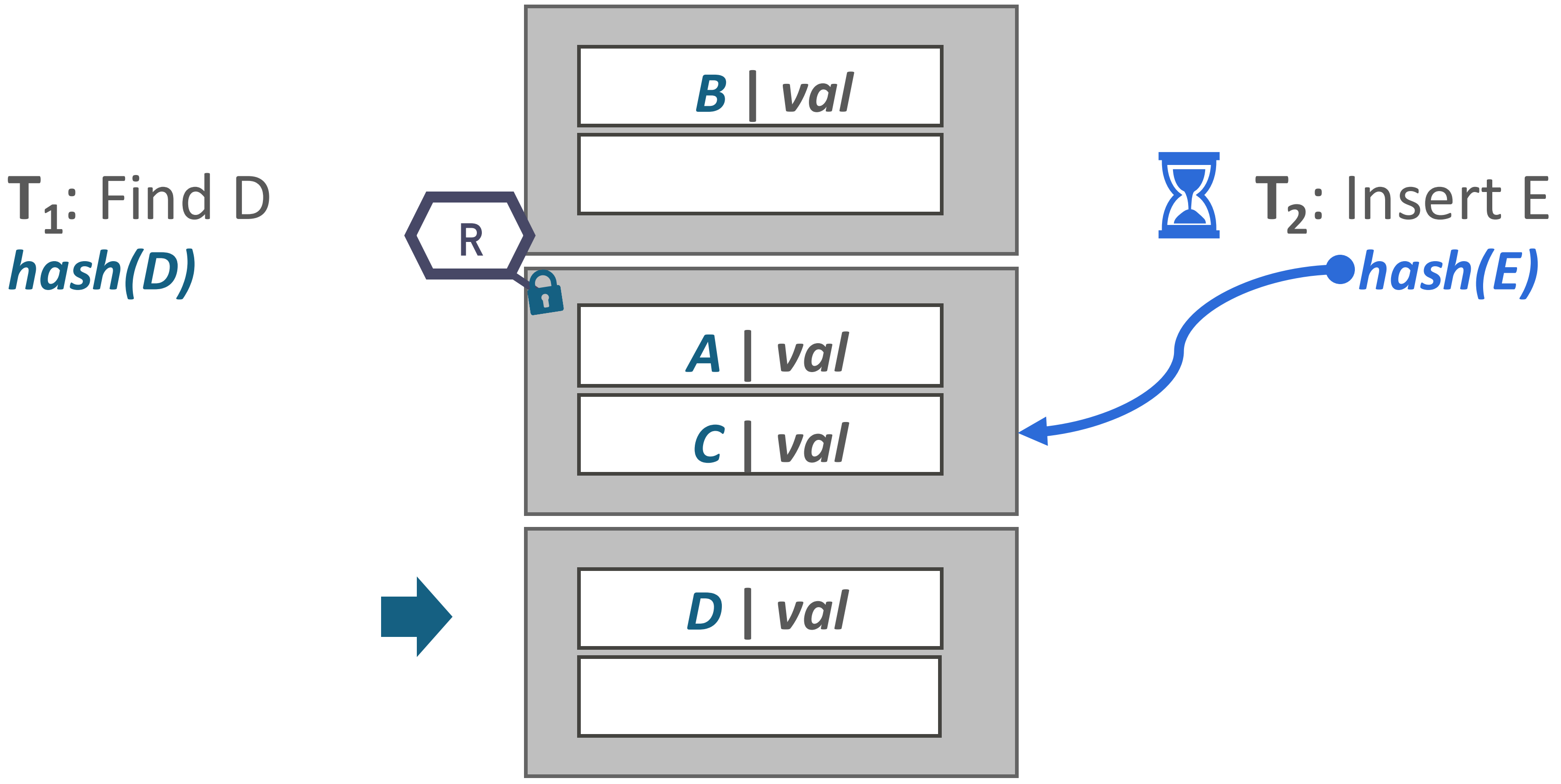
Hash Table - Page/Block Latches



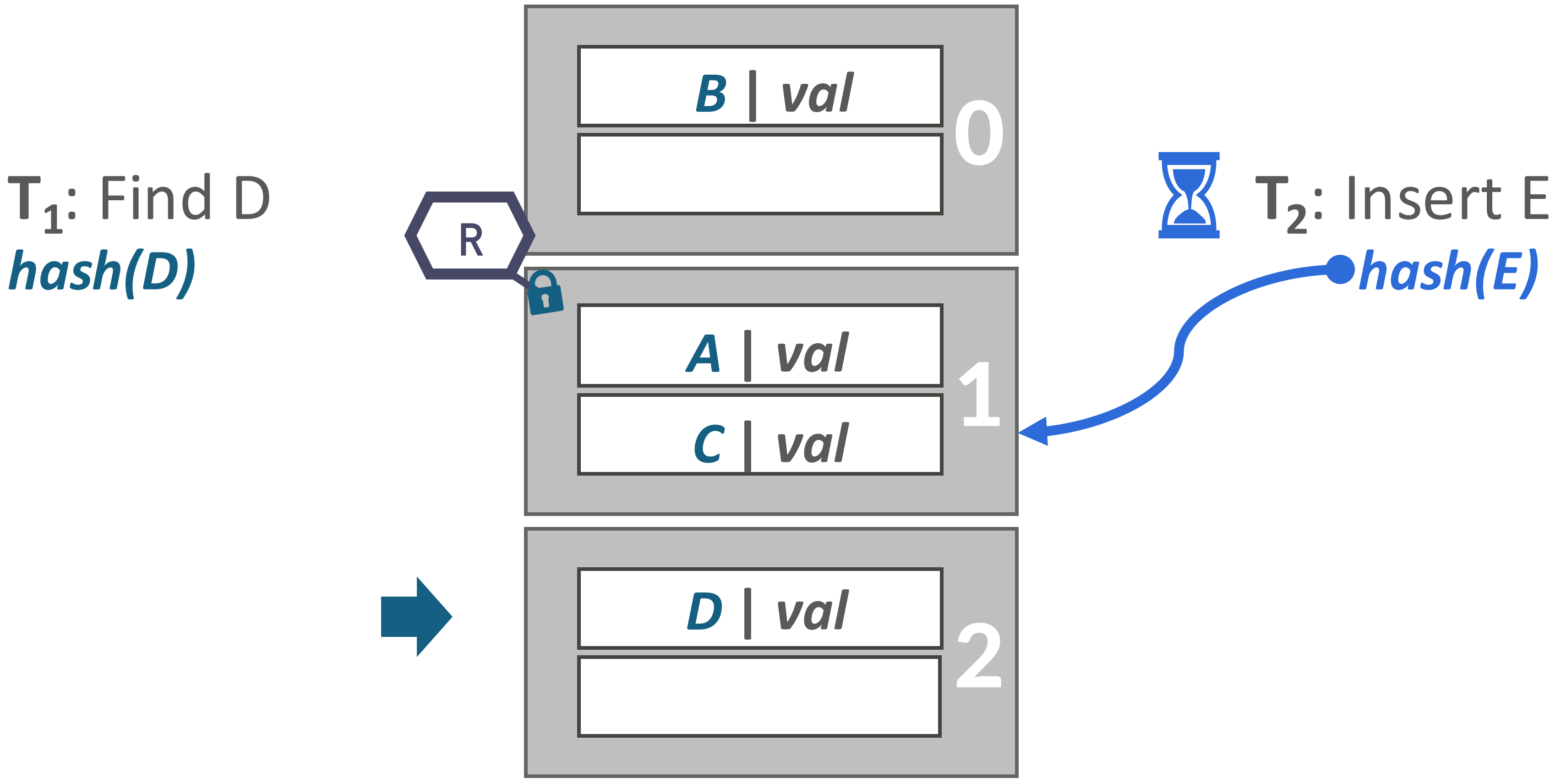
Hash Table - Page/Block Latches



Hash Table - Page/Block Latches

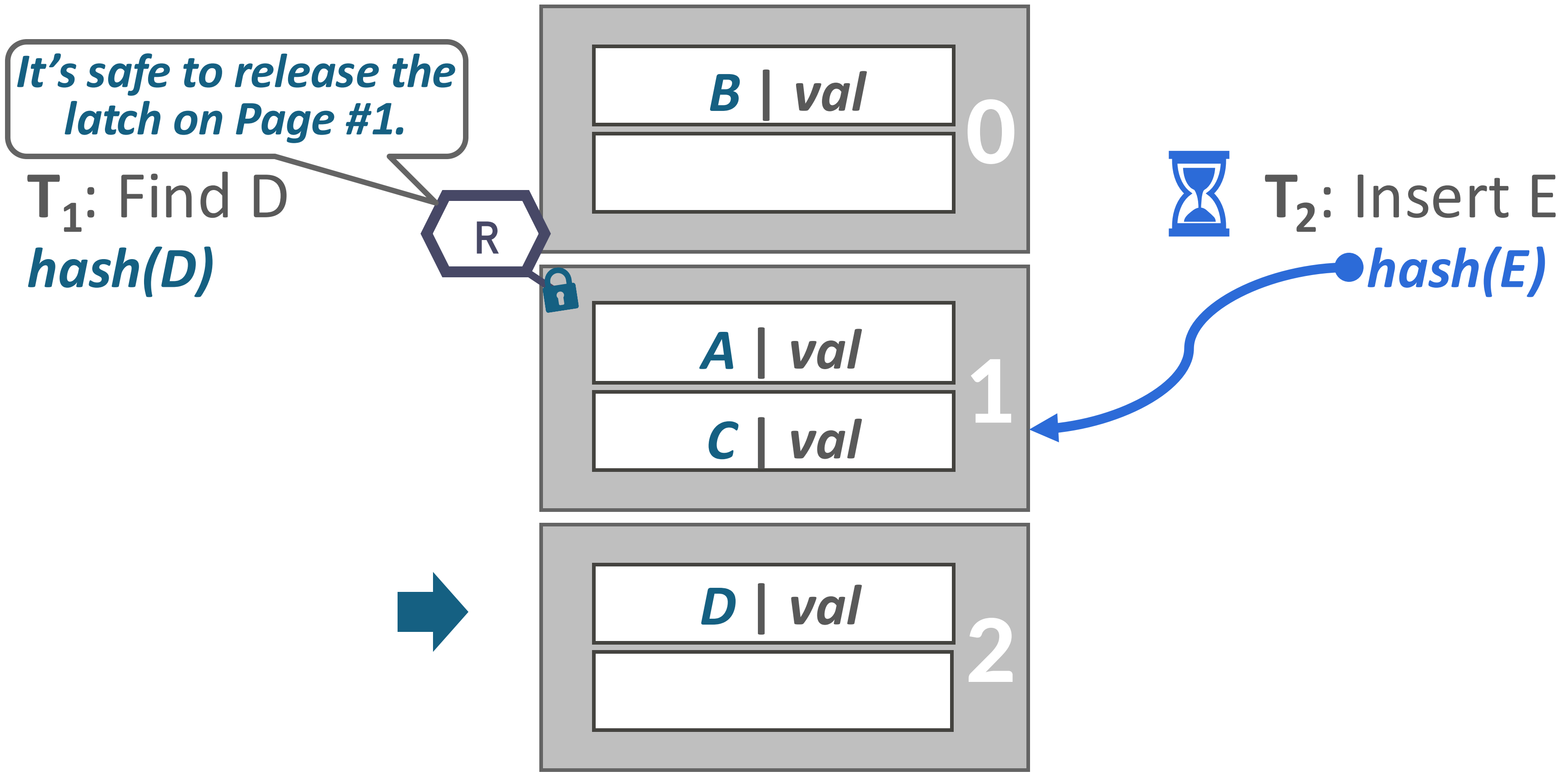


Hash Table - Page/Block Latches

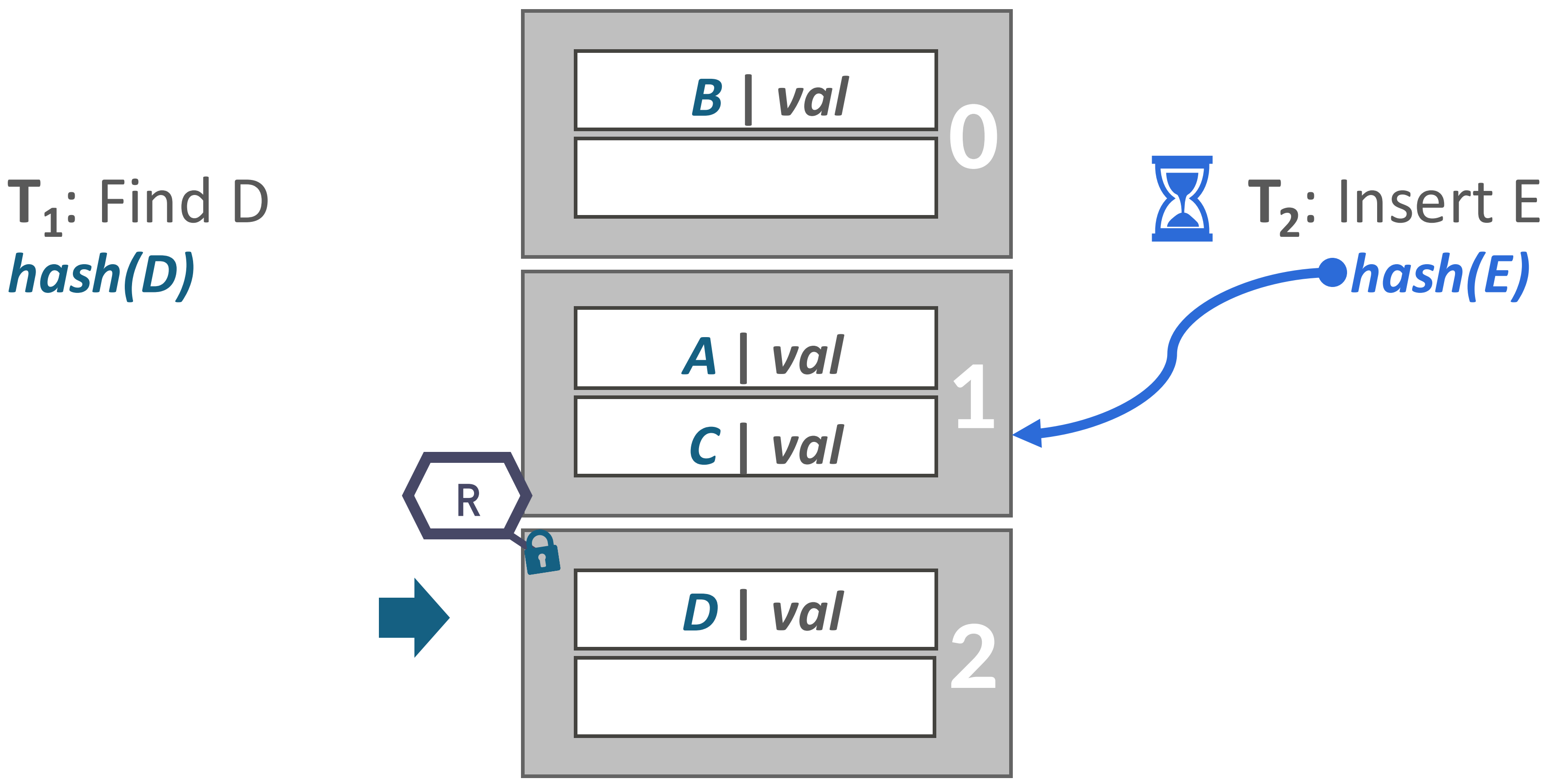


SCHOOL OF DATA SCIENCE

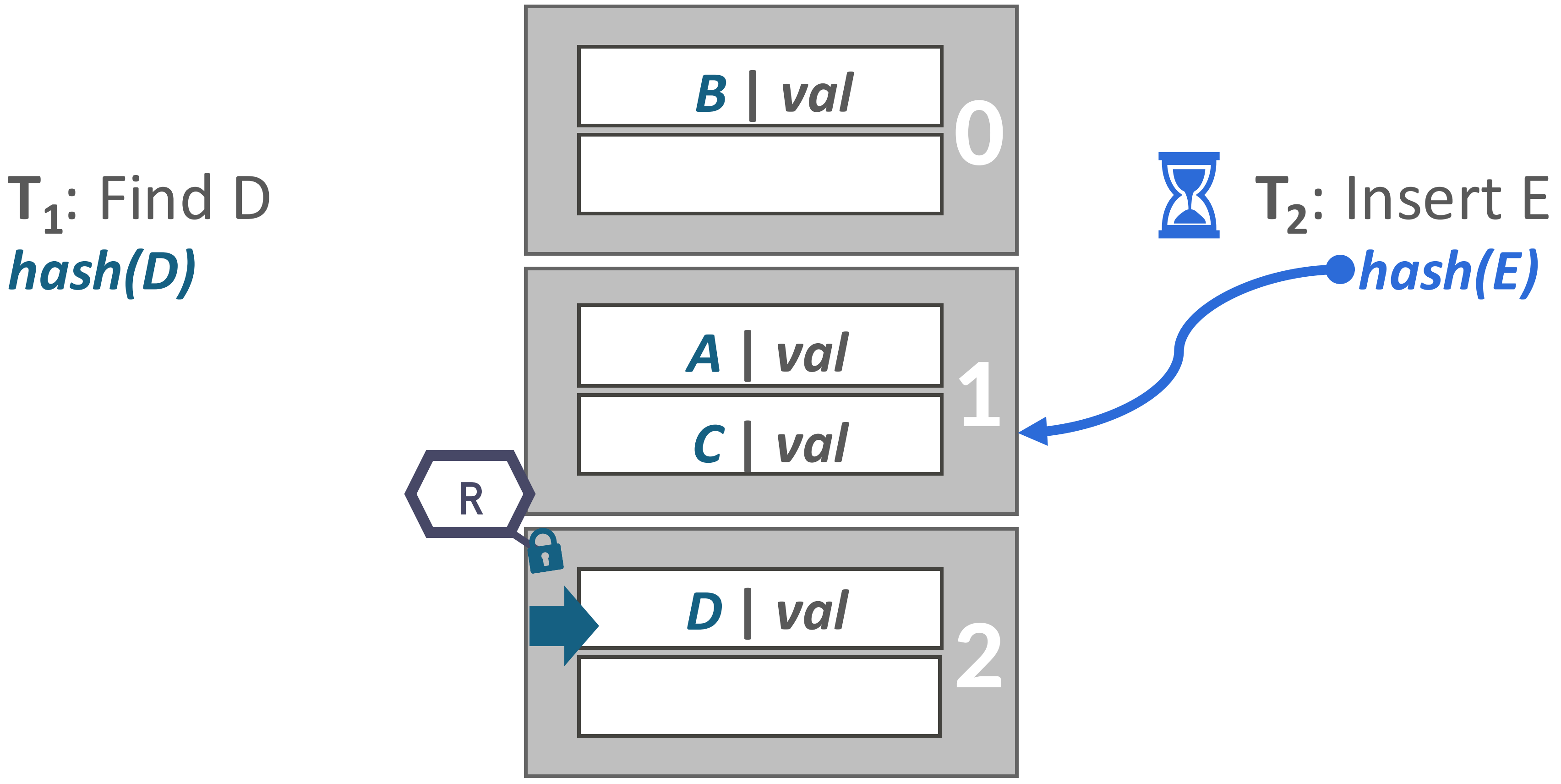
Hash Table - Page/Block Latches



Hash Table - Page/Block Latches

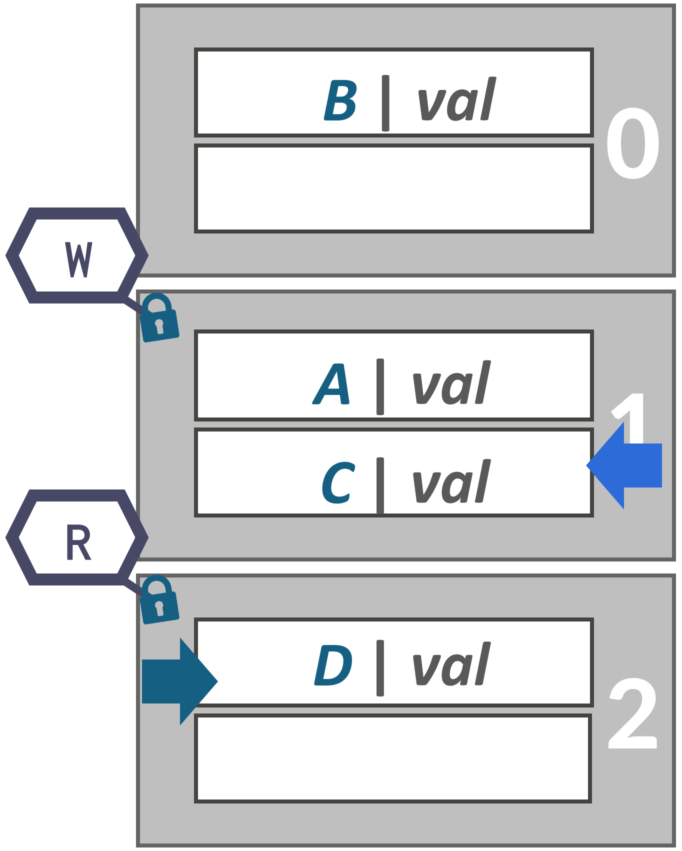


Hash Table - Page/Block Latches



Hash Table - Page/Block Latches

T_1: Find D
hash(D)

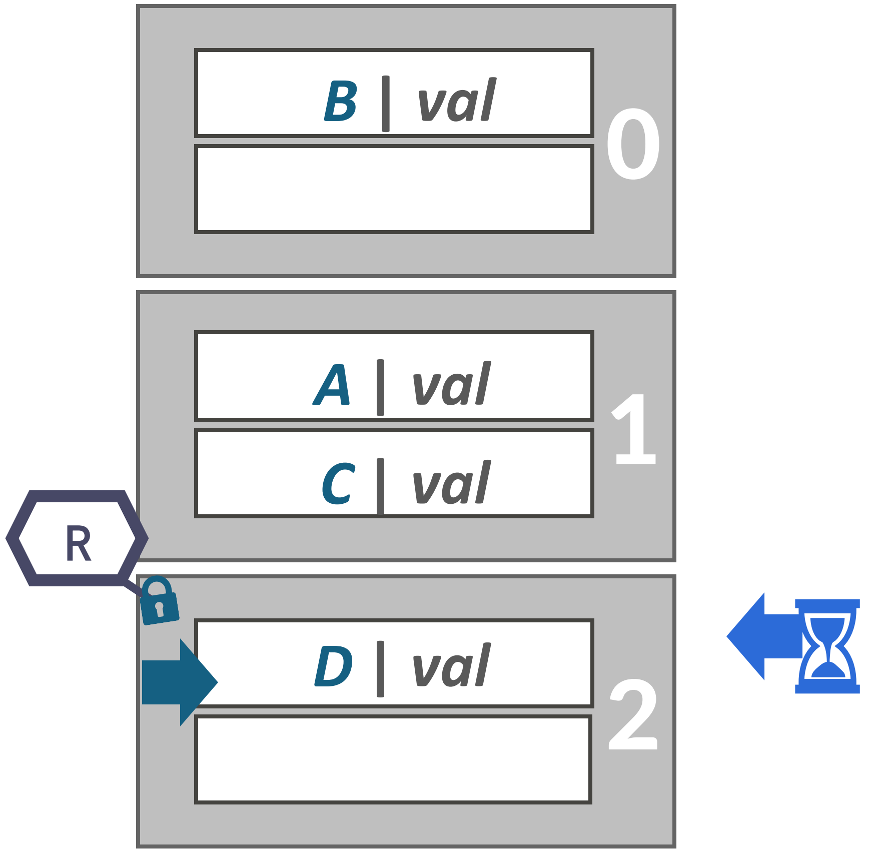


T_2: Insert E
hash(E)

SCHOOL OF DATA SCIENCE

Hash Table - Page/Block Latches

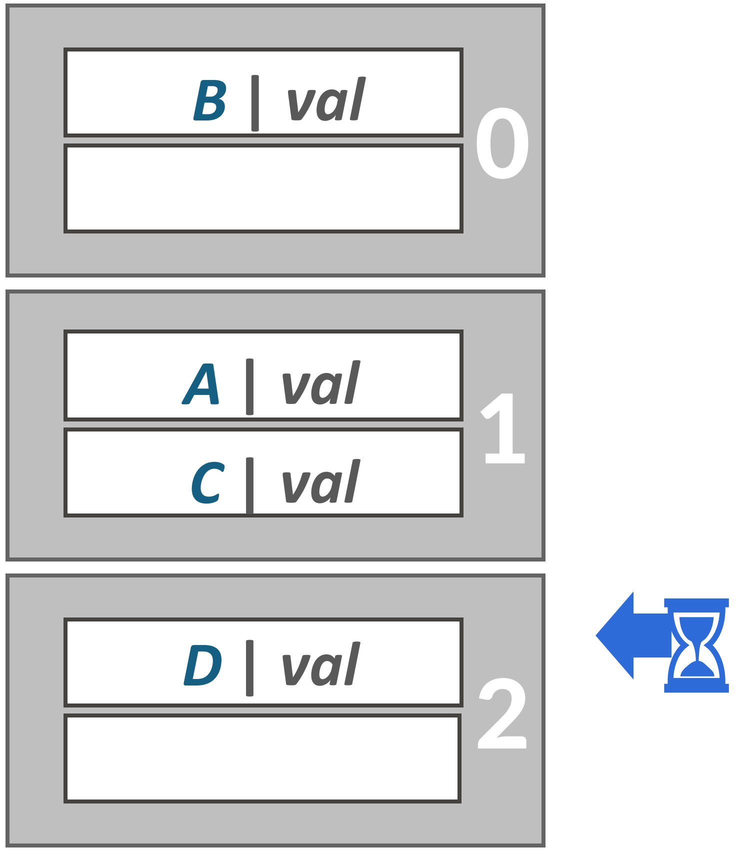
T_1: Find D
hash(D)



T_2: Insert E
hash(E)

Hash Table - Page/Block Latches

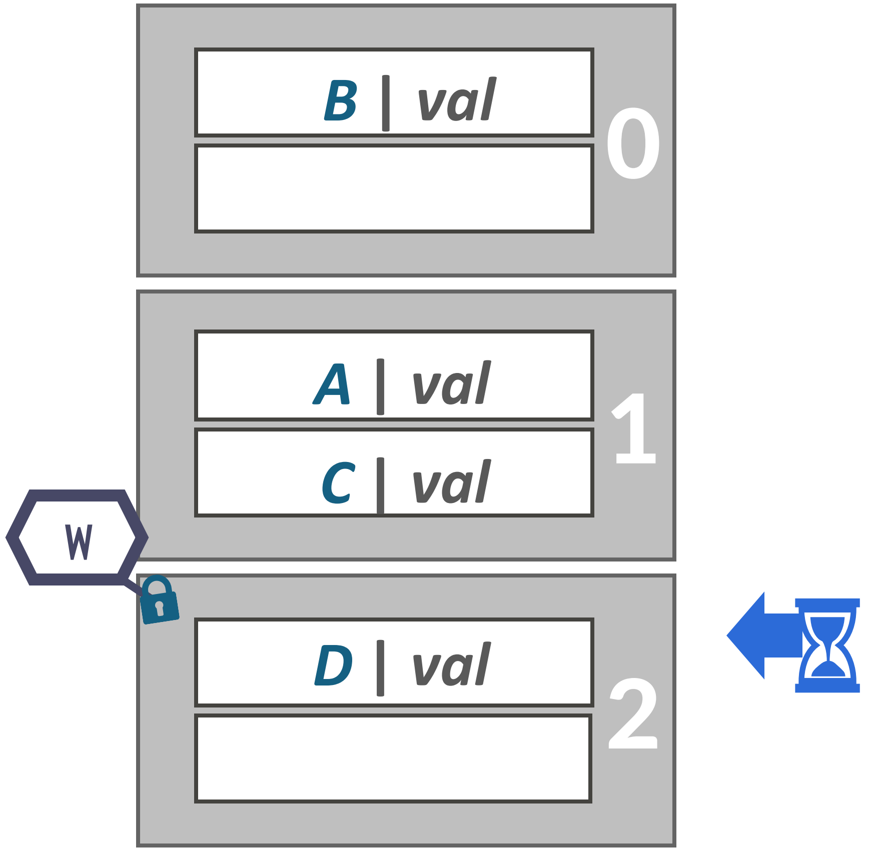
T_1: Find D
hash(D)



T_2: Insert E
hash(E)

Hash Table - Page/Block Latches

T$_1$: Find D
hash(D)

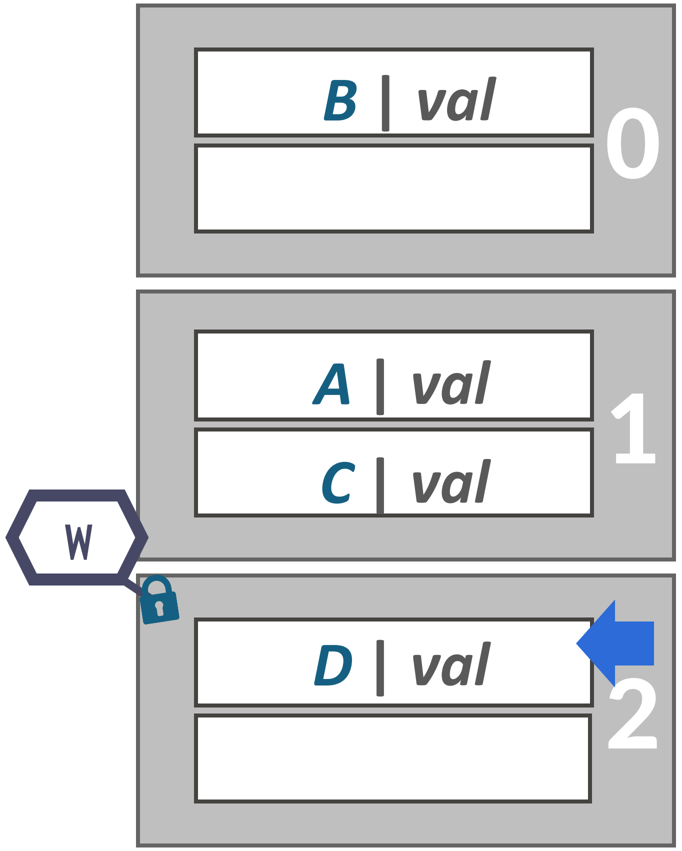


T$_2$: Insert E
hash(E)

Hash Table - Page/Block Latches

T$_1$: Find D
hash(D)

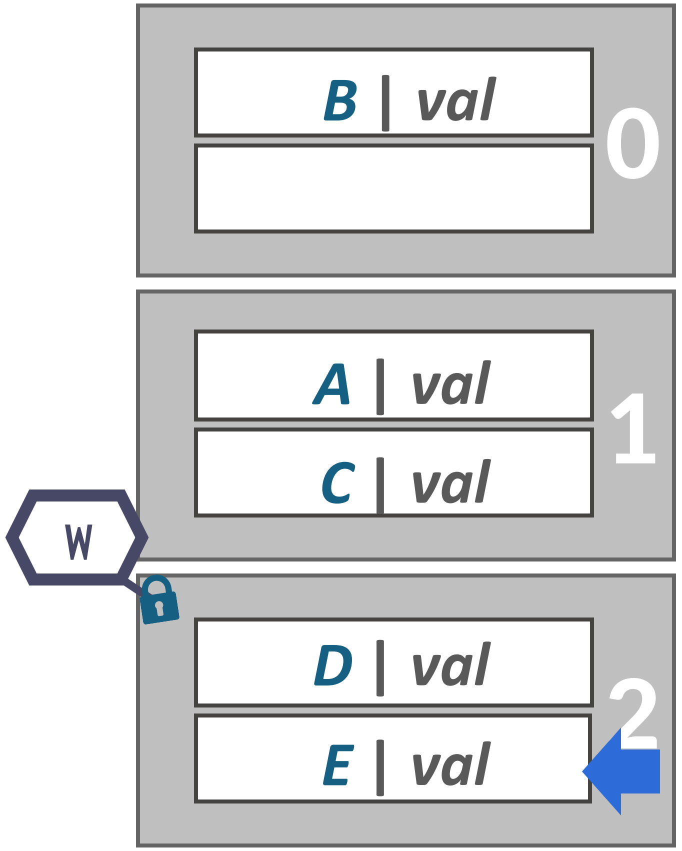
T$_2$: Insert E
hash(E)



SCHOOL OF DATA SCIENCE

Hash Table - Page/Block Latches

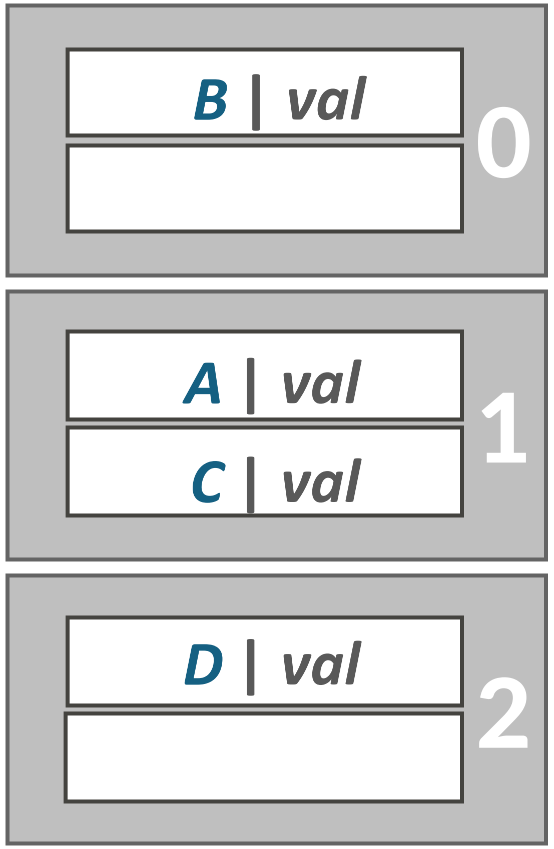
T_1: Find D
hash(D)



T_2: Insert E
hash(E)

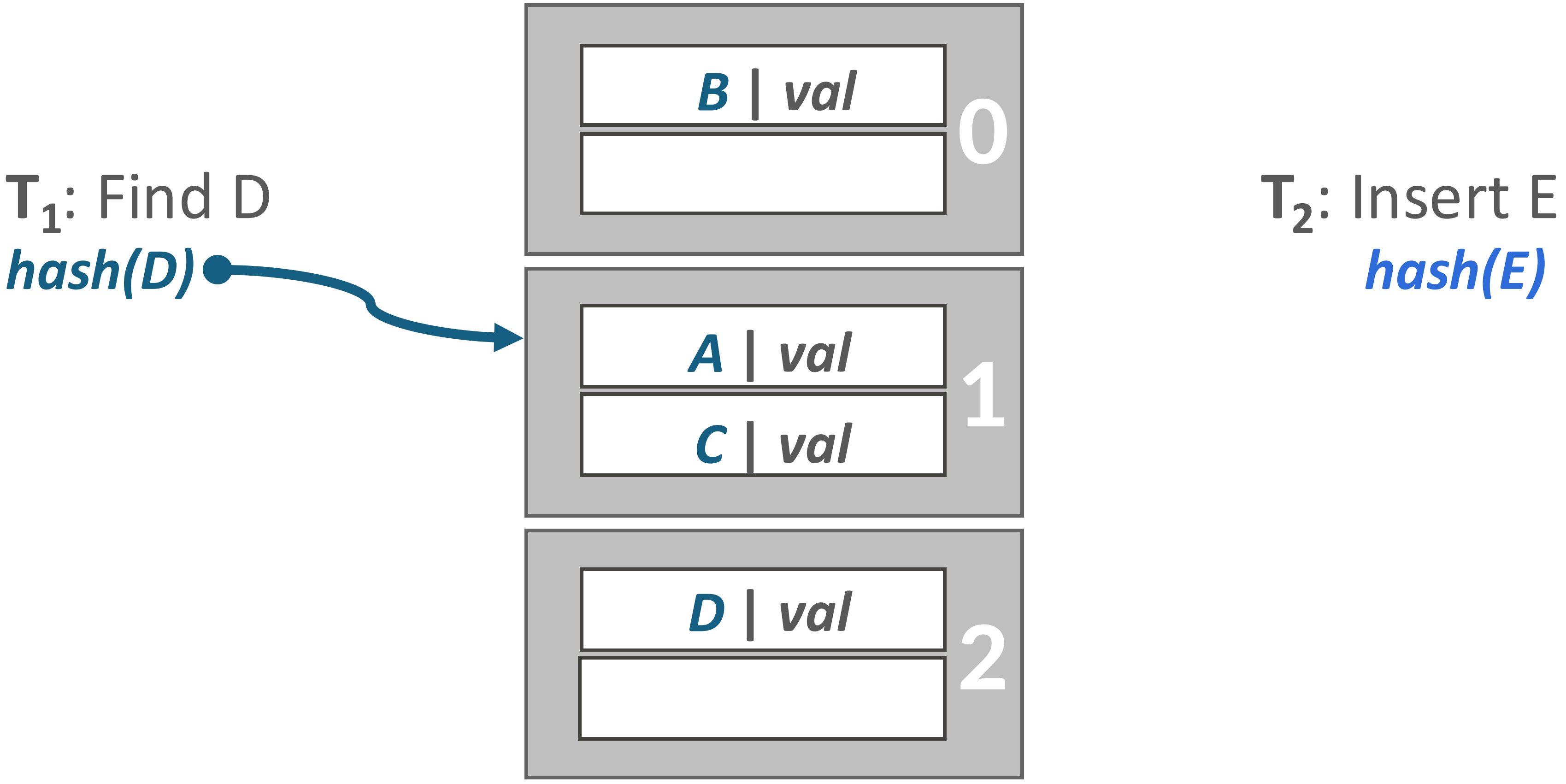
Hash Table – Slot Latches

T_1: Find D
hash(D)

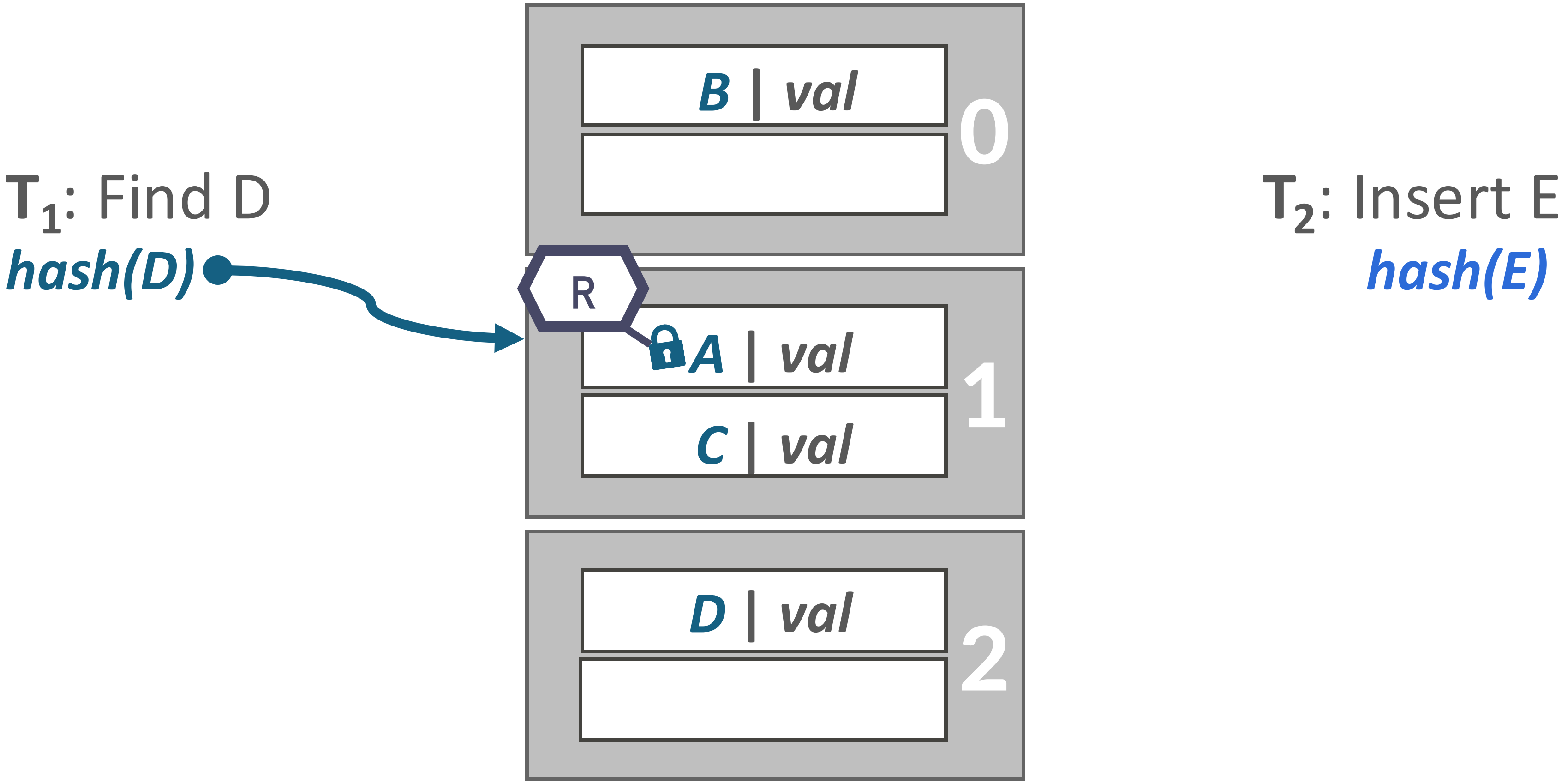


T_2: Insert E
hash(E)

Hash Table – Slot Latches

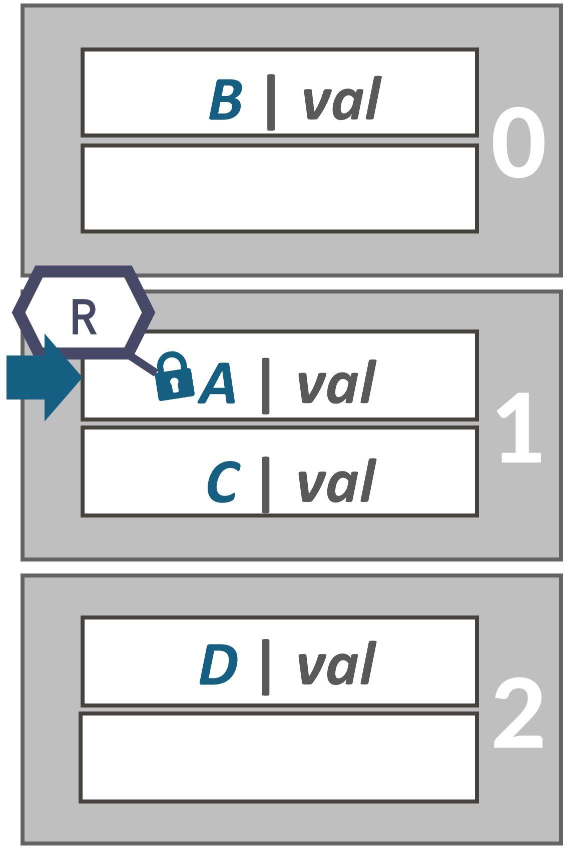


Hash Table – Slot Latches



Hash Table – Slot Latches

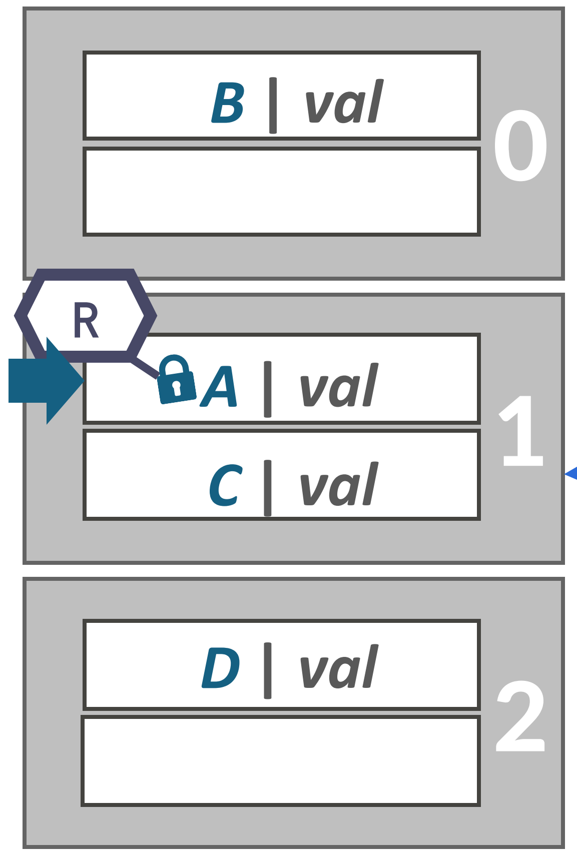
T$_1$: Find D
hash(D)



T$_2$: Insert E
hash(E)

Hash Table – Slot Latches

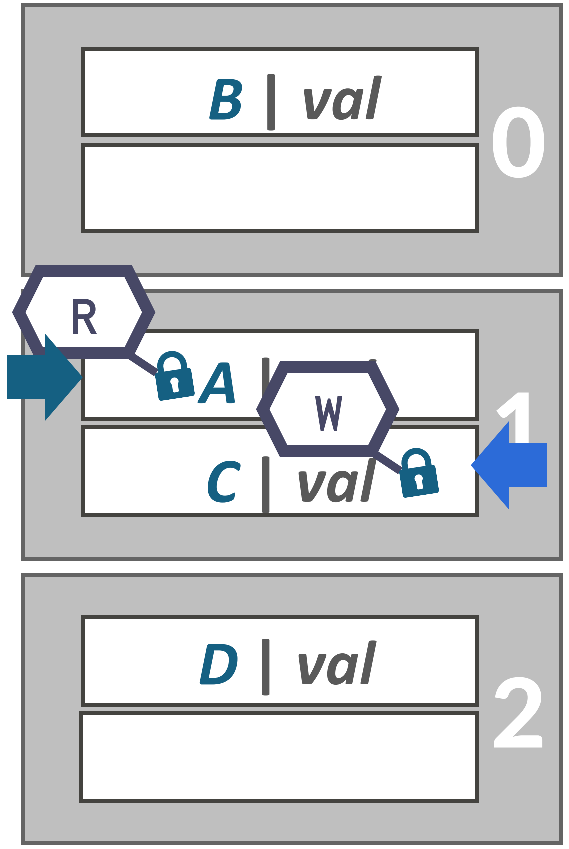
T_1: Find D
hash(D)



T_2: Insert E
hash(E)

Hash Table – Slot Latches

T$_1$: Find D
hash(D)

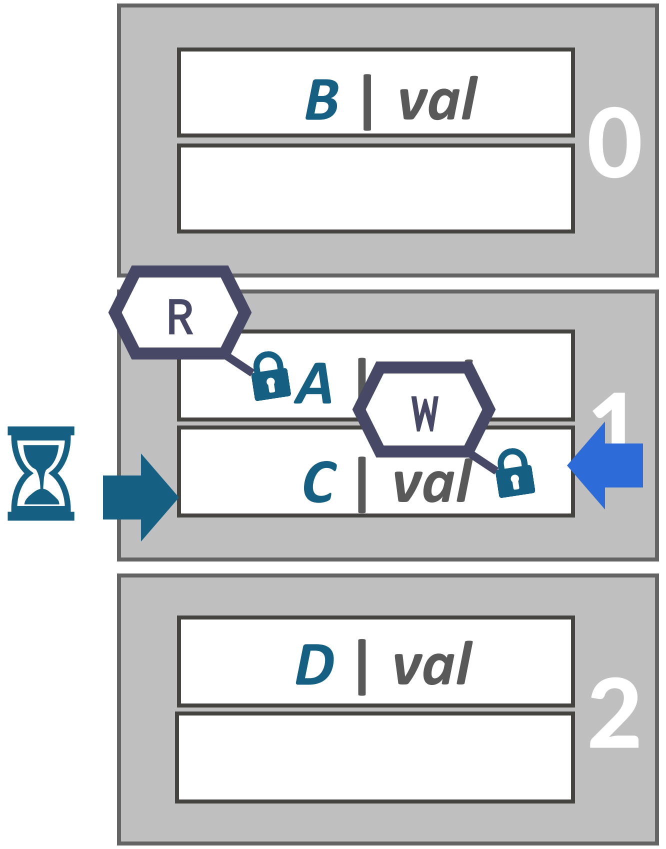


T$_2$: Insert E
hash(E)

Hash Table – Slot Latches

T$_1$: Find D
hash(D)

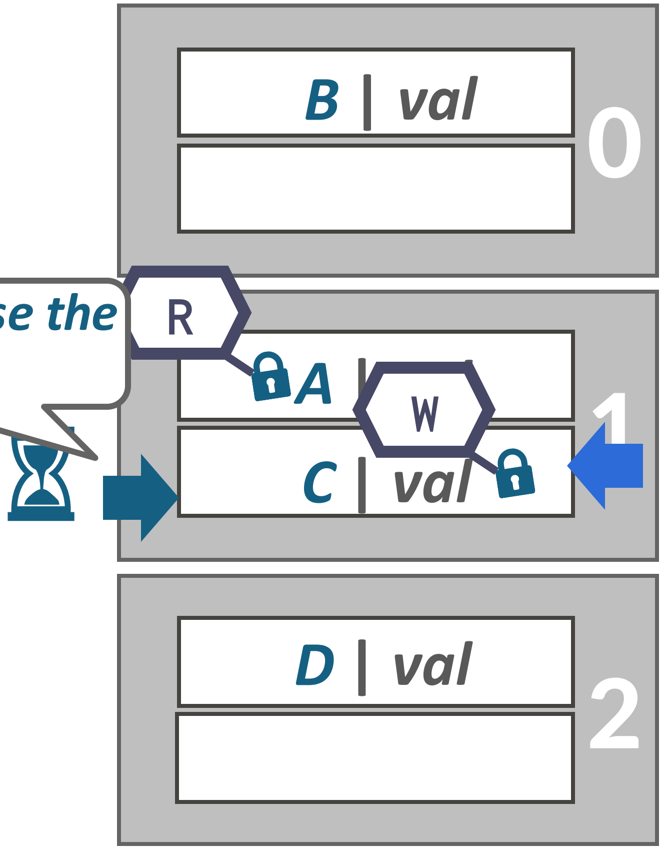
T$_2$: Insert E
hash(E)



Hash Table – Slot Latches

T$_1$: Find D

ha

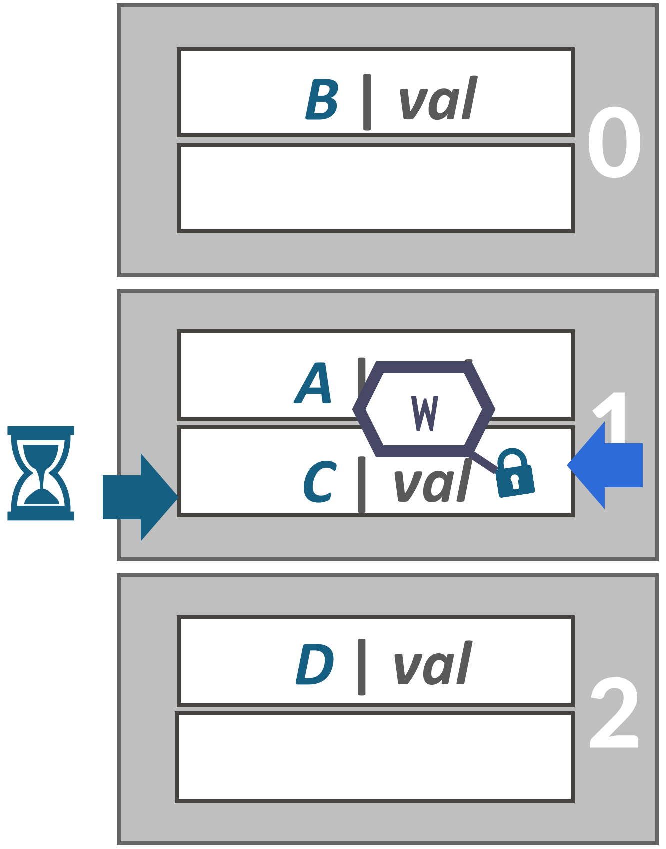


T$_2$: Insert E

hash(E)

Hash Table – Slot Latches

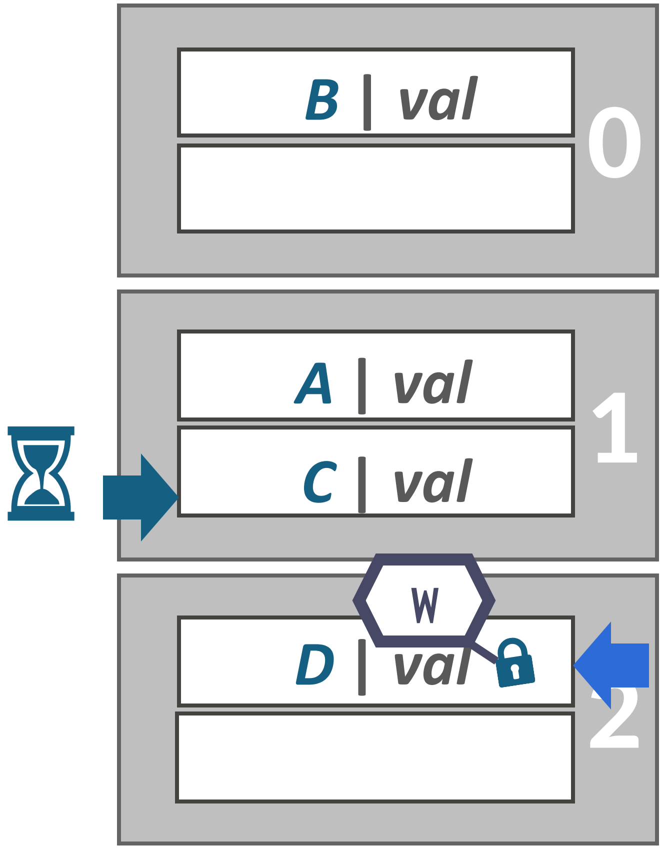
T$_1$: Find D
hash(D)



T$_2$: Insert E
hash(E)

Hash Table – Slot Latches

T_1: Find D
hash(D)

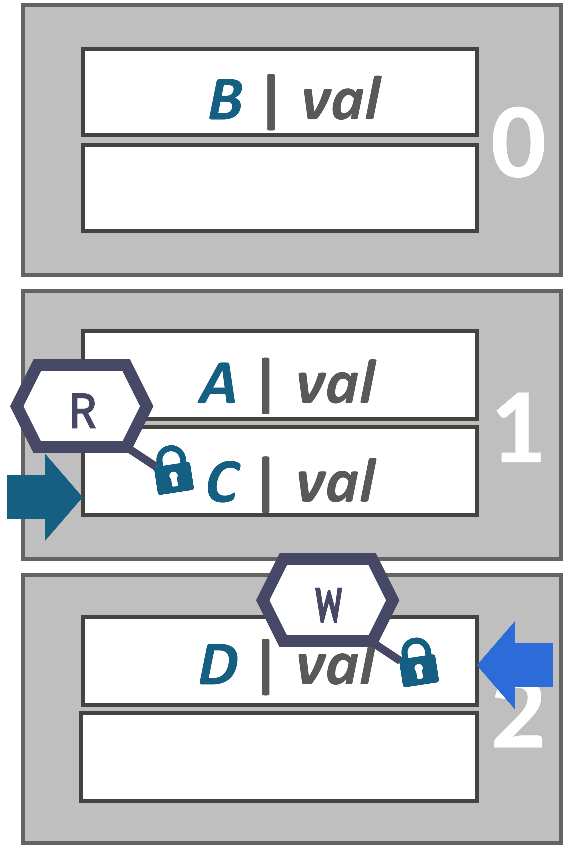


T_2: Insert E
hash(E)

SCHOOL OF DATA SCIENCE

Hash Table – Slot Latches

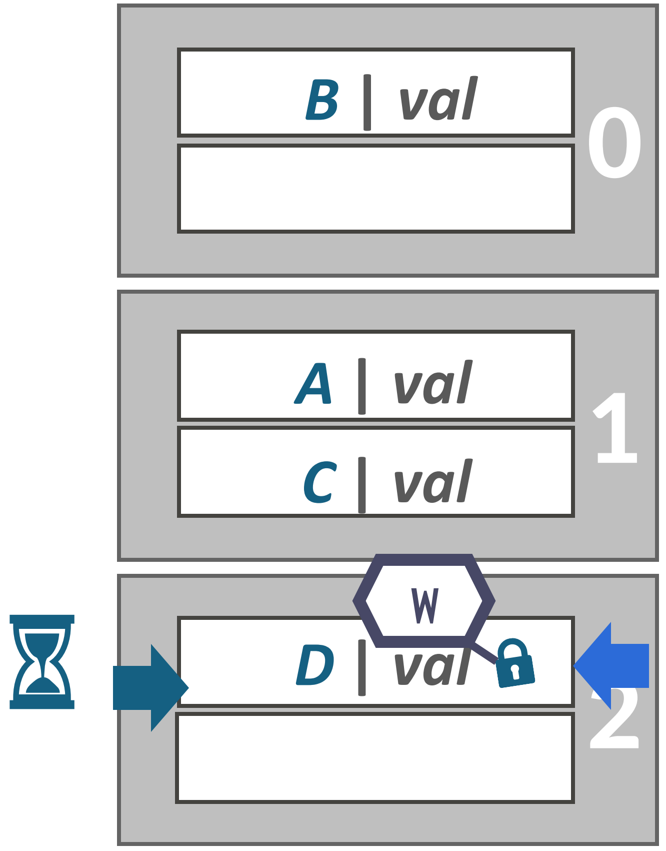
T_1: Find D
hash(D)



T_2: Insert E
hash(E)

Hash Table – Slot Latches

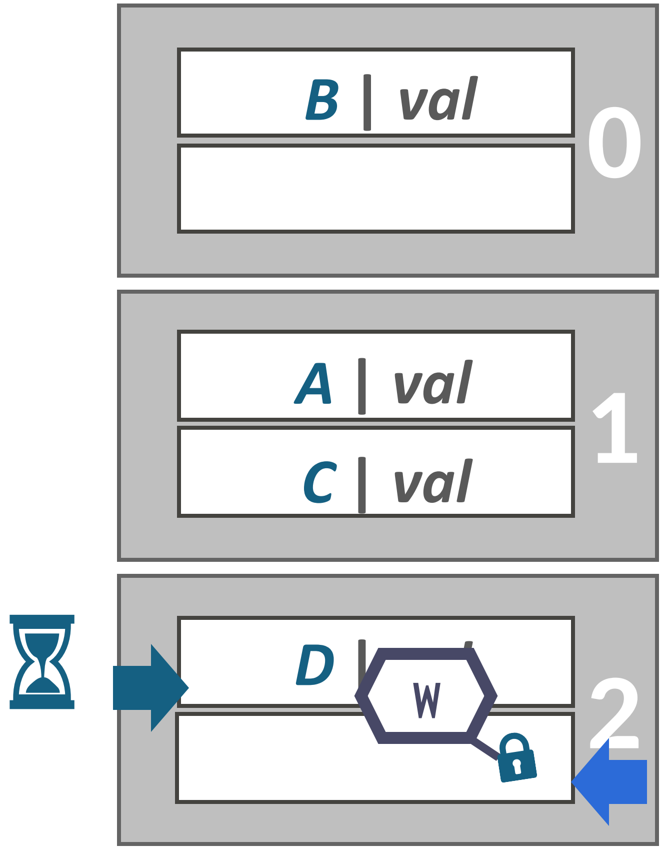
T_1: Find D
hash(D)



T_2: Insert E
hash(E)

Hash Table – Slot Latches

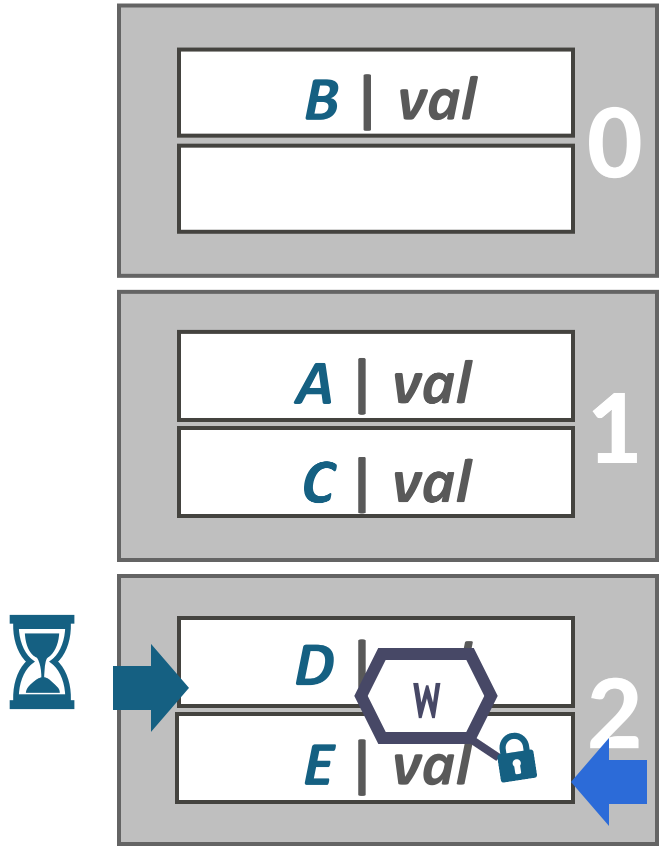
T_1: Find D
hash(D)



T_2: Insert E
hash(E)

Hash Table – Slot Latches

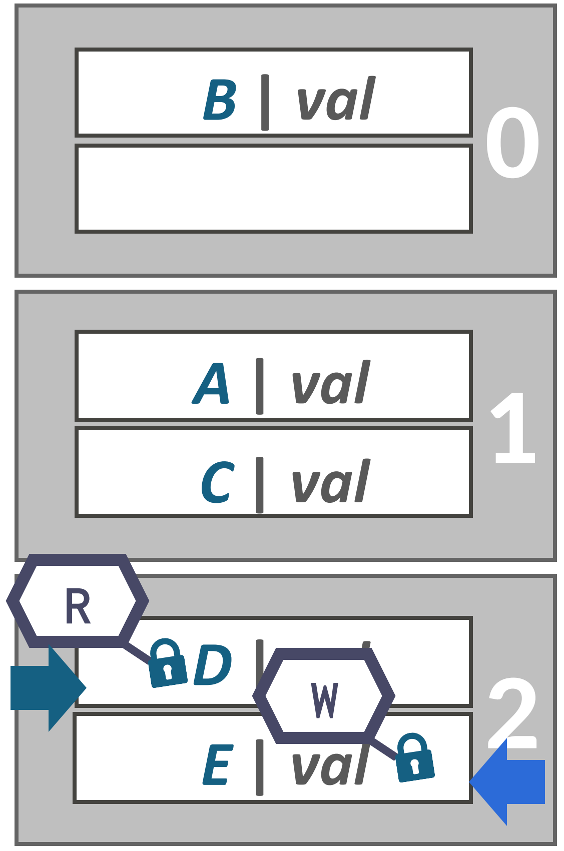
T$_1$: Find D
hash(D)



T$_2$: Insert E
hash(E)

Hash Table – Slot Latches

T_1: Find D
hash(D)



T_2: Insert E
hash(E)

SCHOOL OF
DATA SCIENCE

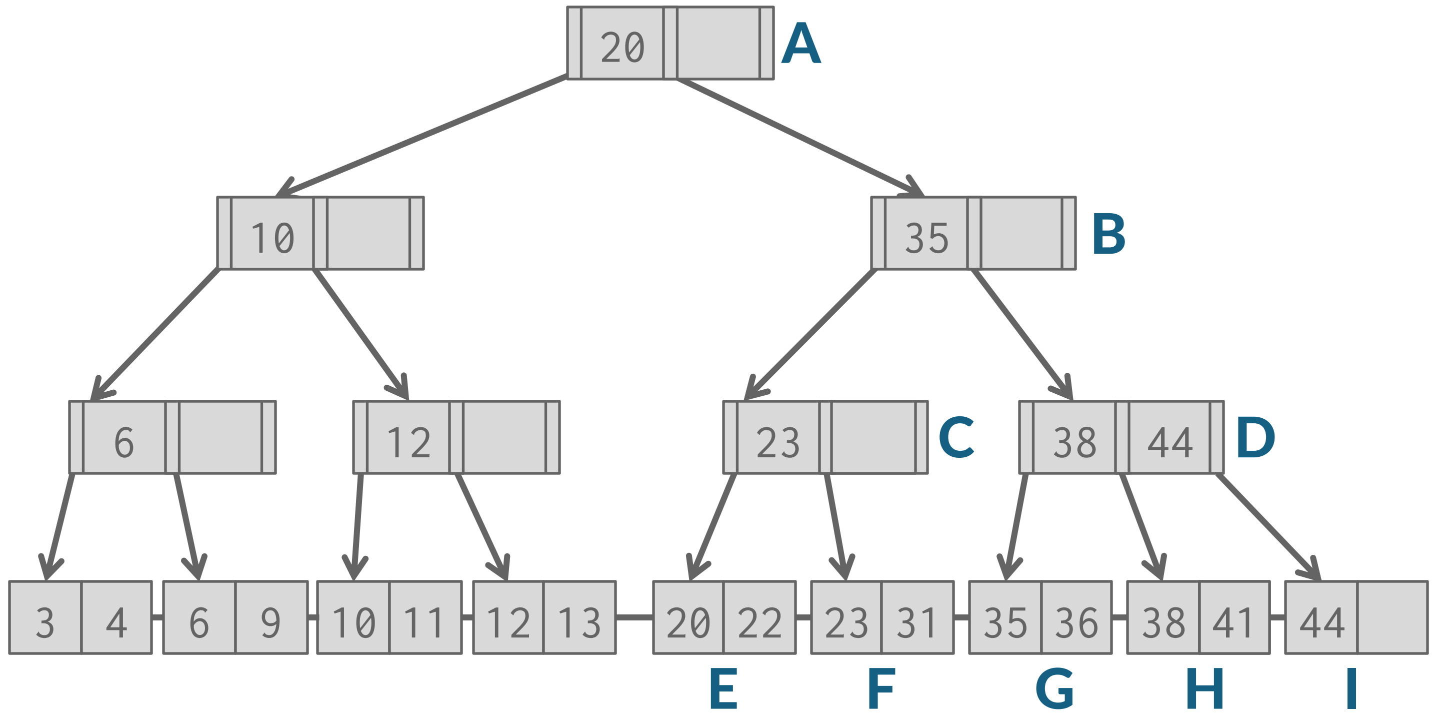
B+Tree Concurrency Control

B+Tree Concurrency Control

- 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.

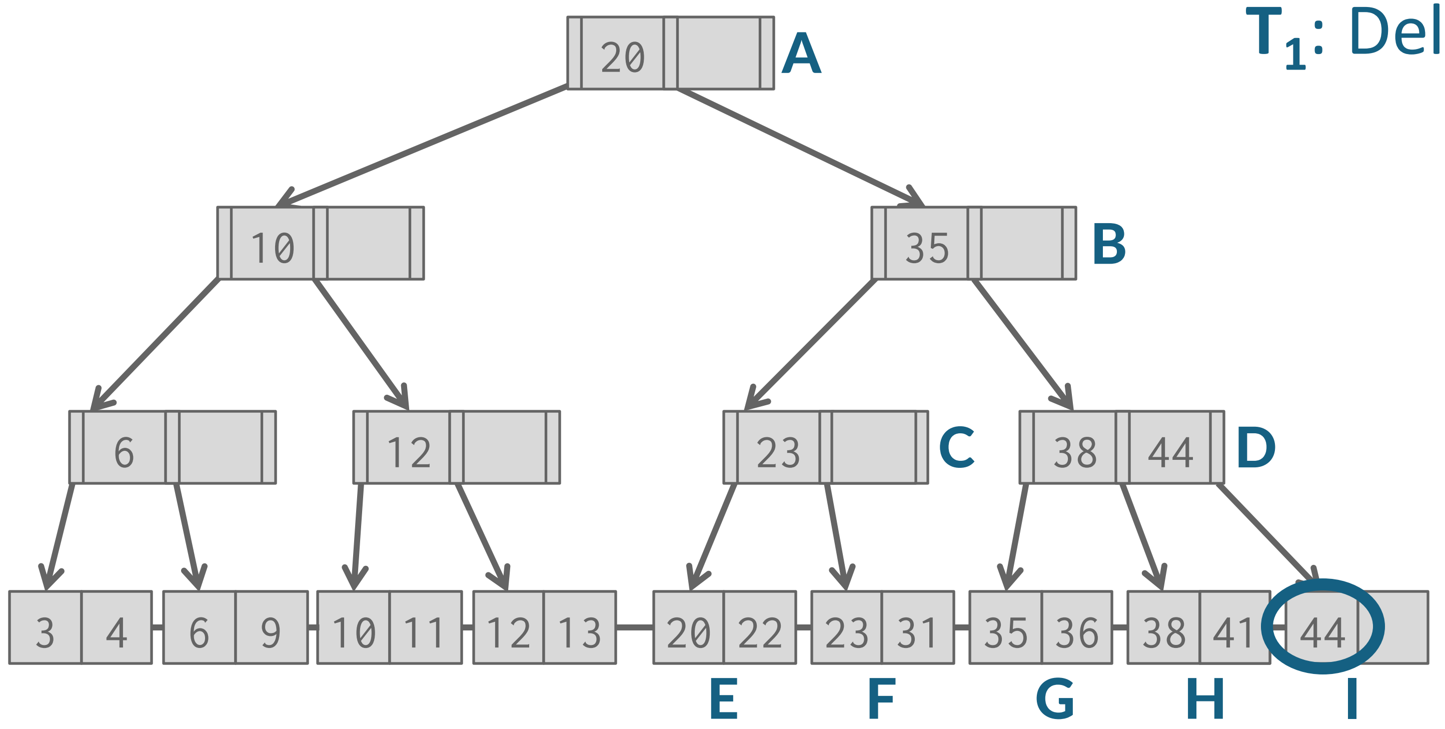
SCHOOL OF DATA SCIENCE

B+Tree Multi-Threaded Example



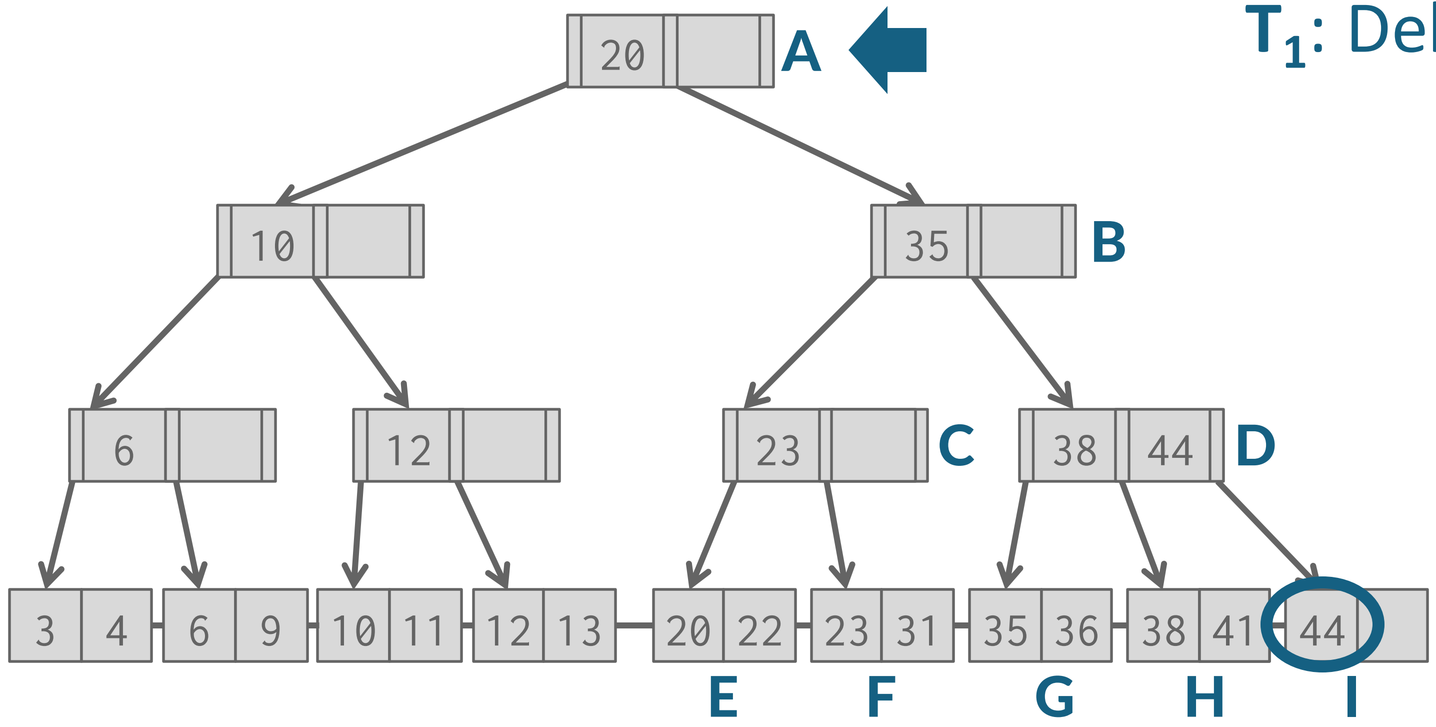
SCHOOL OF DATA SCIENCE

B+Tree Multi-Threaded Example



T_1: Delete 44

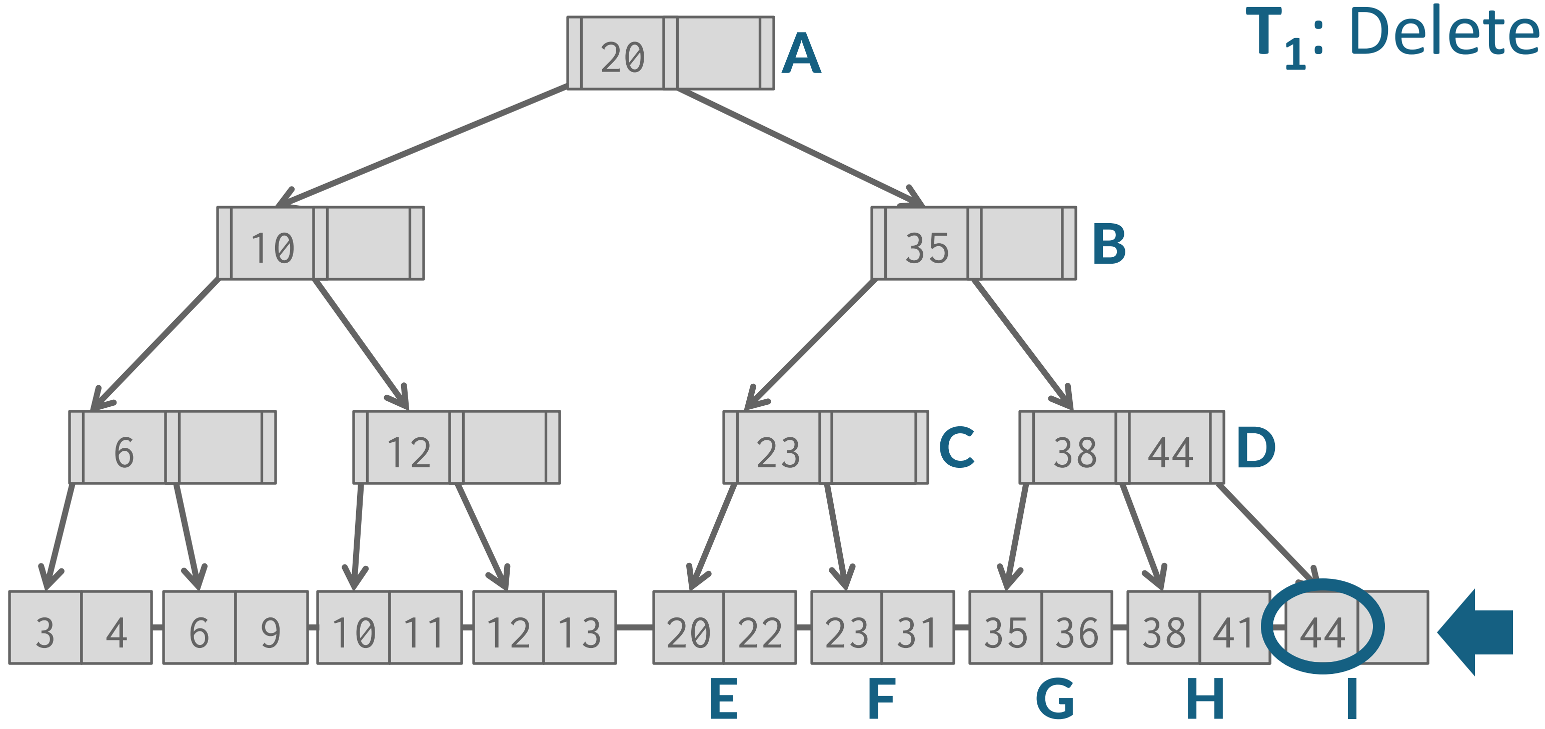
B+Tree Multi-Threaded Example



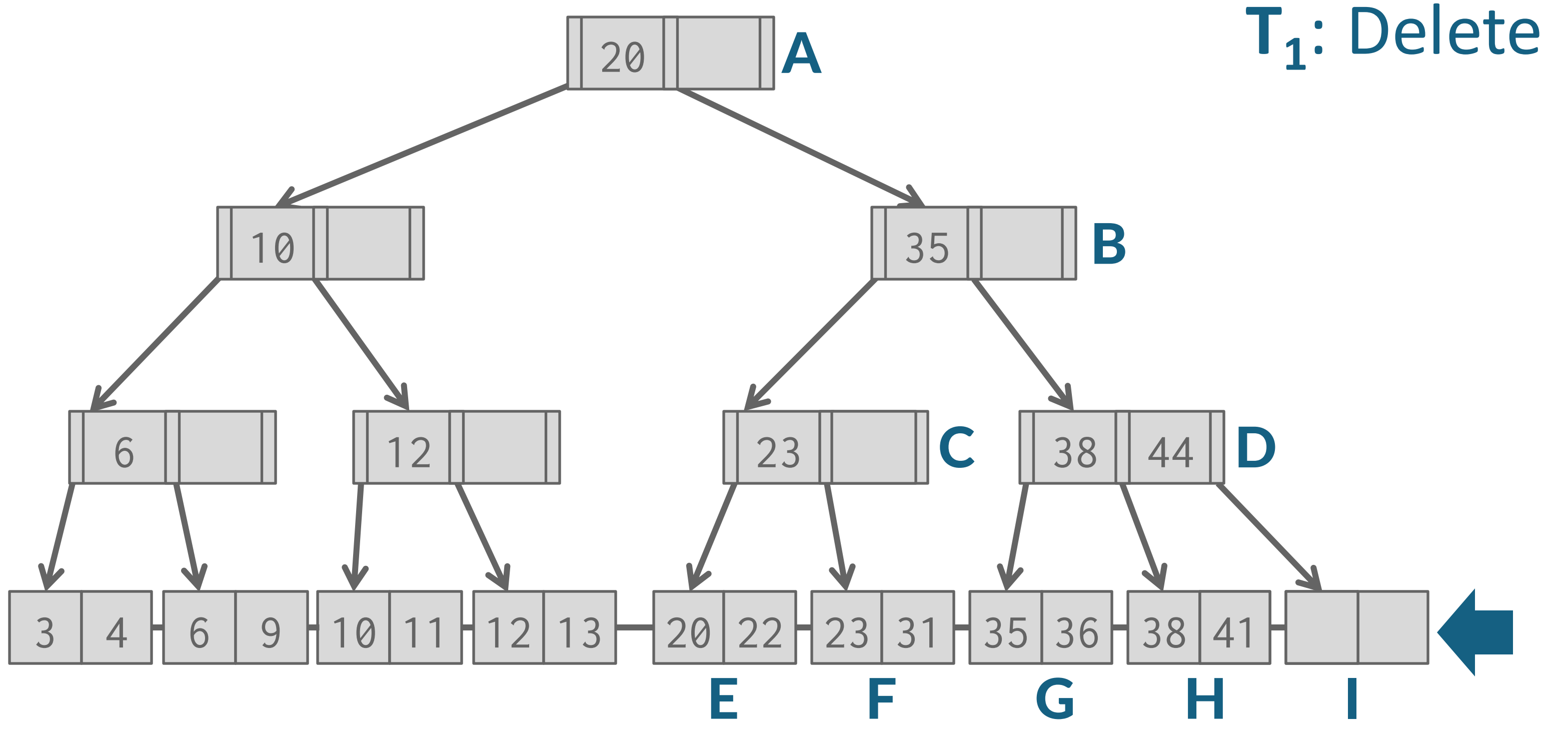
T$_1$: Delete 44

B+Tree Multi-Threaded Example

T$_1$: Delete 44



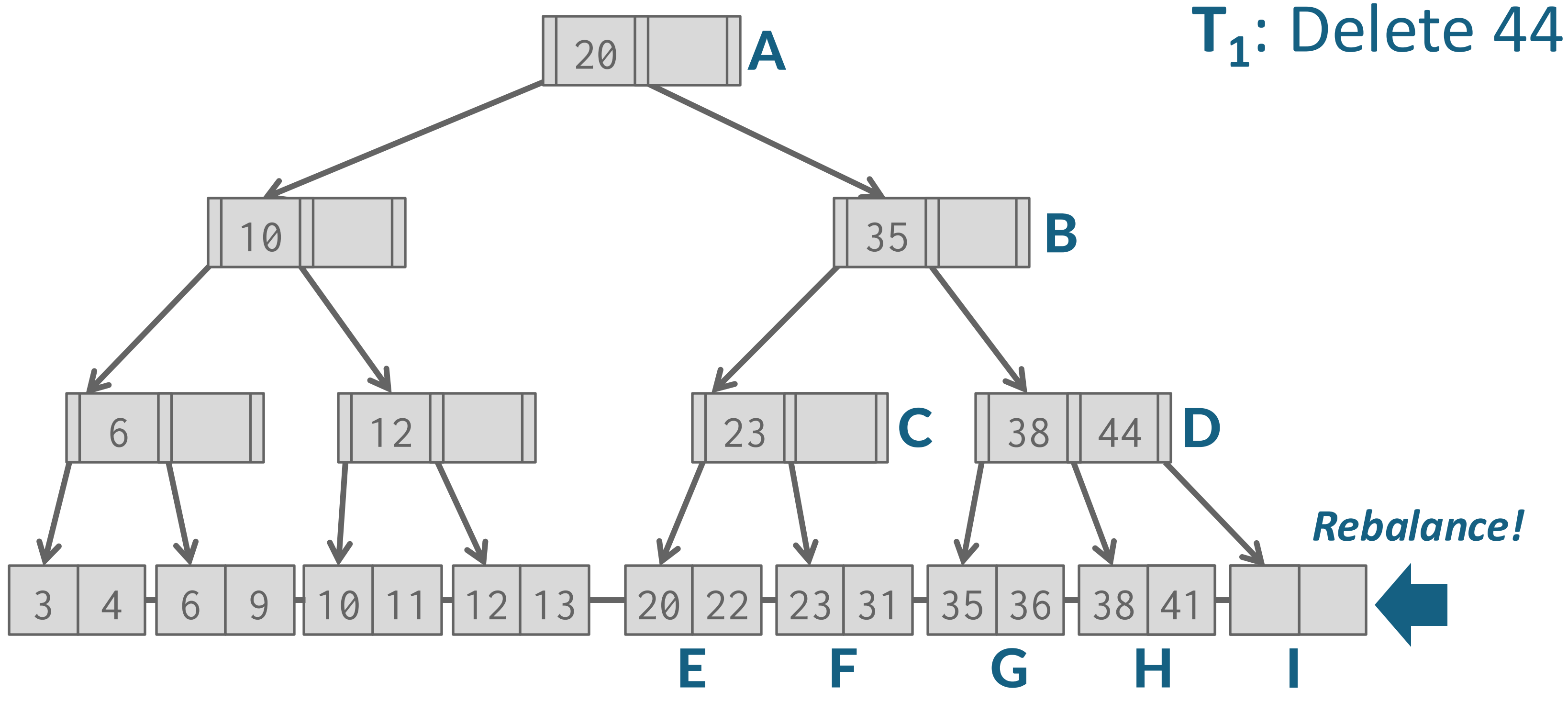
B+Tree Multi-Threaded Example



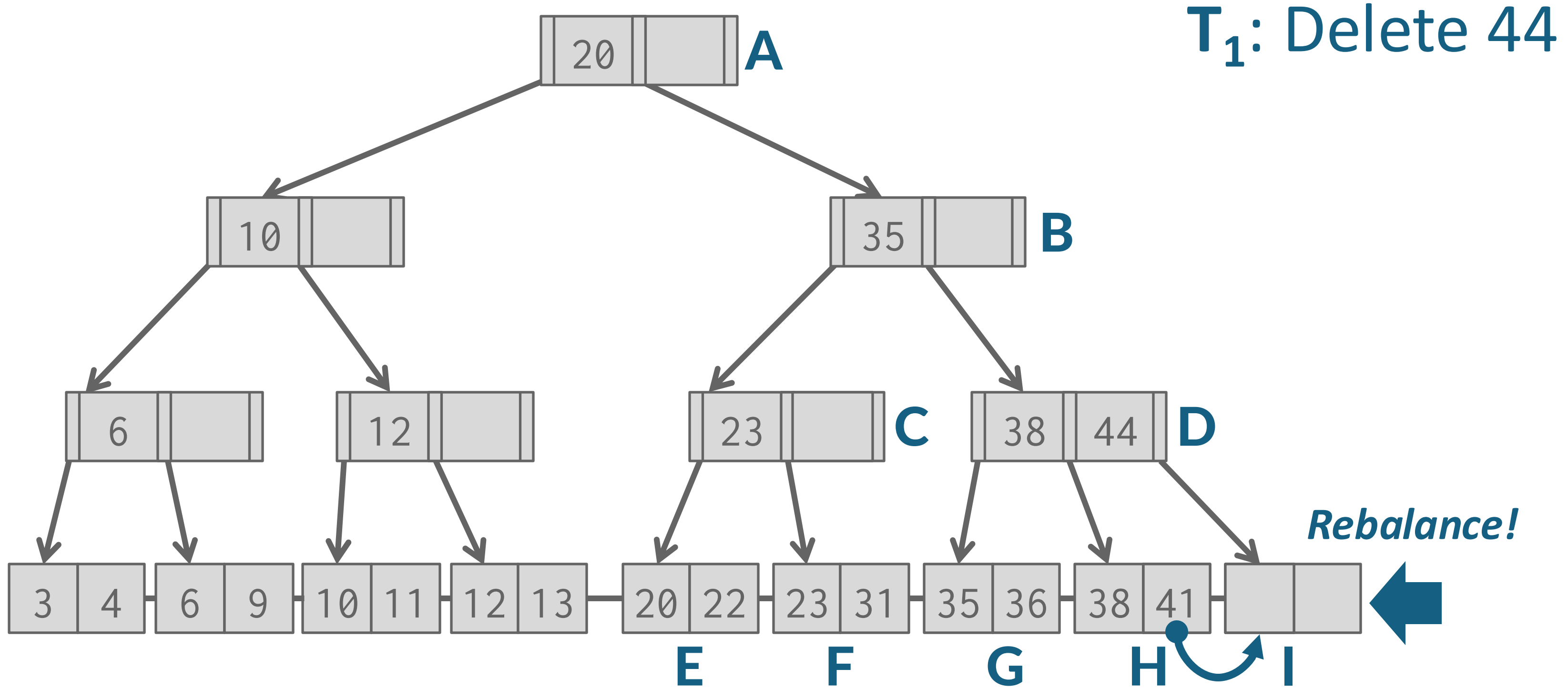
T_1: Delete 44

B+Tree Multi-Threaded Example

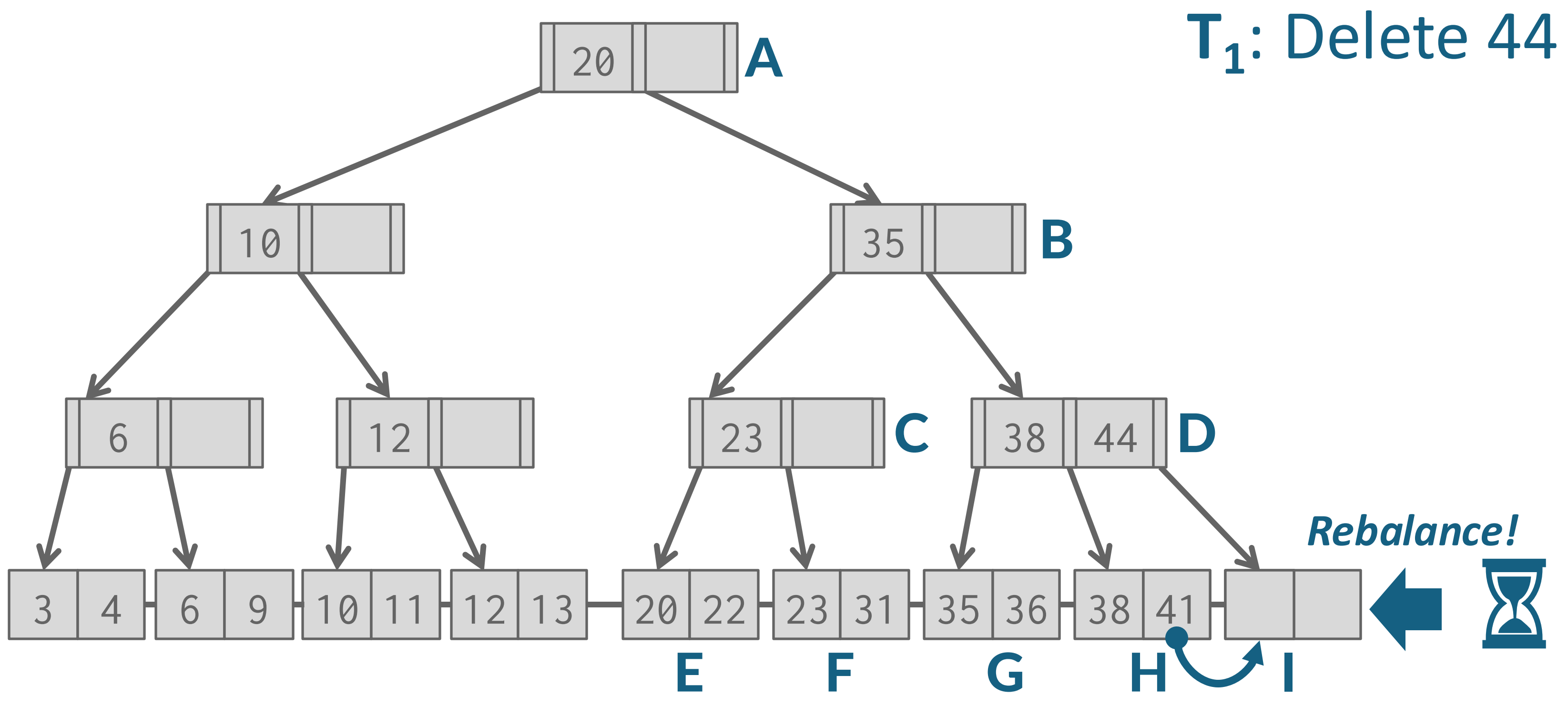
T$_1$: Delete 44



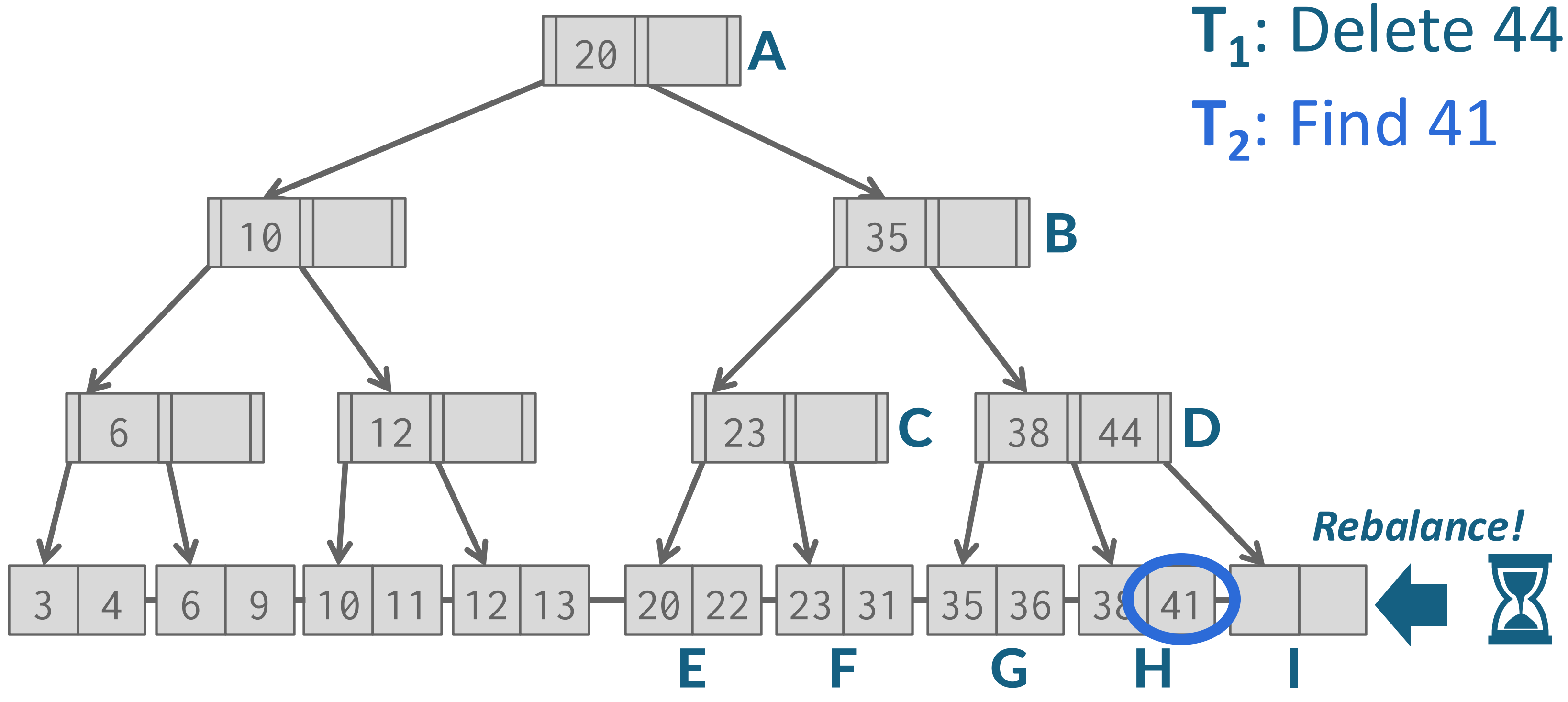
B+Tree Multi-Threaded Example



B+Tree Multi-Threaded Example



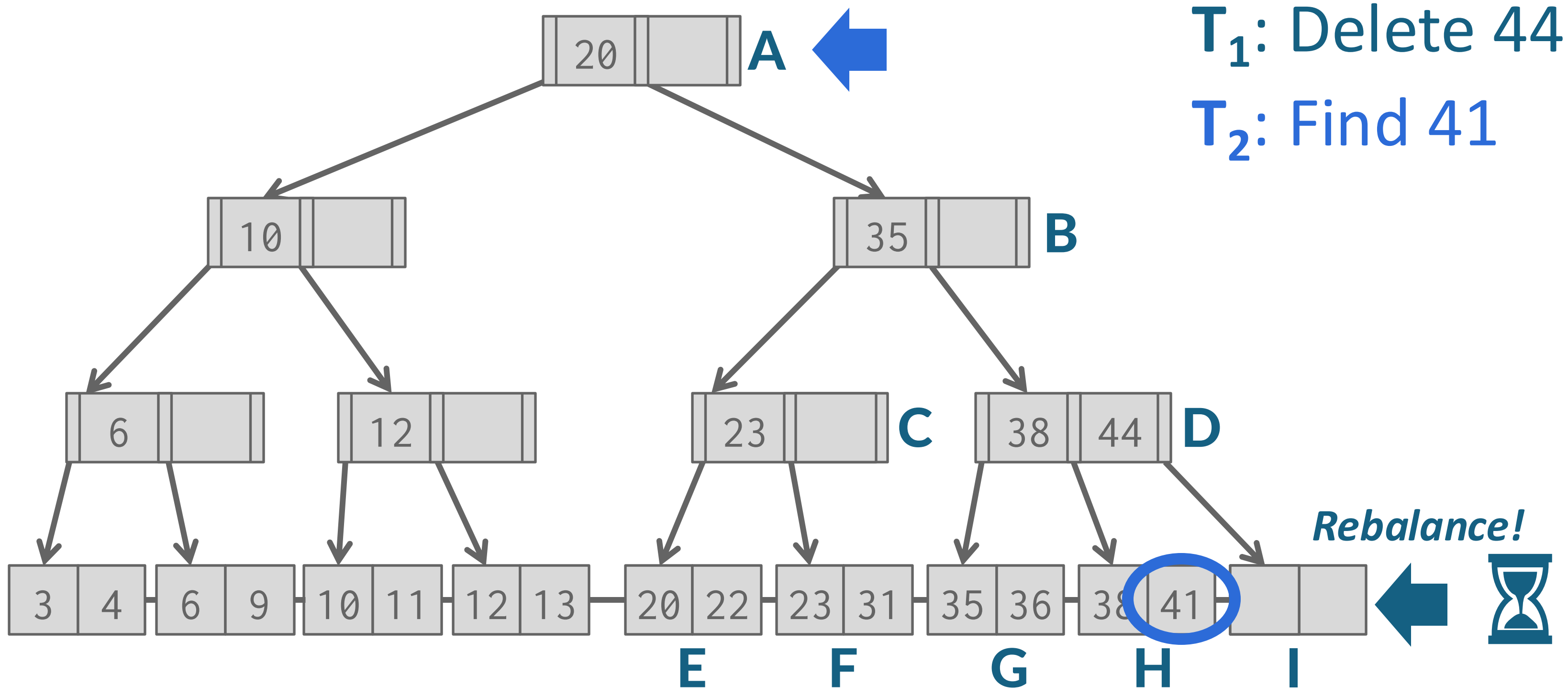
B+Tree Multi-Threaded Example



T_1: Delete 44

T_2: Find 41

B+Tree Multi-Threaded Example

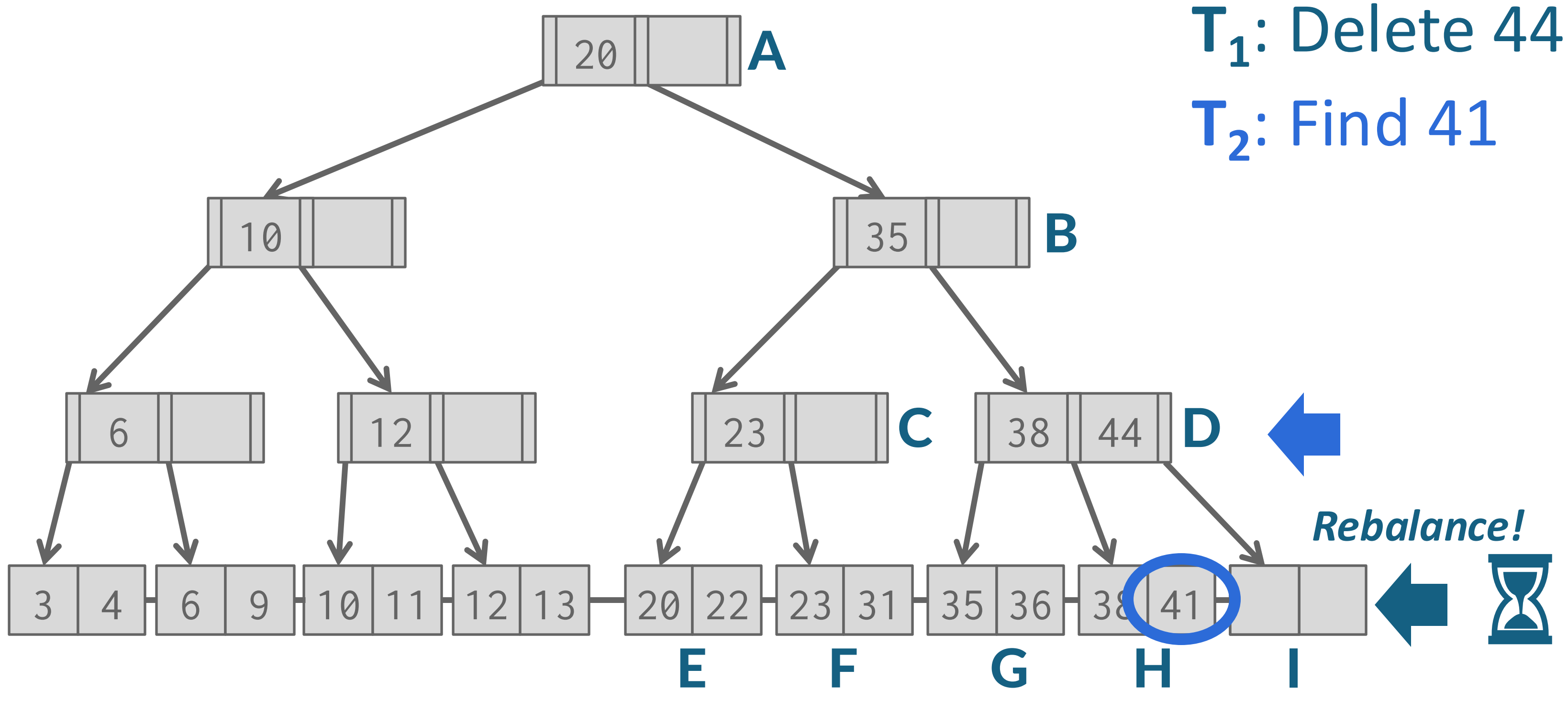


T_1: Delete 44

T_2: Find 41

SCHOOL OF
DATA SCIENCE

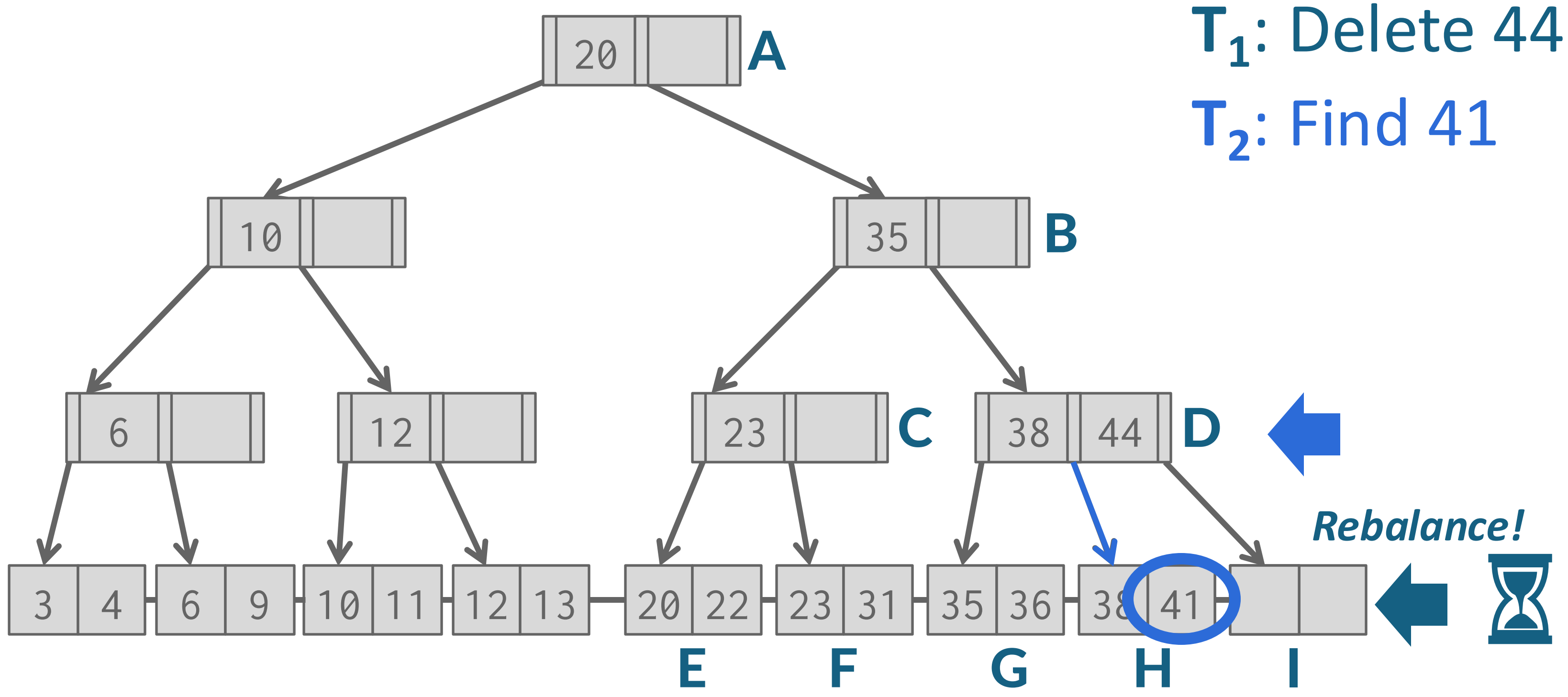
B+Tree Multi-Threaded Example



T_1: Delete 44

T_2: Find 41

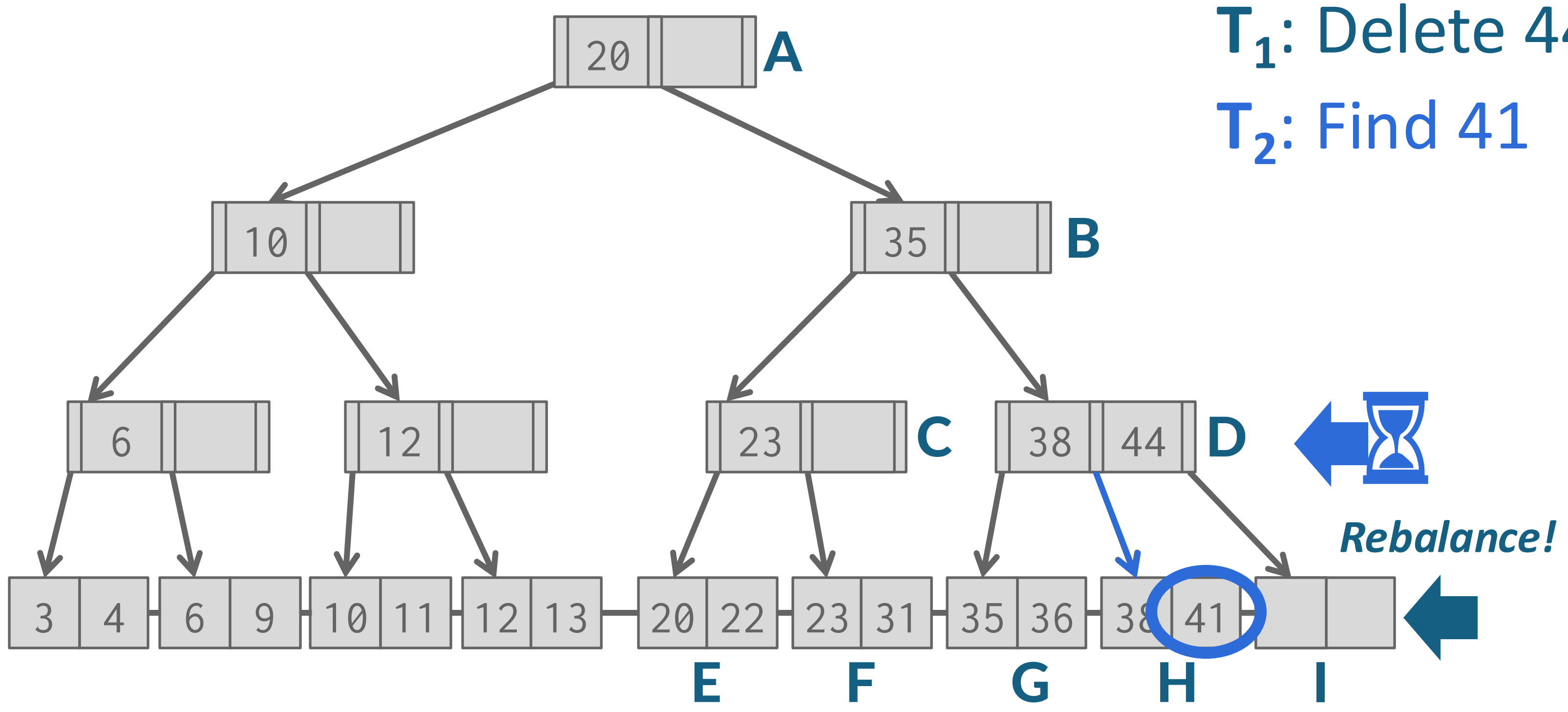
B+Tree Multi-Threaded Example



T$_1$: Delete 44

T$_2$: Find 41

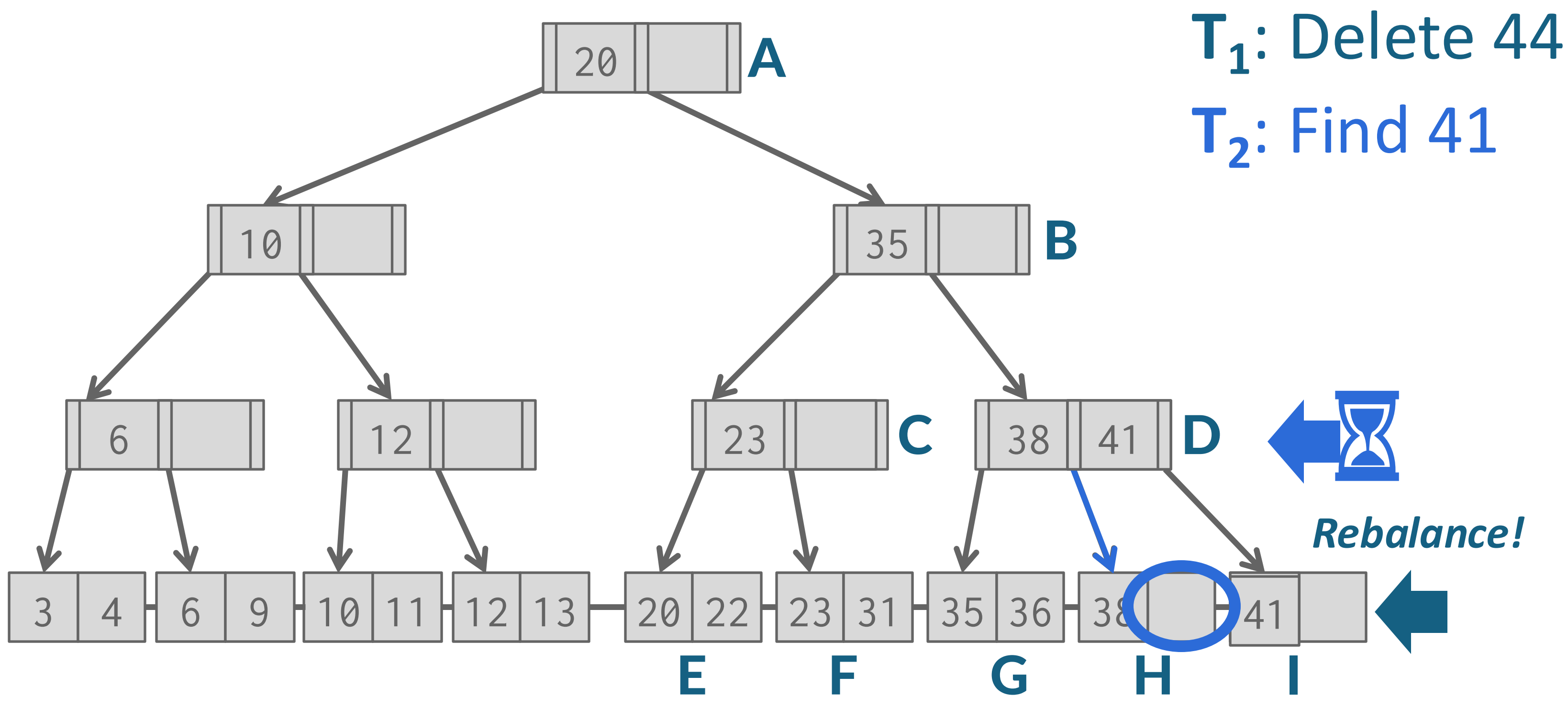
B+Tree Multi-Threaded Example



T_1: Delete 44

T_2: Find 41

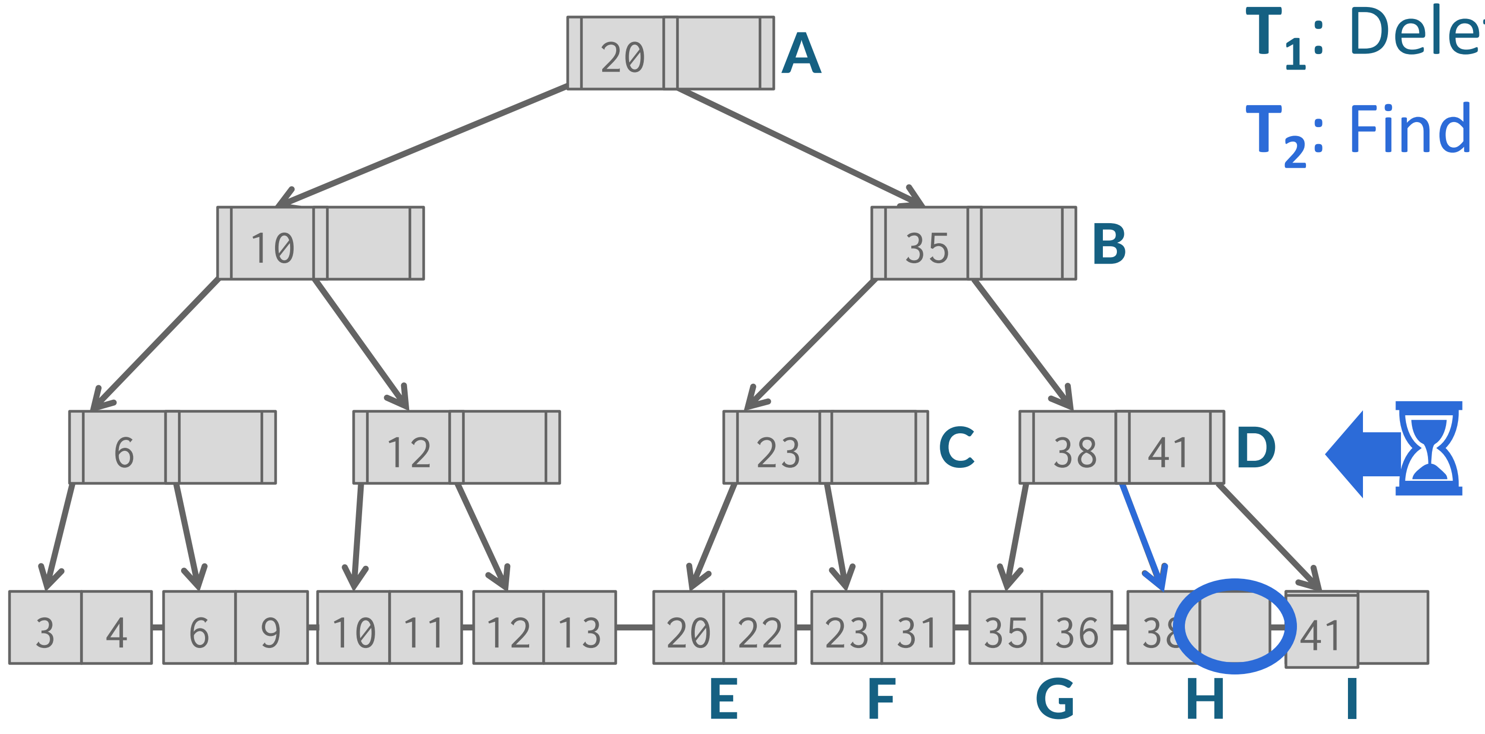
B+Tree Multi-Threaded Example



T_1: Delete 44

T_2: Find 41

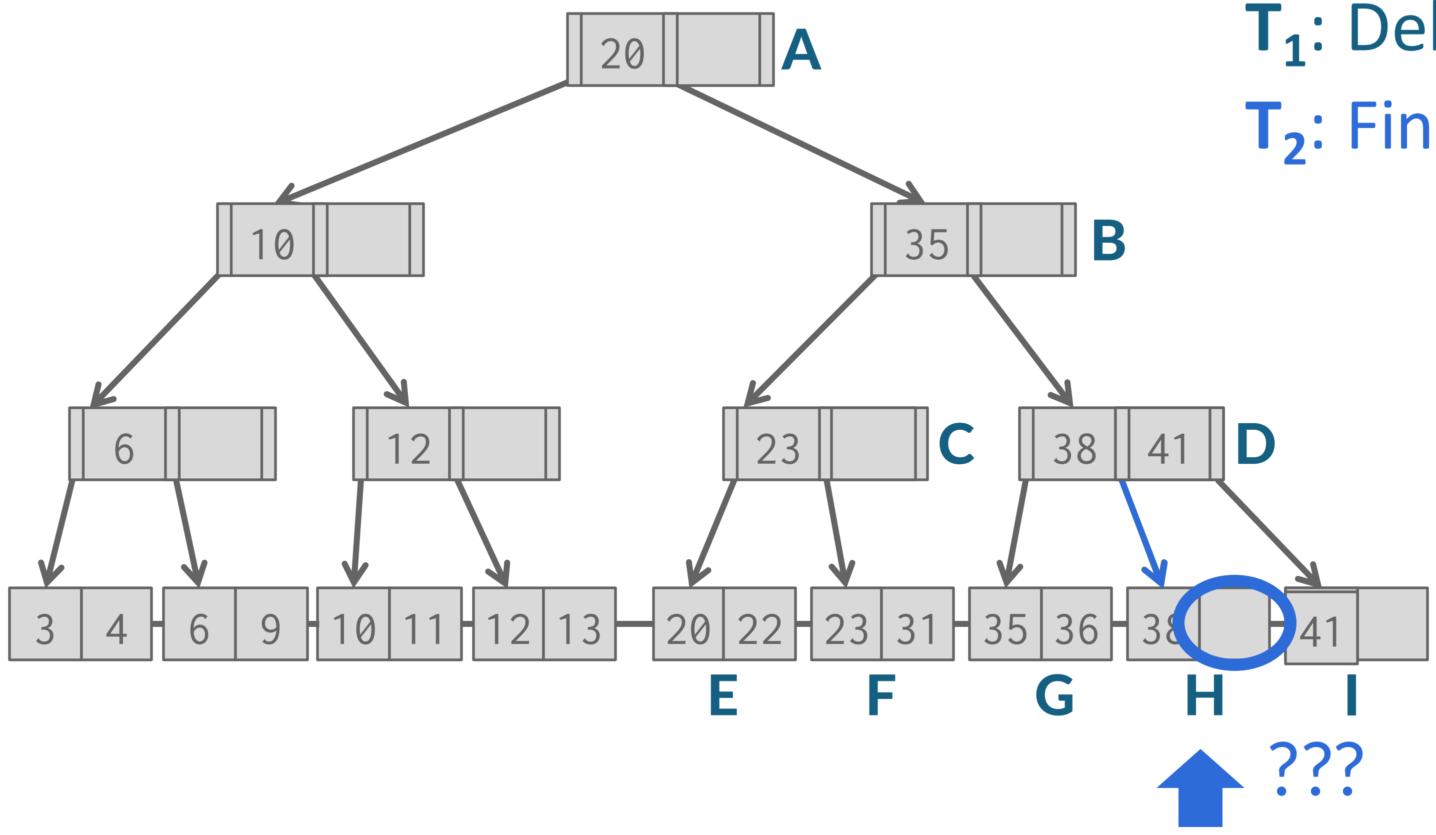
B+Tree Multi-Threaded Example



T$_1$: Delete 44

T$_2$: Find 41

B+Tree Multi-Threaded Example



T_1: Delete 44

T_2: Find 41

Latch Crabbling/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”

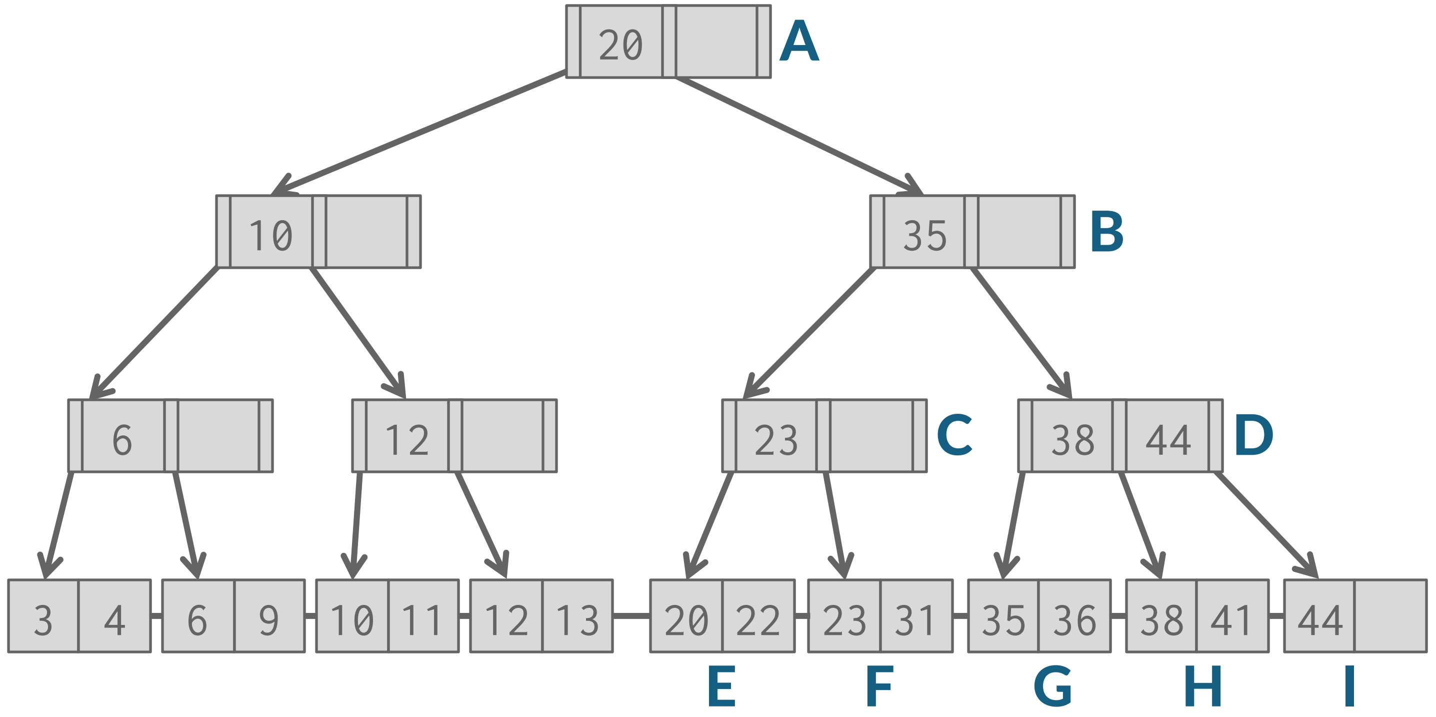
- A **safe node** is one that will not split or merge when updated.
    - Not full (on insertion)
    - More than half-full (on deletion)

Latch Crabbling/Coupling

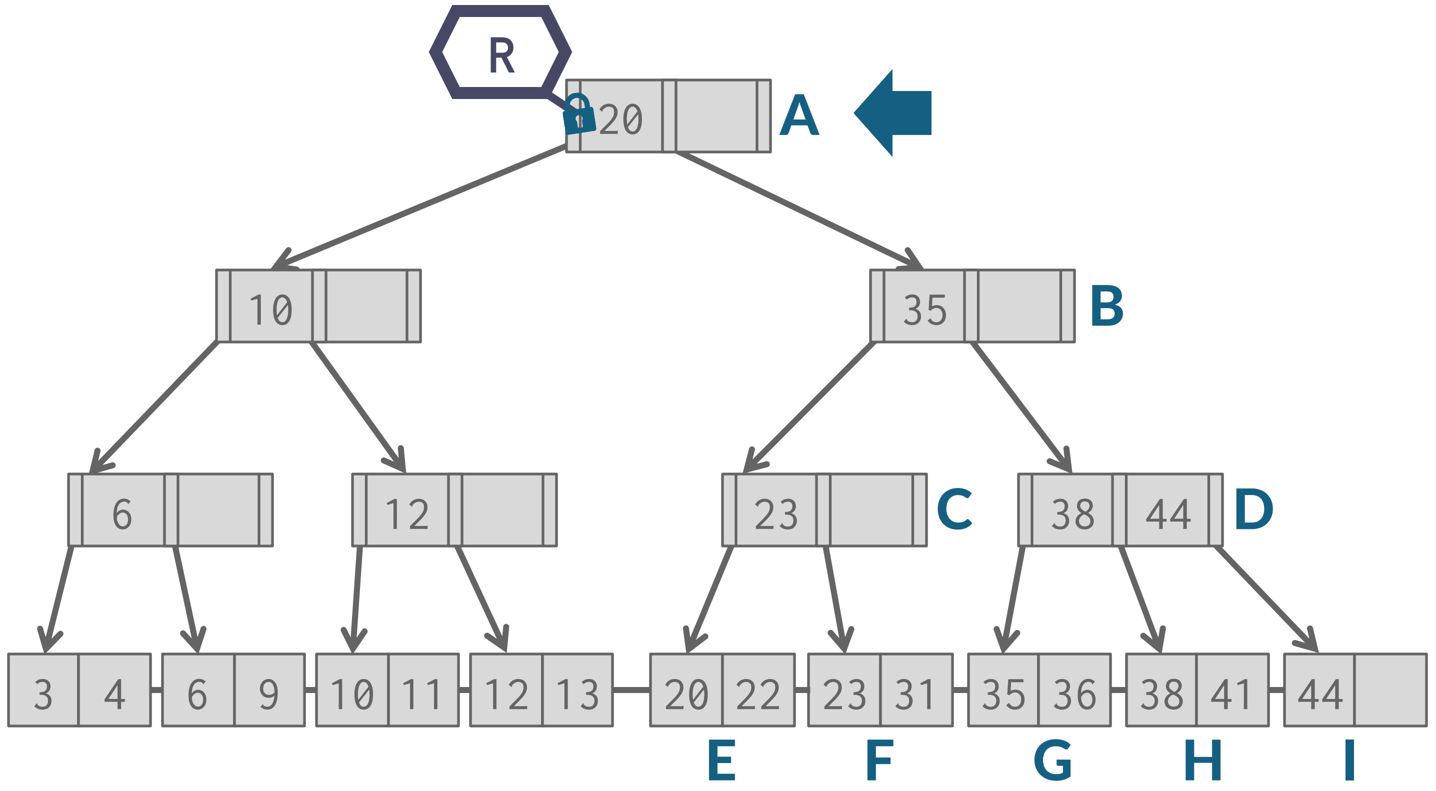
- **Find**: Start at root and traverse down the tree:
    - Acquire `R` latch on child,
    - Then unlatch parent.
    - Repeat until we reach the leaf node.
- **Insert/Delete**: Start at root and go down, obtaining `W` latches as needed. Once child is latched, check if it is safe:
    - If child is safe, release all latches on ancestors

SCHOOL OF DATA SCIENCE

Example #1 – Find 38

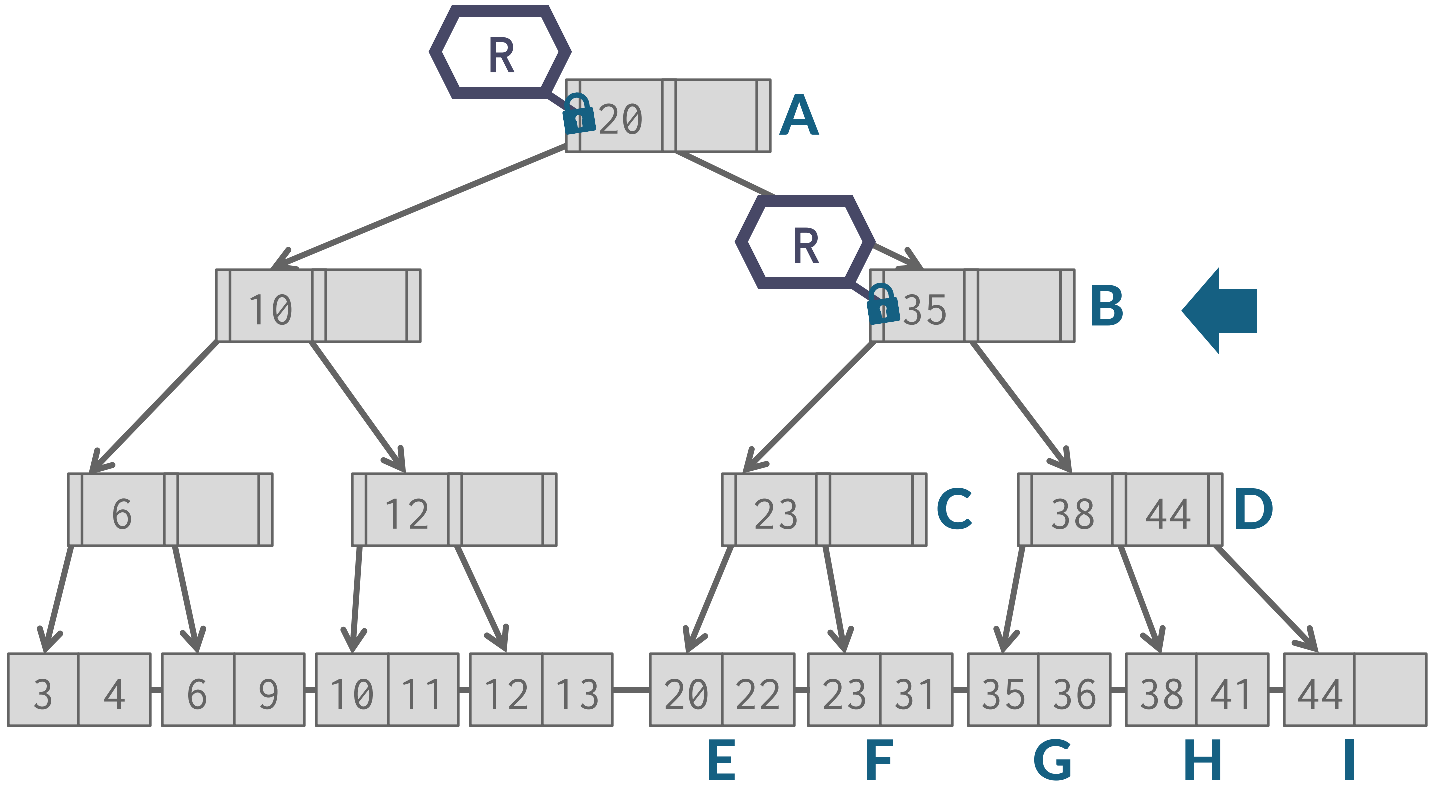


Example #1 – Find 38

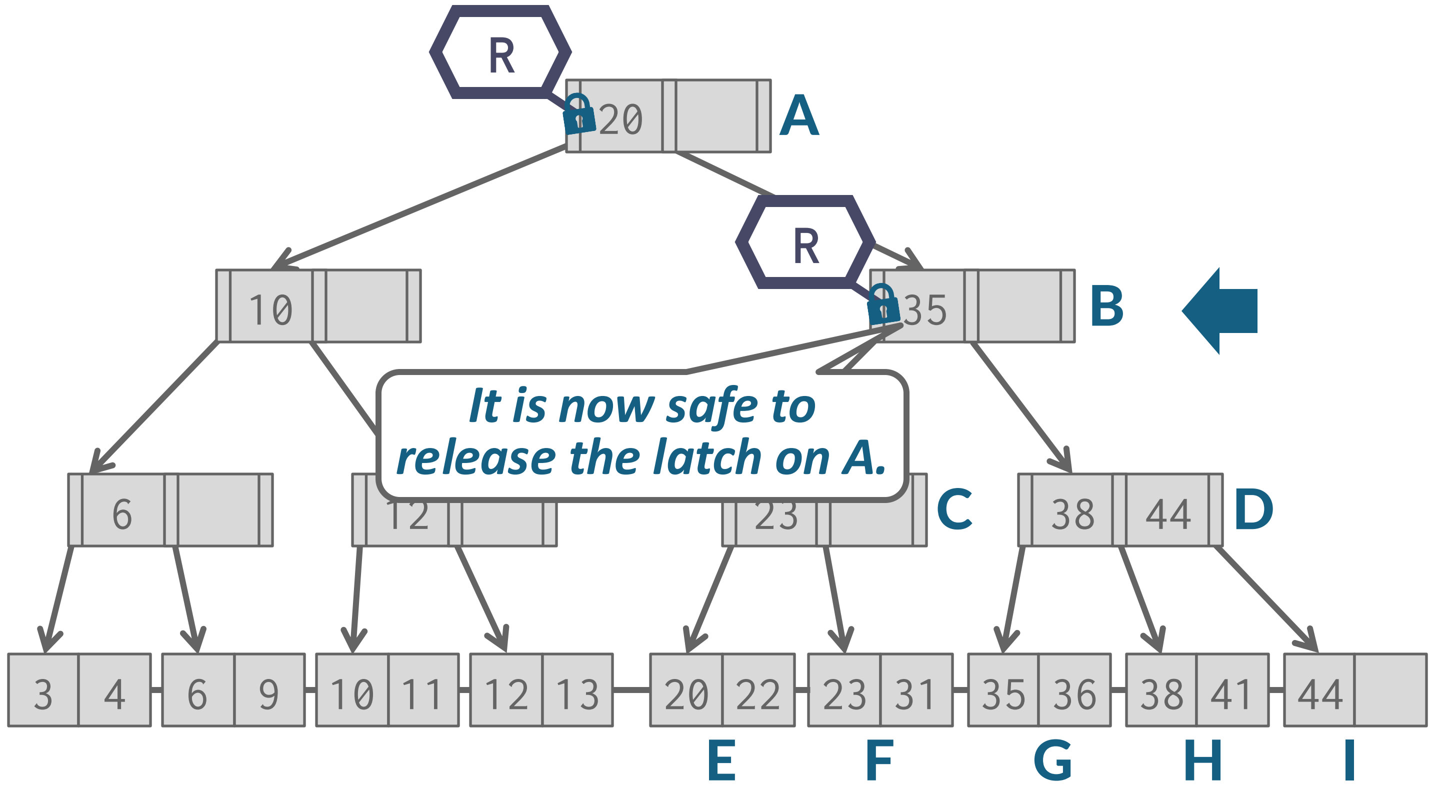


SCHOOL OF DATA SCIENCE

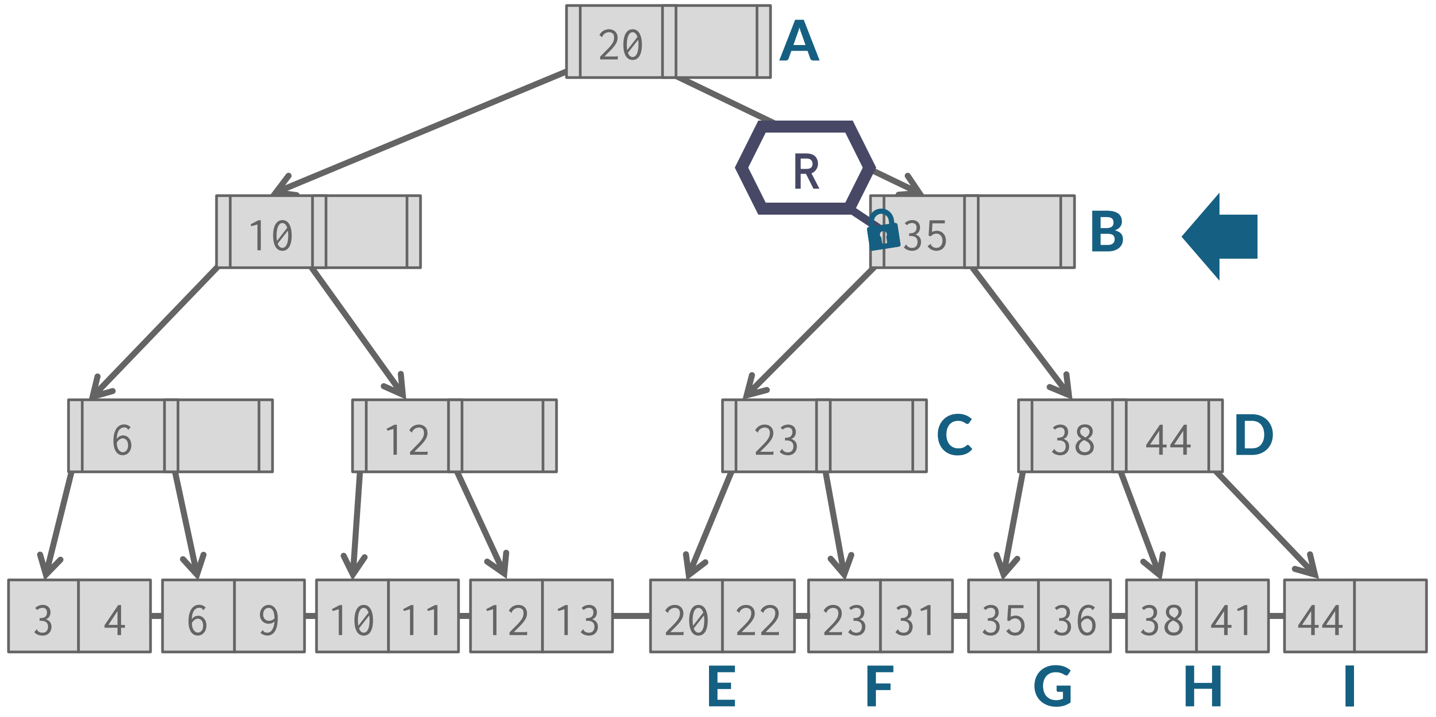
Example #1 – Find 38



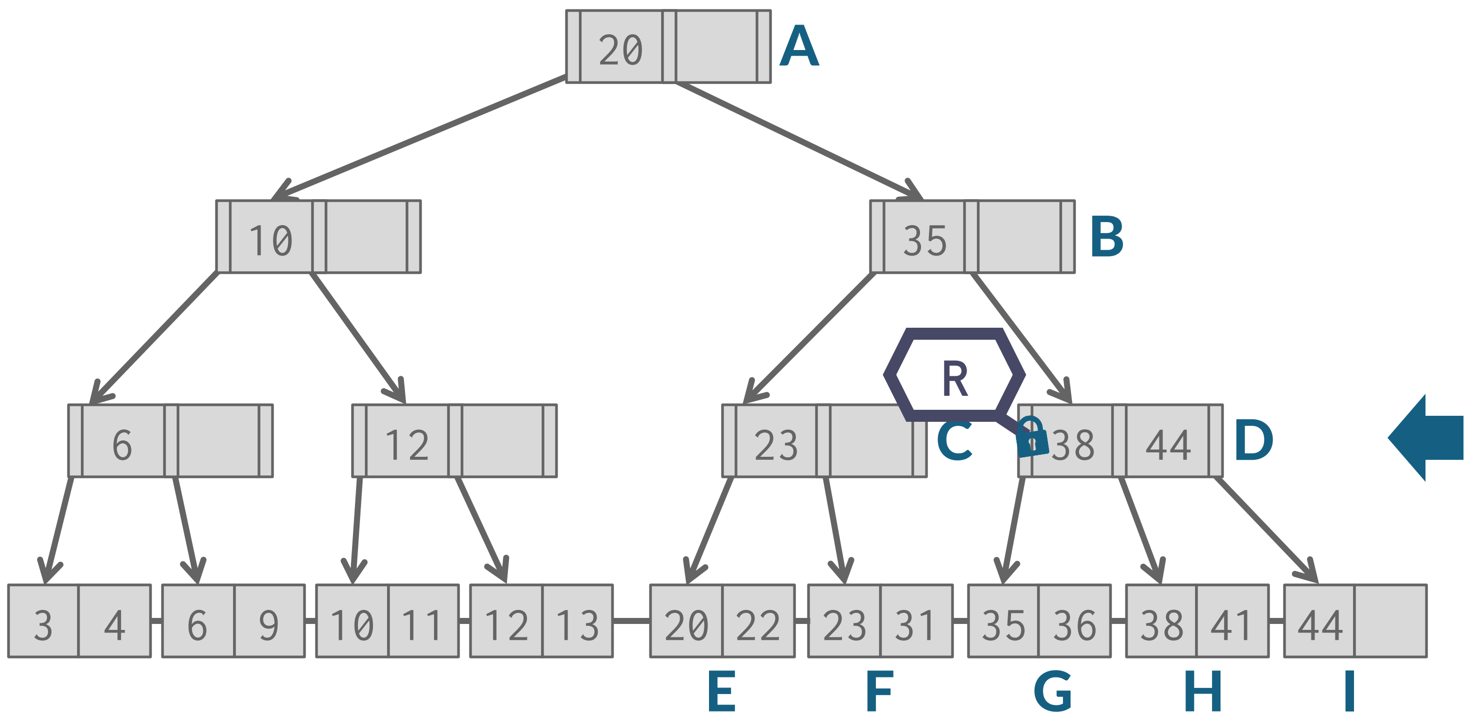
Example #1 – Find 38



Example #1 – Find 38

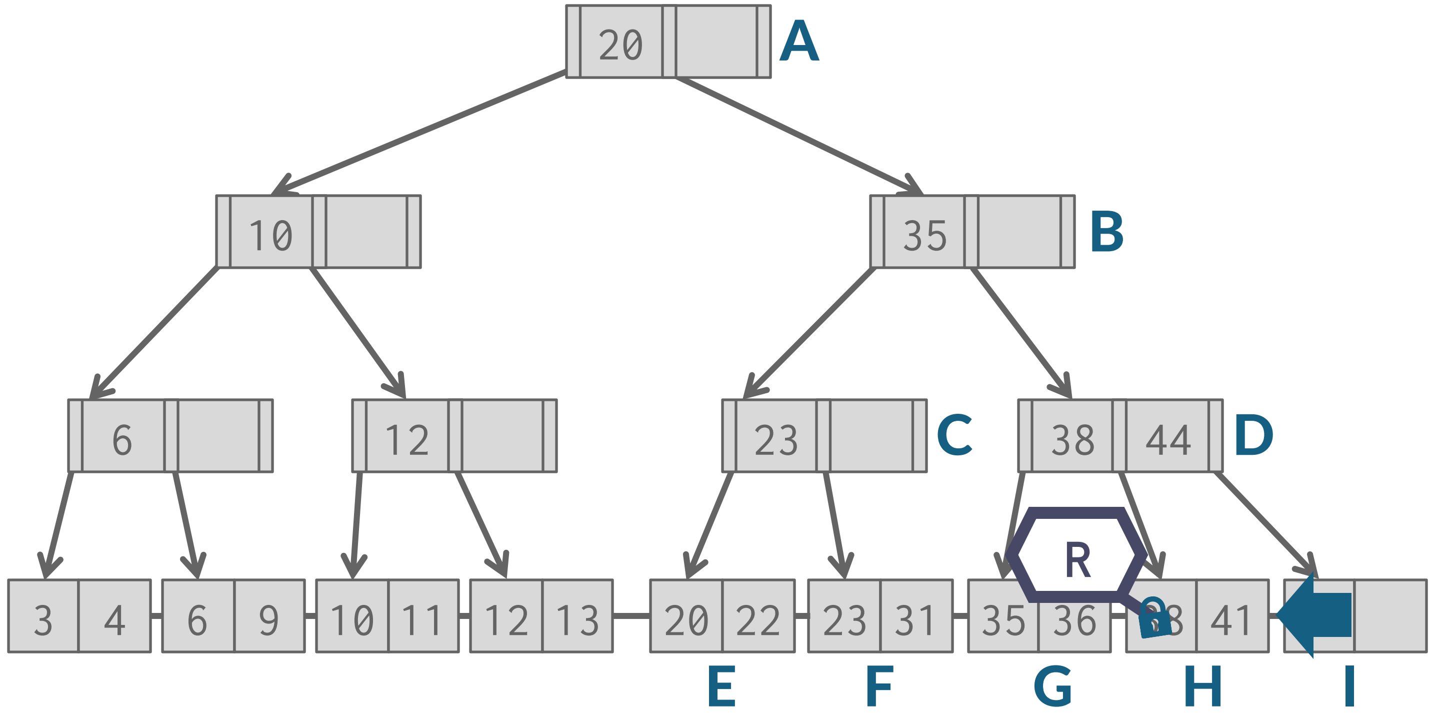


Example #1 – Find 38

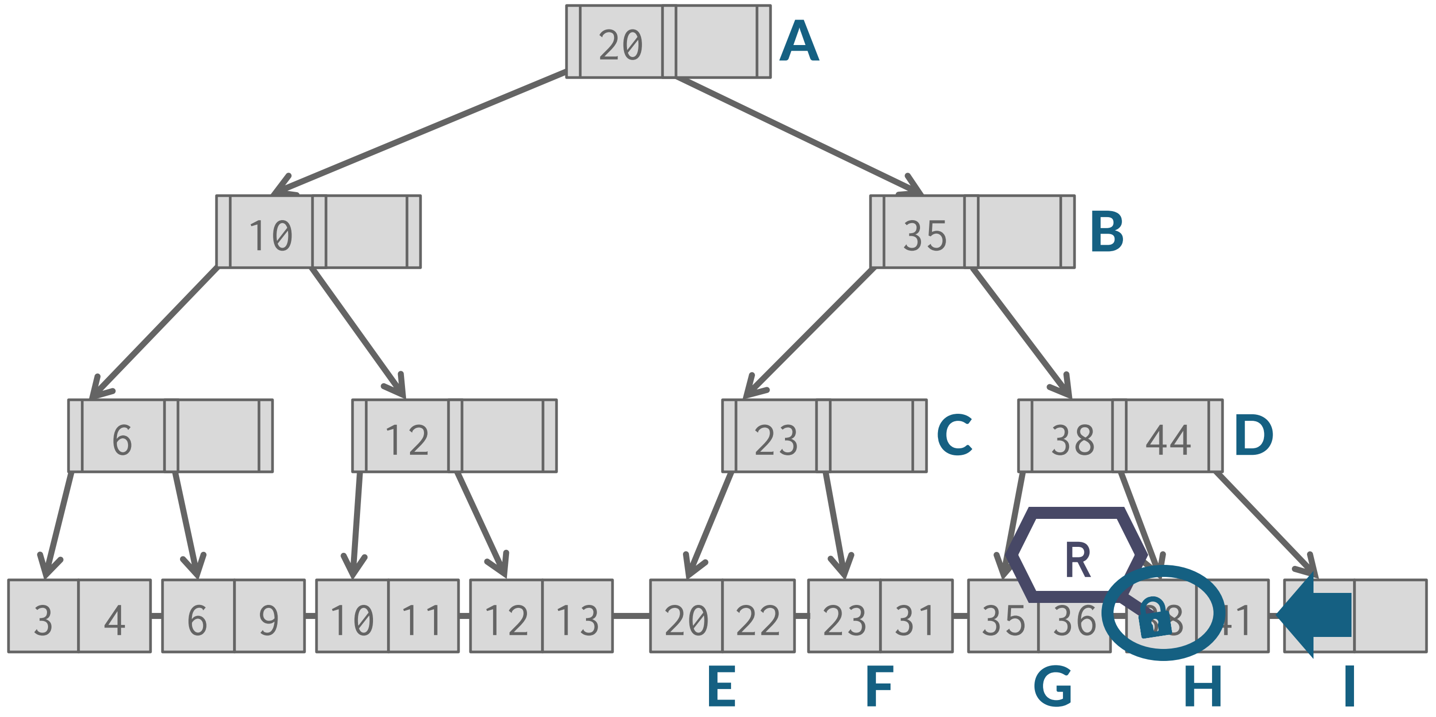


SCHOOL OF DATA SCIENCE

Example #1 – Find 38

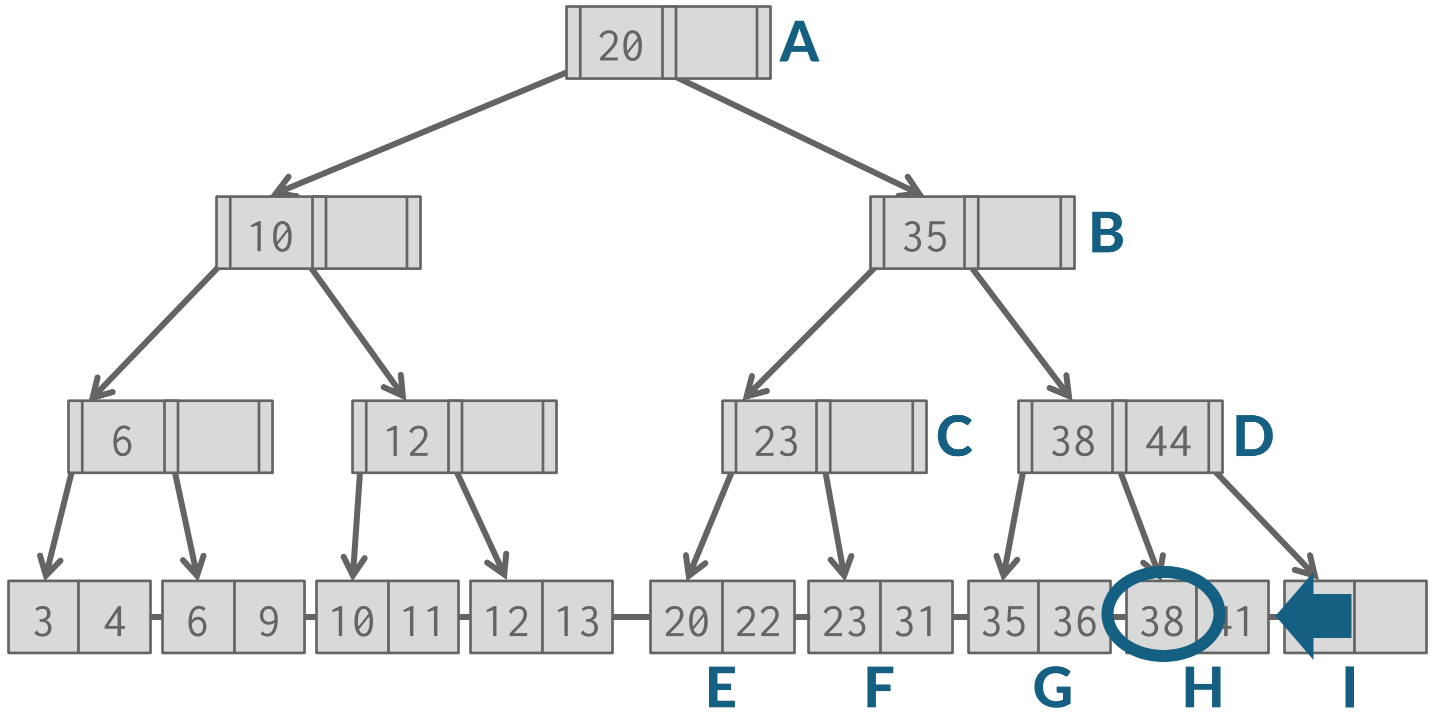


Example #1 – Find 38

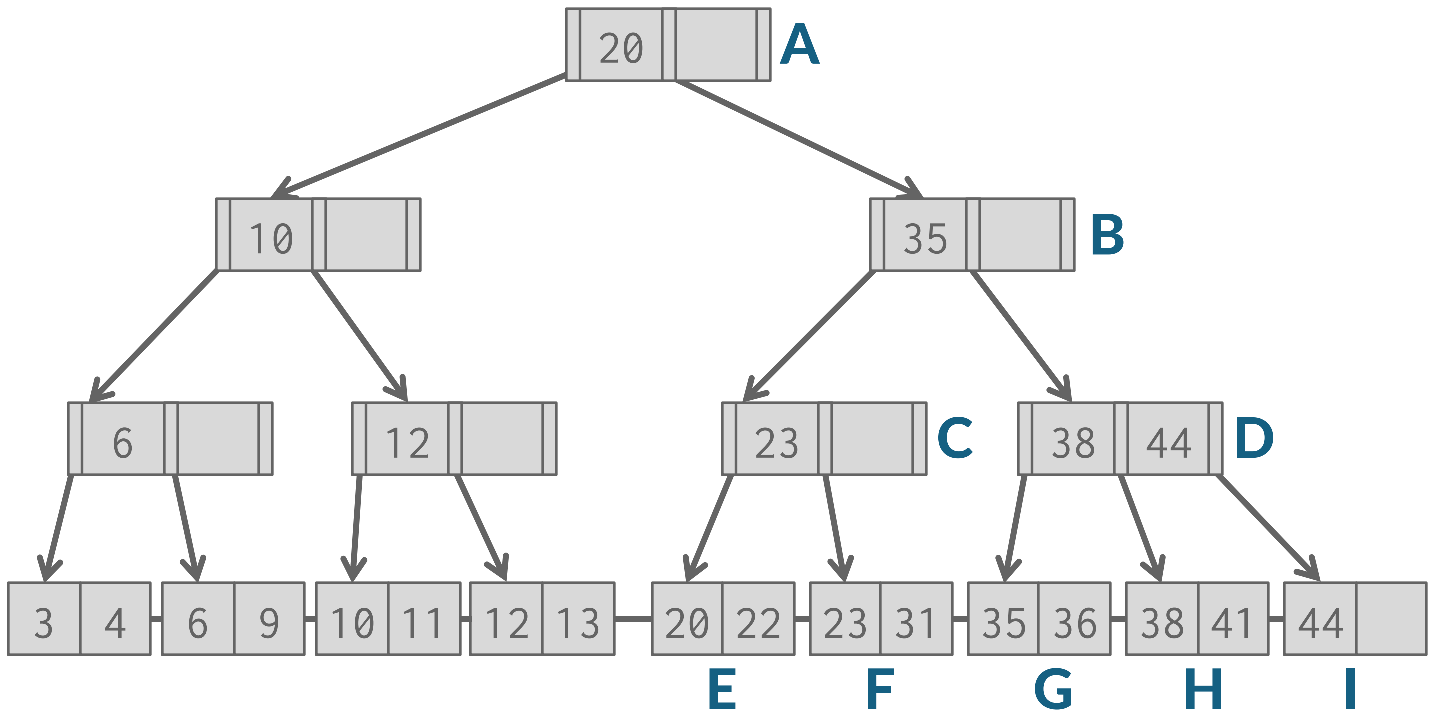


SCHOOL OF
DATA SCIENCE

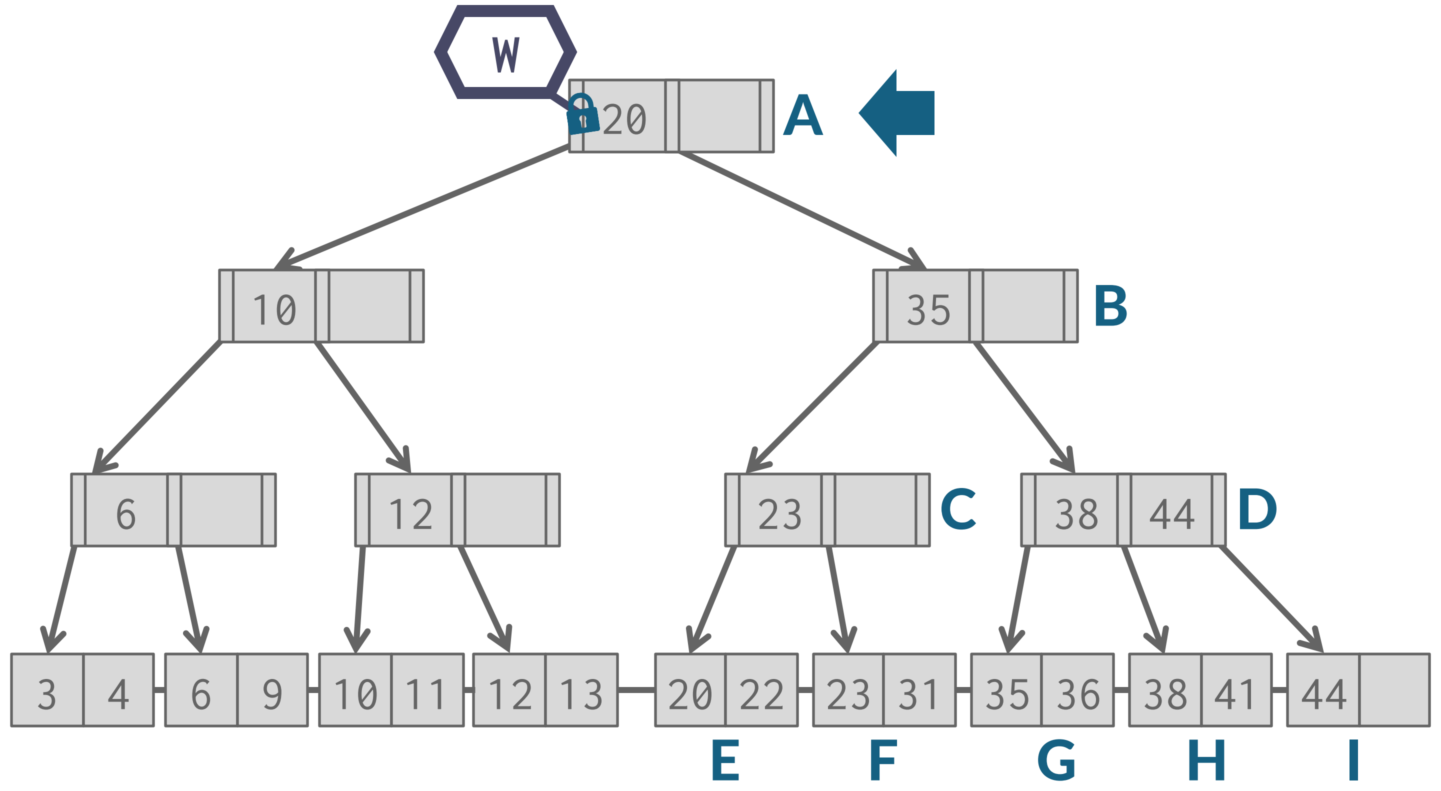
Example #1 – Find 38



Example #1 – Delete 38

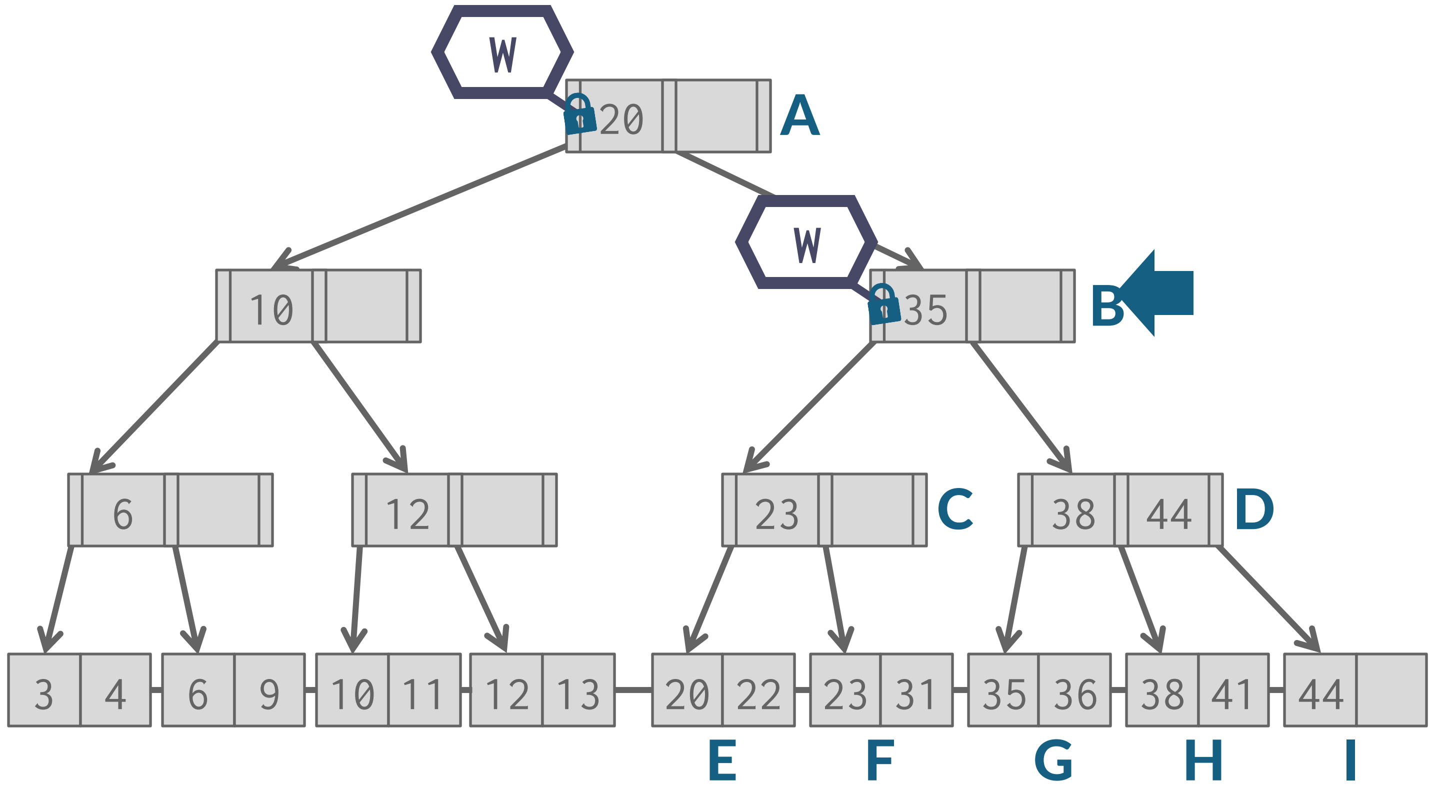


Example #1 – Delete 38



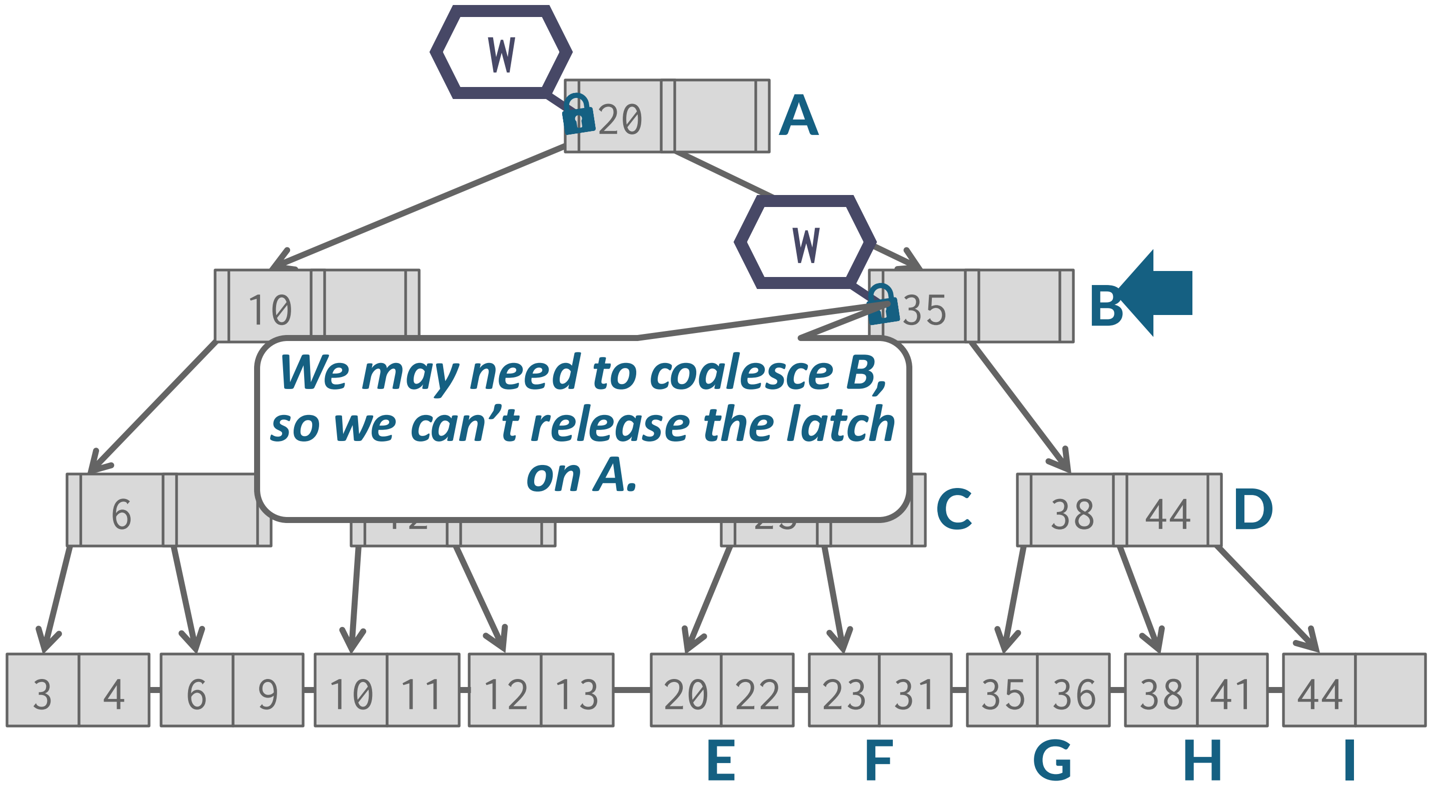
SCHOOL OF DATA SCIENCE

Example #1 – Delete 38



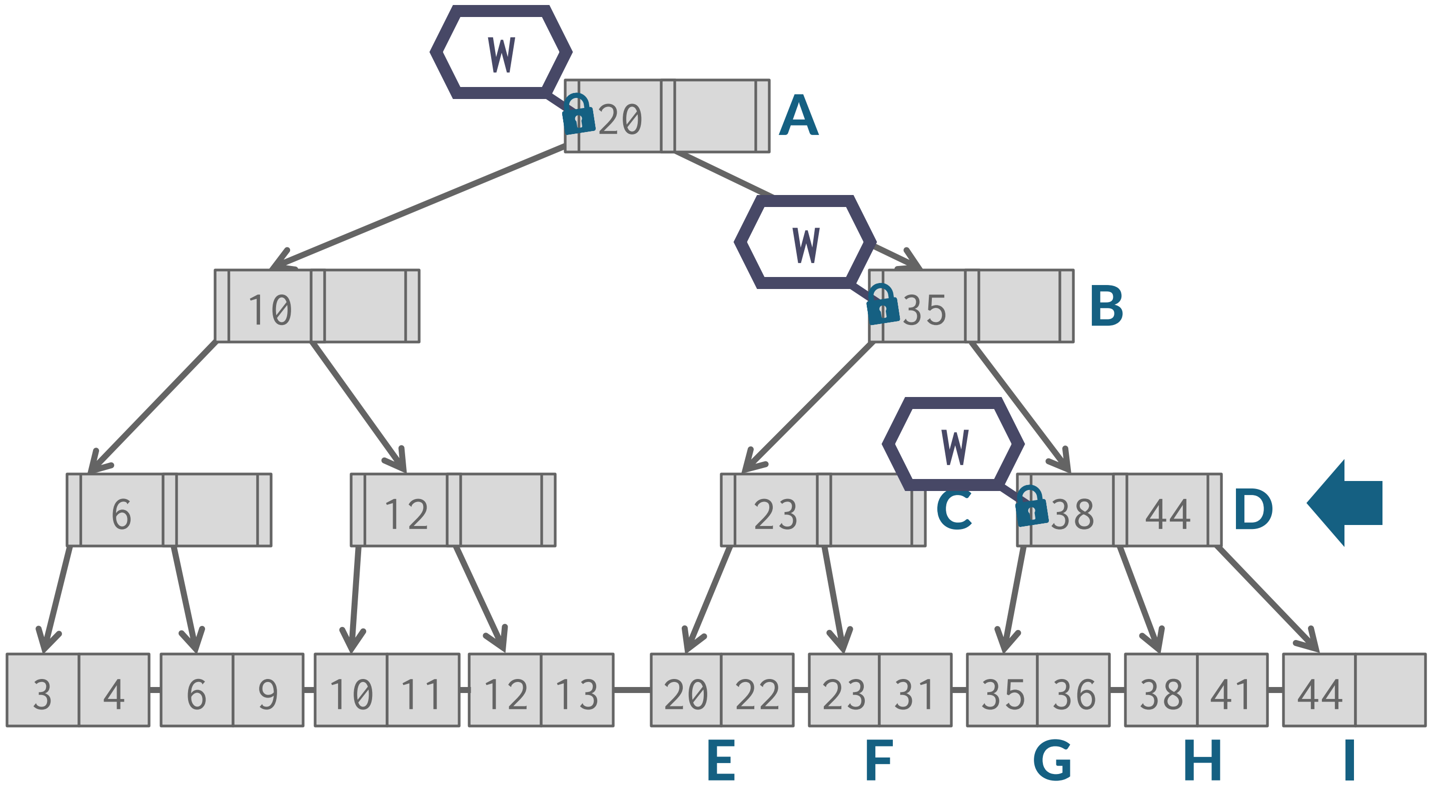
SCHOOL OF DATA SCIENCE

Example #1 – Delete 38

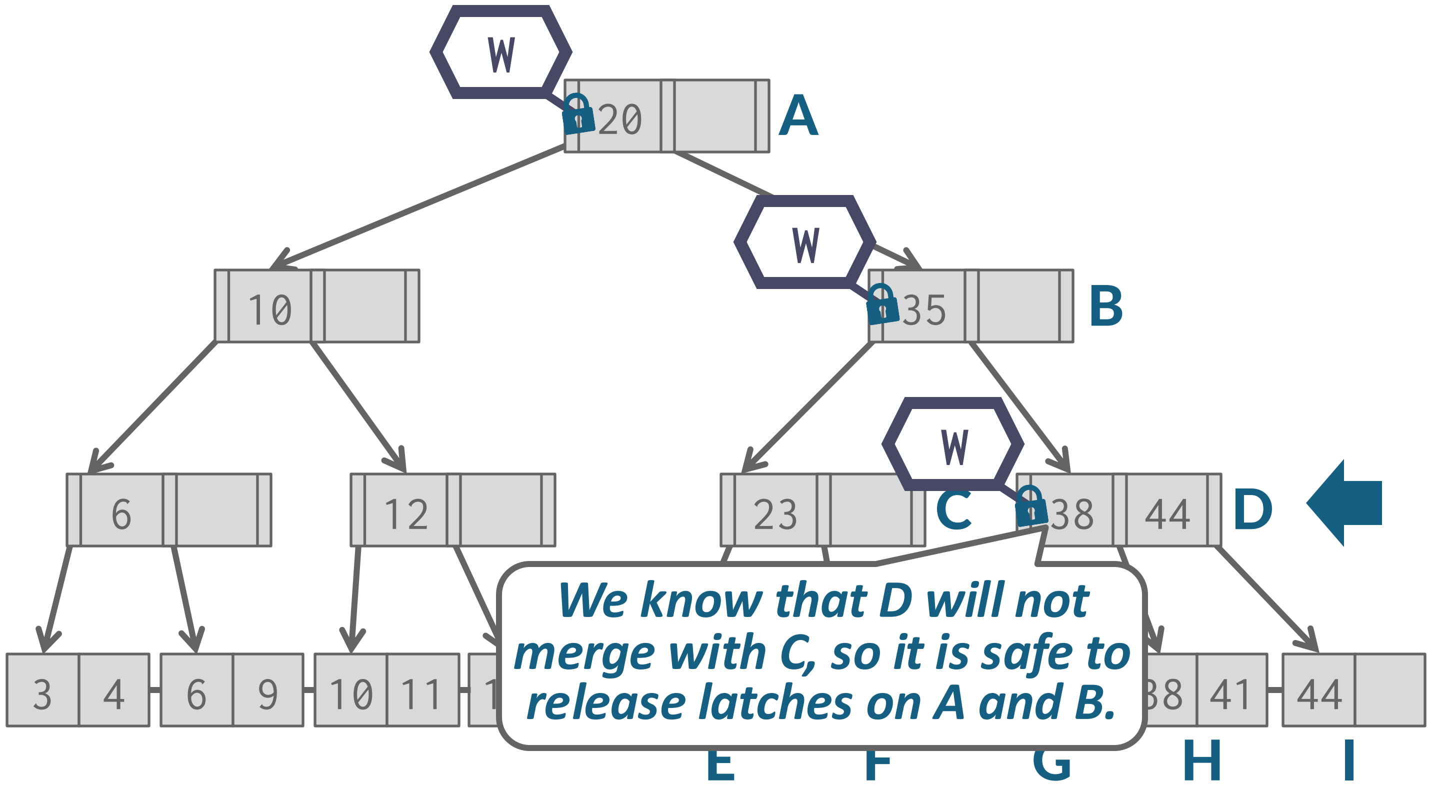


We may need to coalesce B, so we can't release the latch on A.

Example #1 – Delete 38

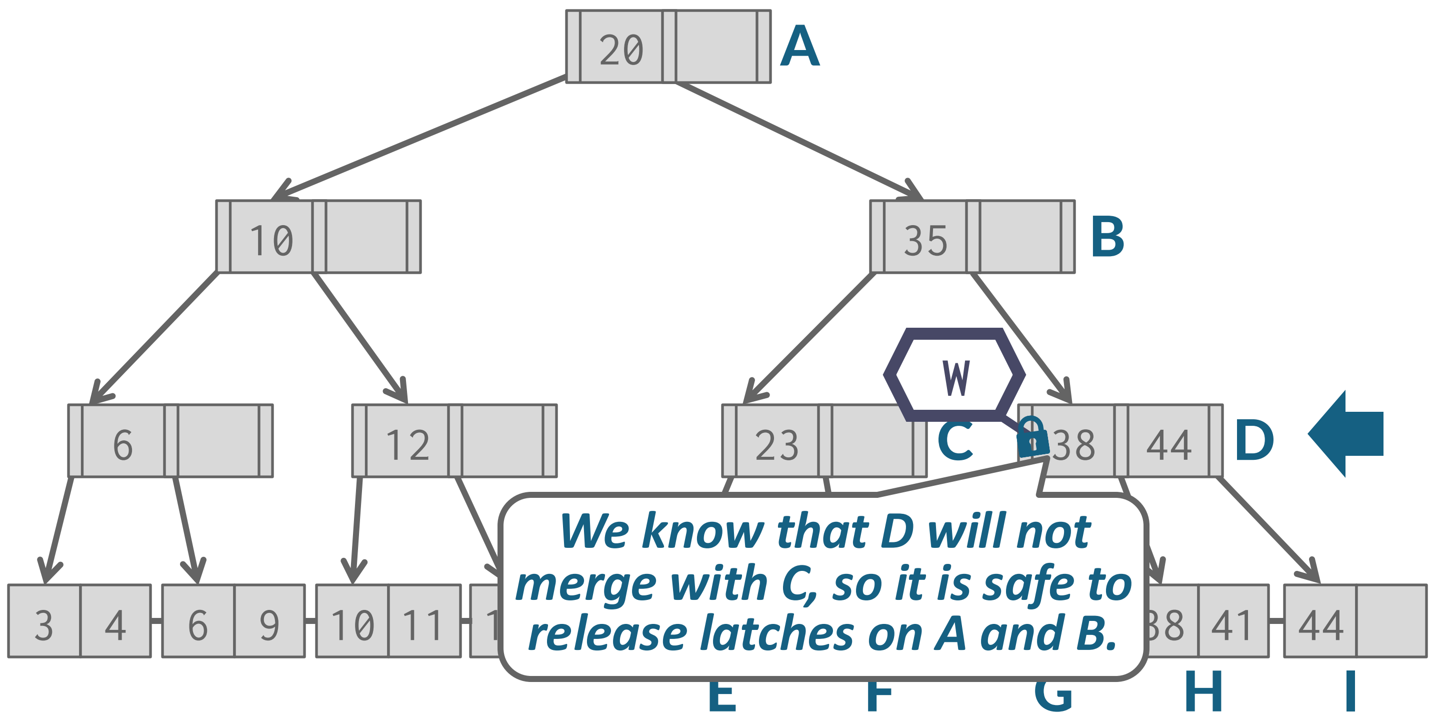


Example #1 – Delete 38



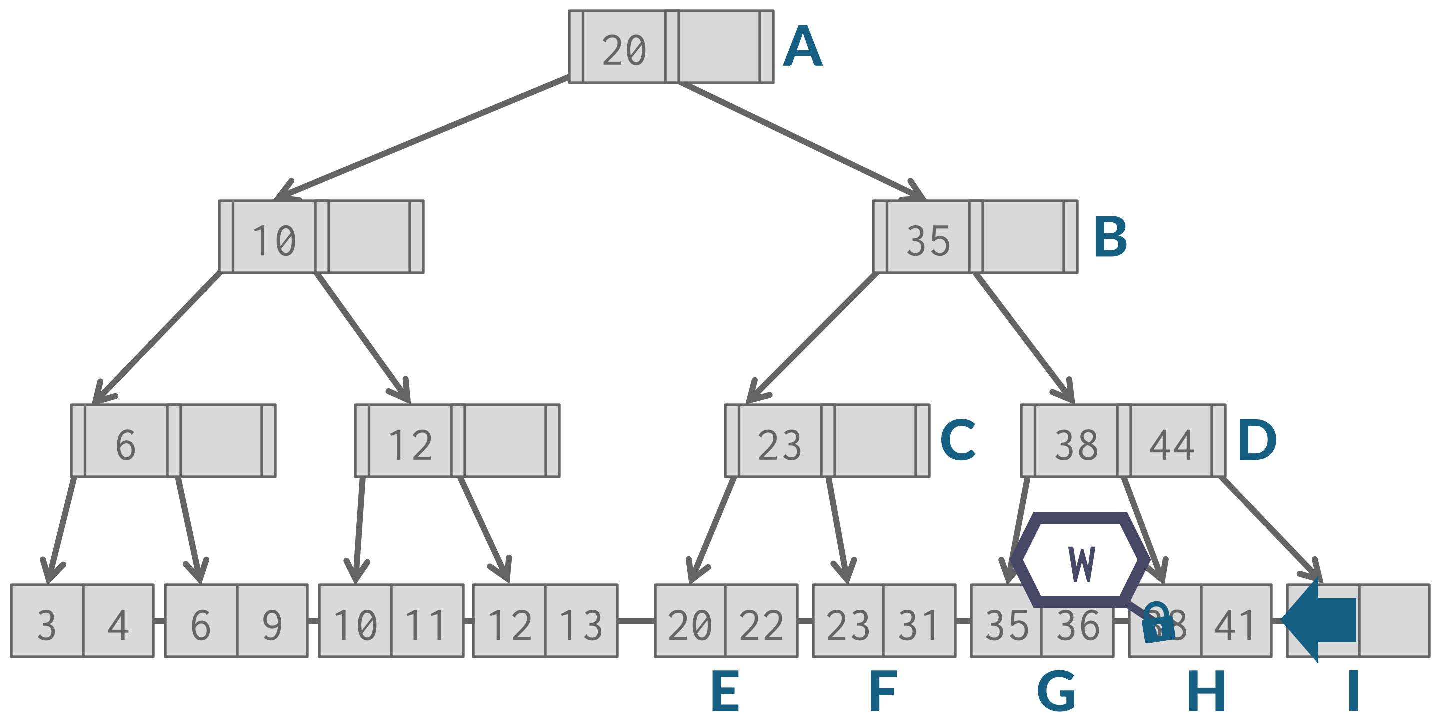
We know that D will not merge with C, so it is safe to release latches on A and B.

Example #1 – Delete 38



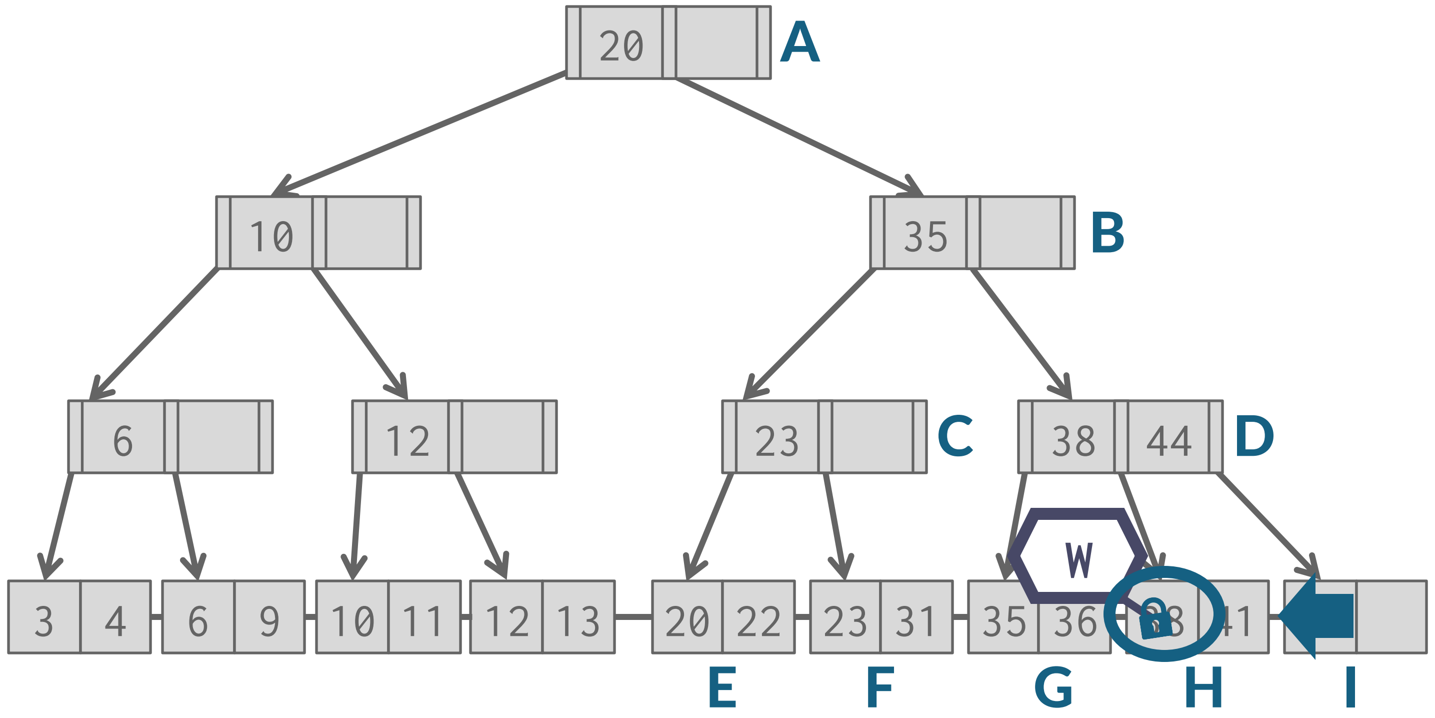
We know that D will not merge with C, so it is safe to release latches on A and B.

Example #1 – Delete 38

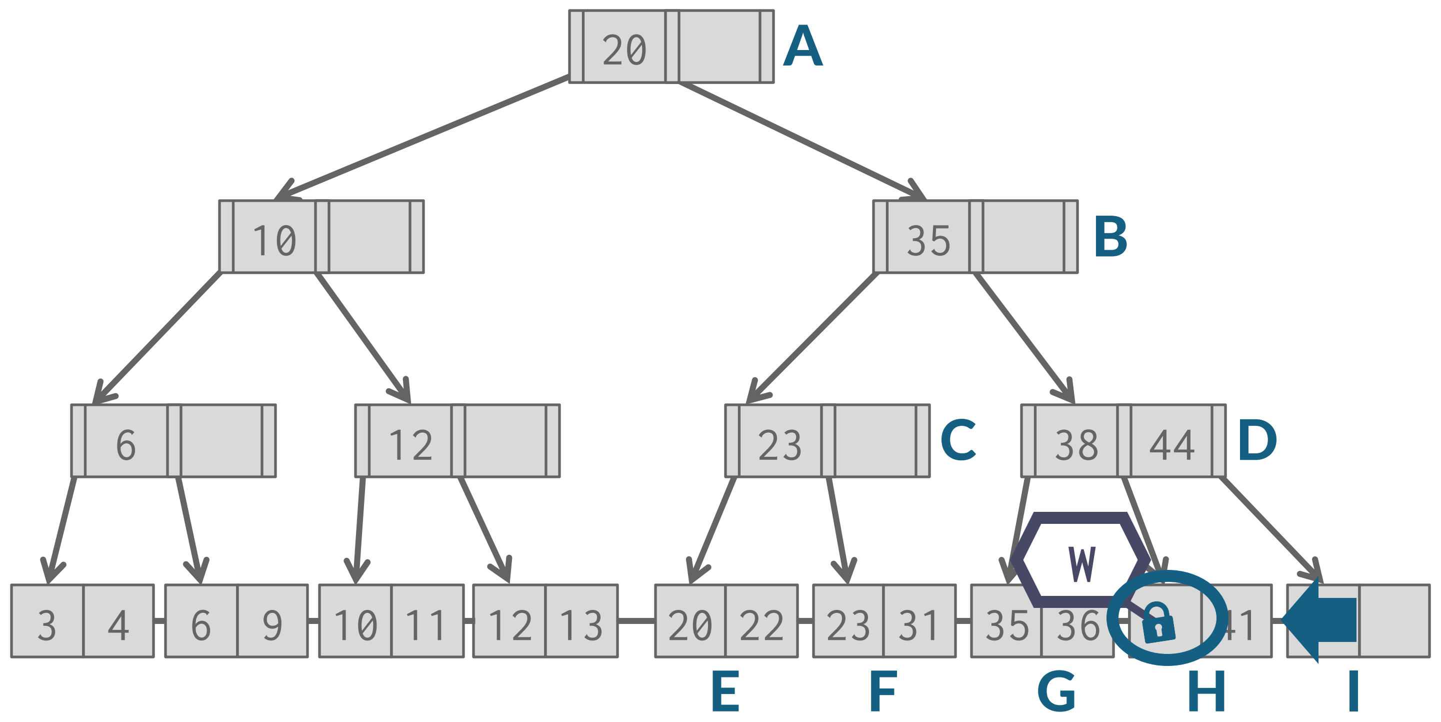


SCHOOL OF
DATA SCIENCE

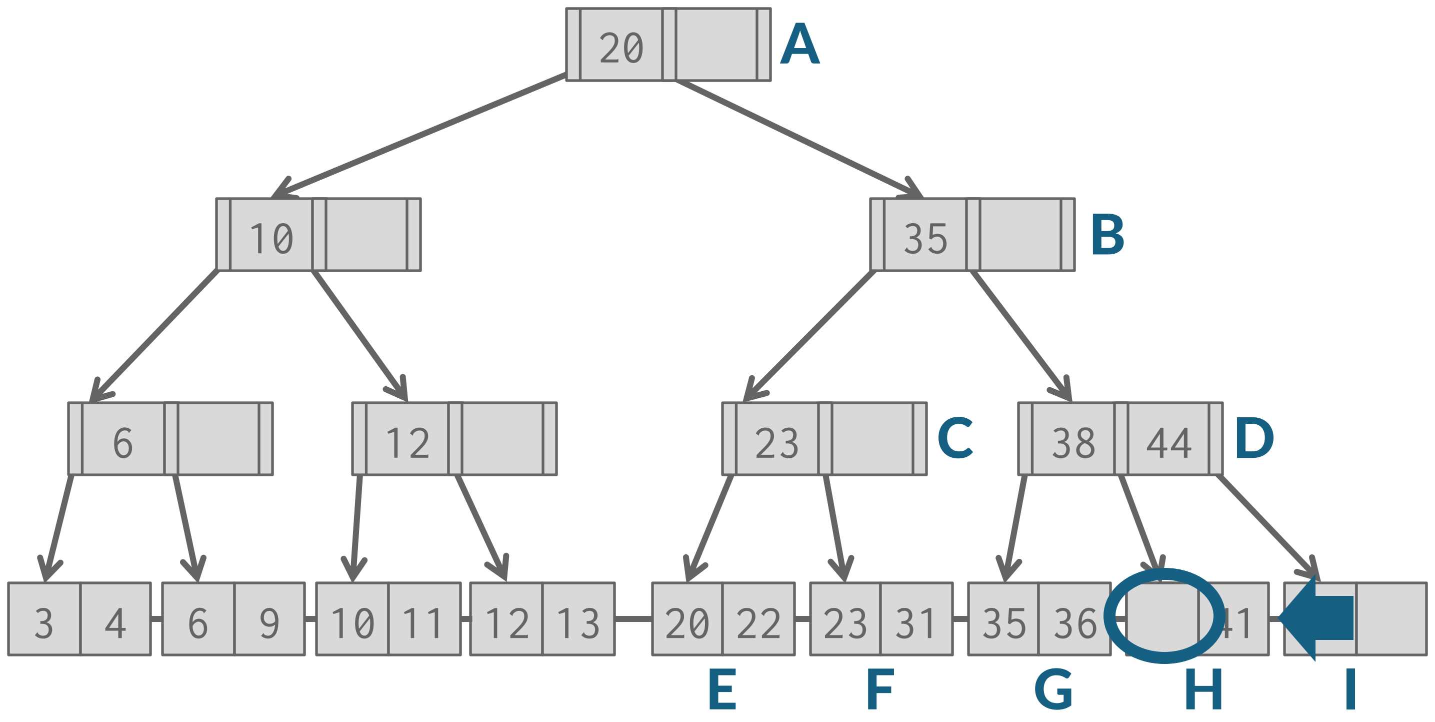
Example #1 – Delete 38



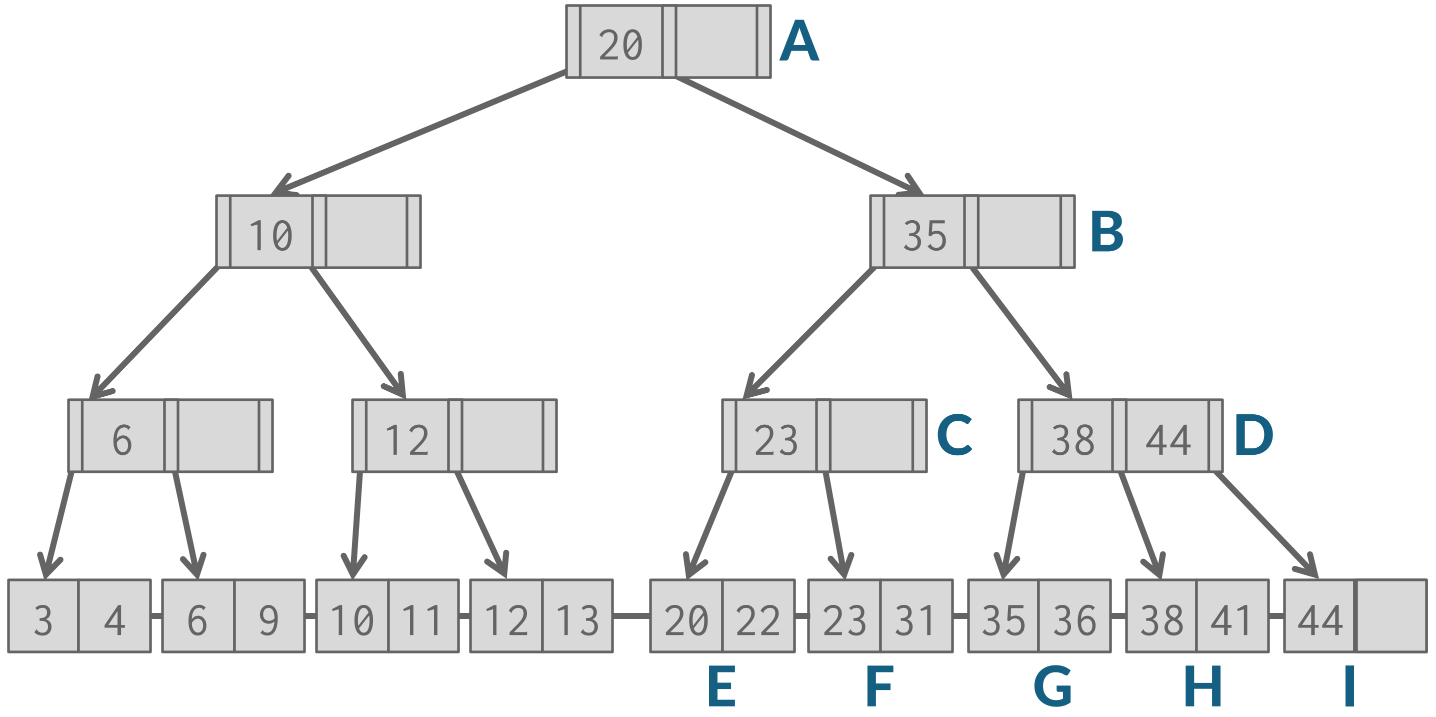
Example #1 – Delete 38



Example #1 – Delete 38

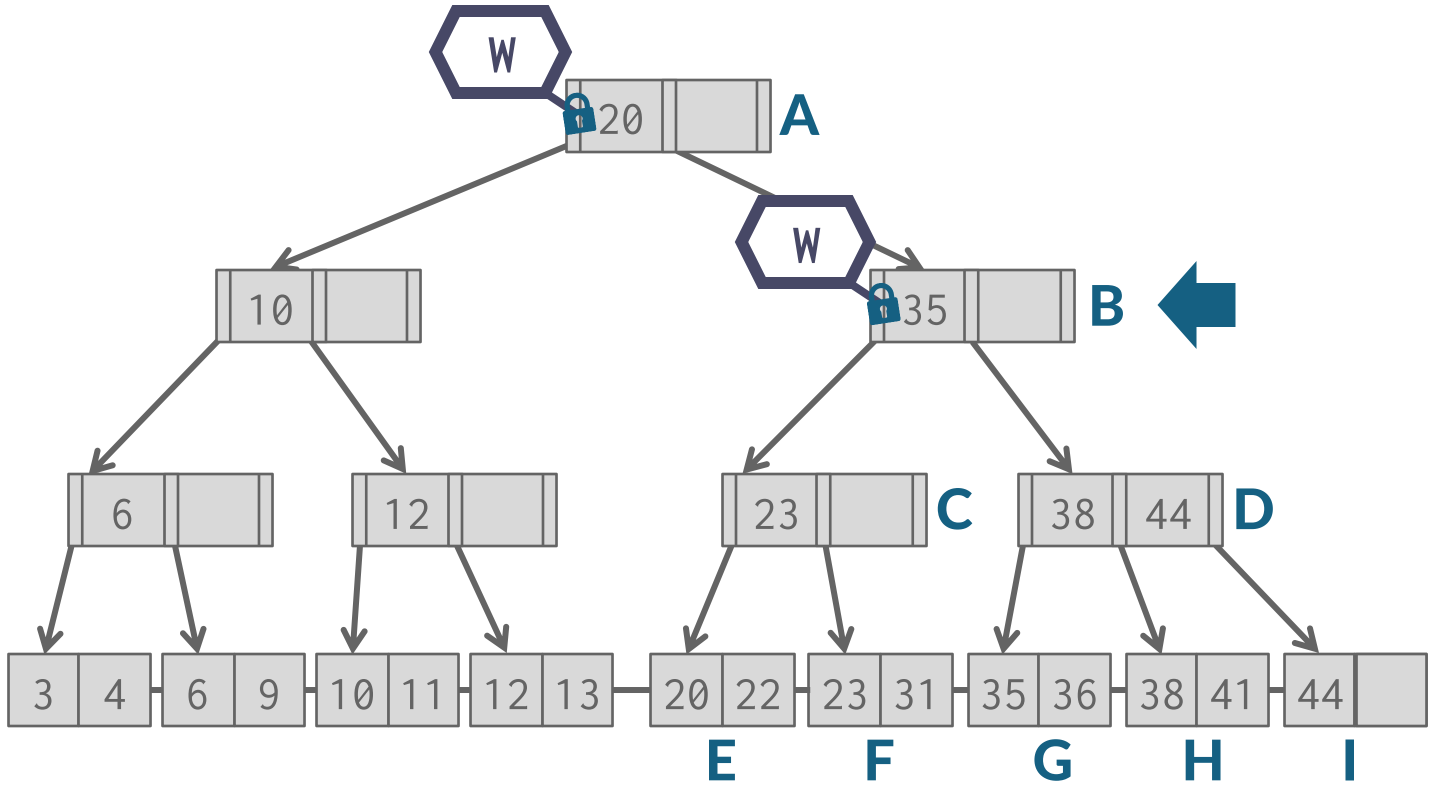


Example #1 – Insert 45

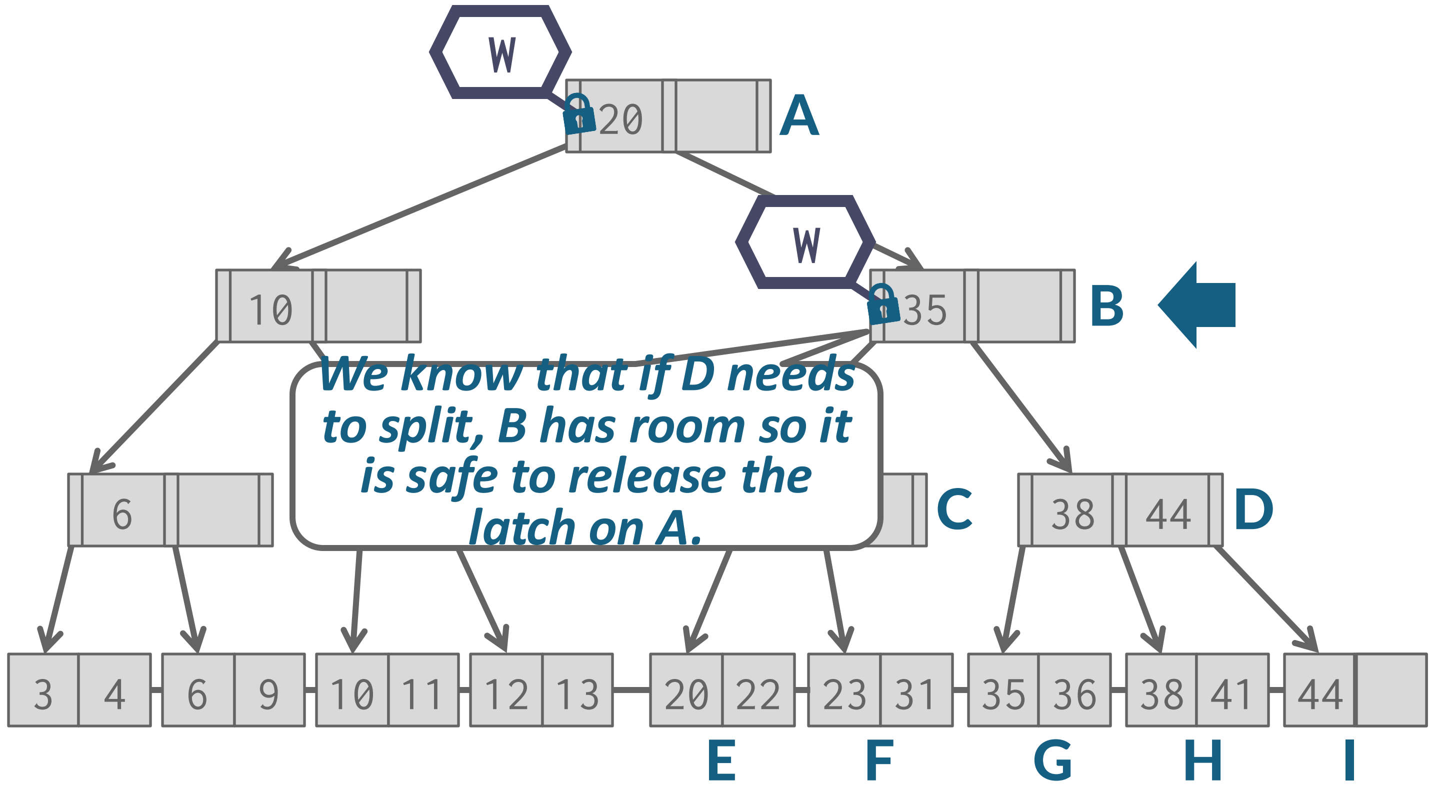


SCHOOL OF DATA SCIENCE

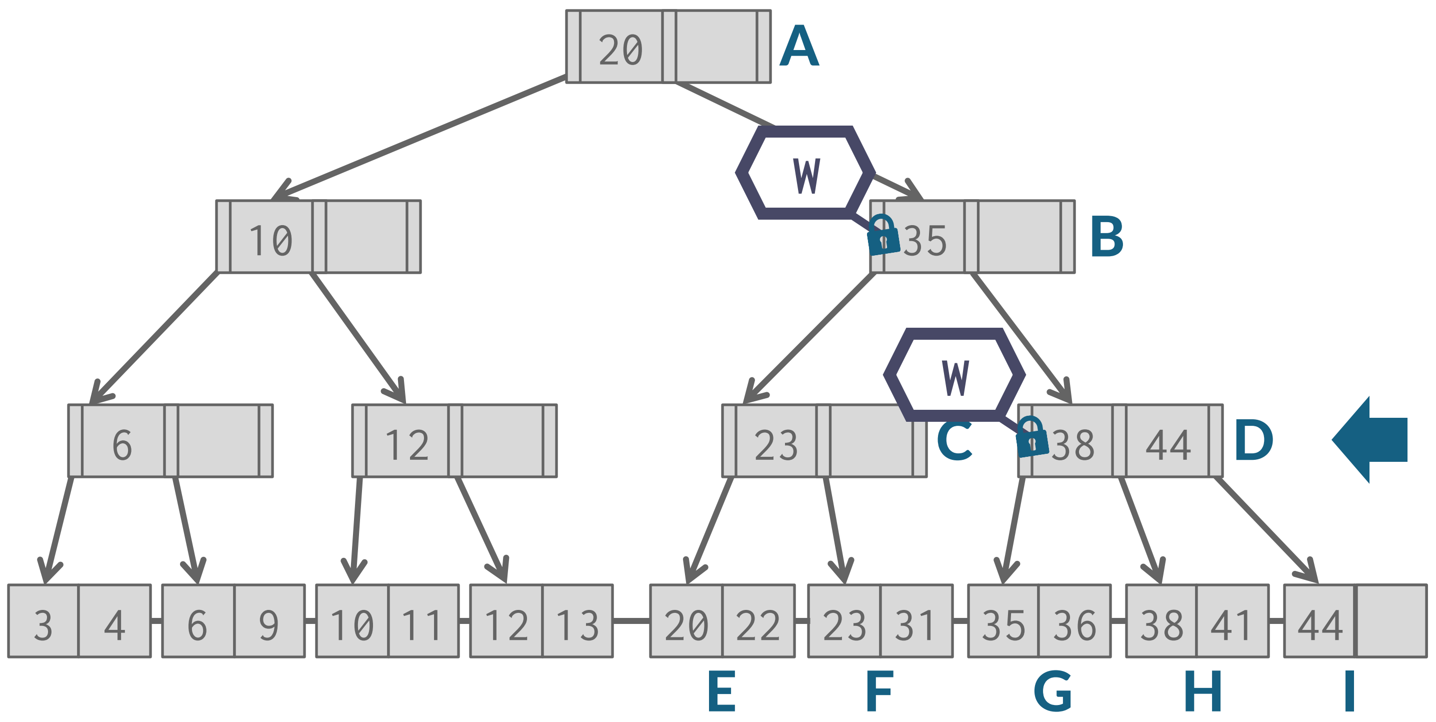
Example #1 – Insert 45



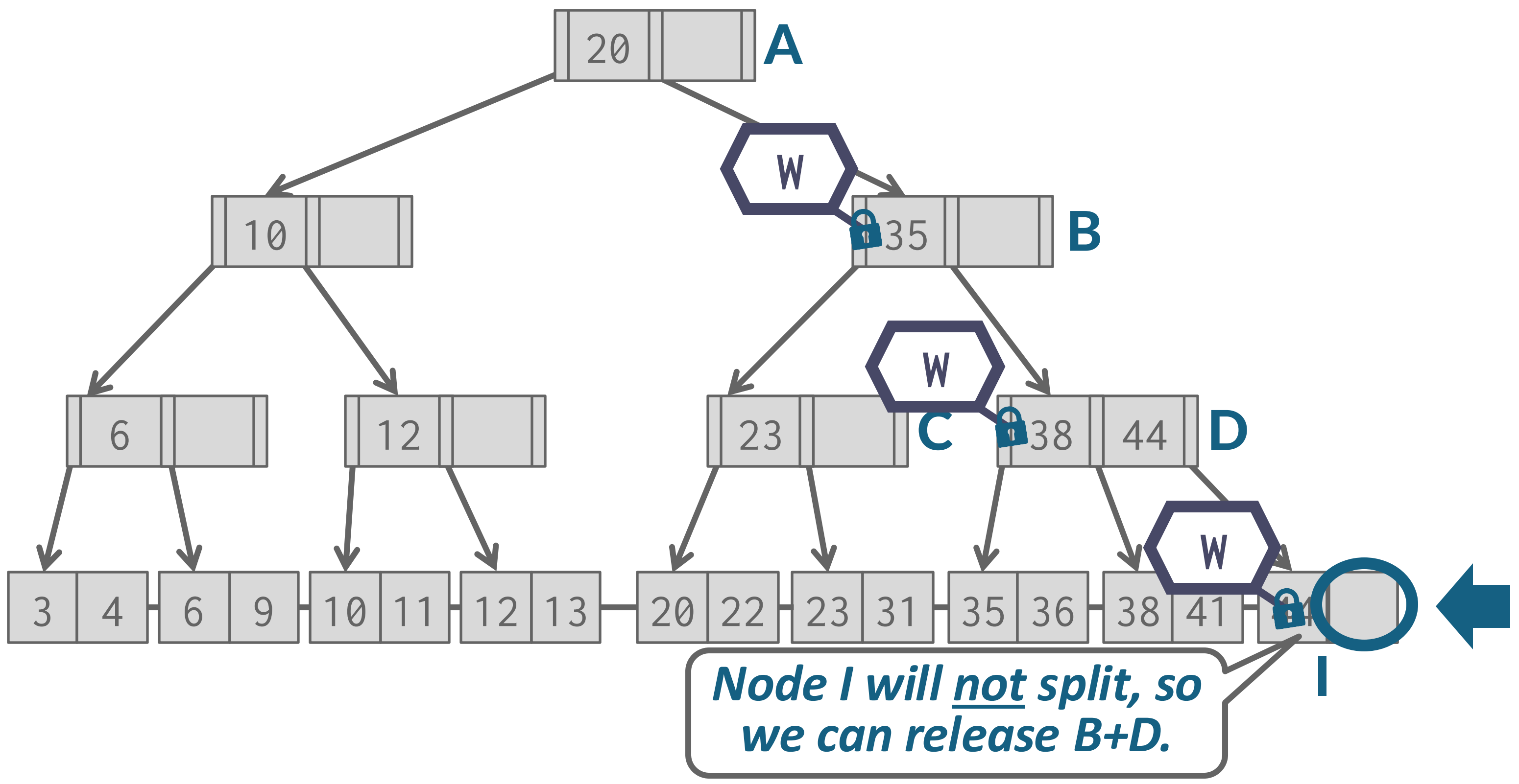
Example #1 – Insert 45


W
20
A
W
35
B
10
We know that if D needs to split, B has room so it is safe to release the latch on A.
6
C
38
44
D
3
4
6
9
10
11
12
13
20
22
23
31
35
36
38
41
44
E
F
G
H
I

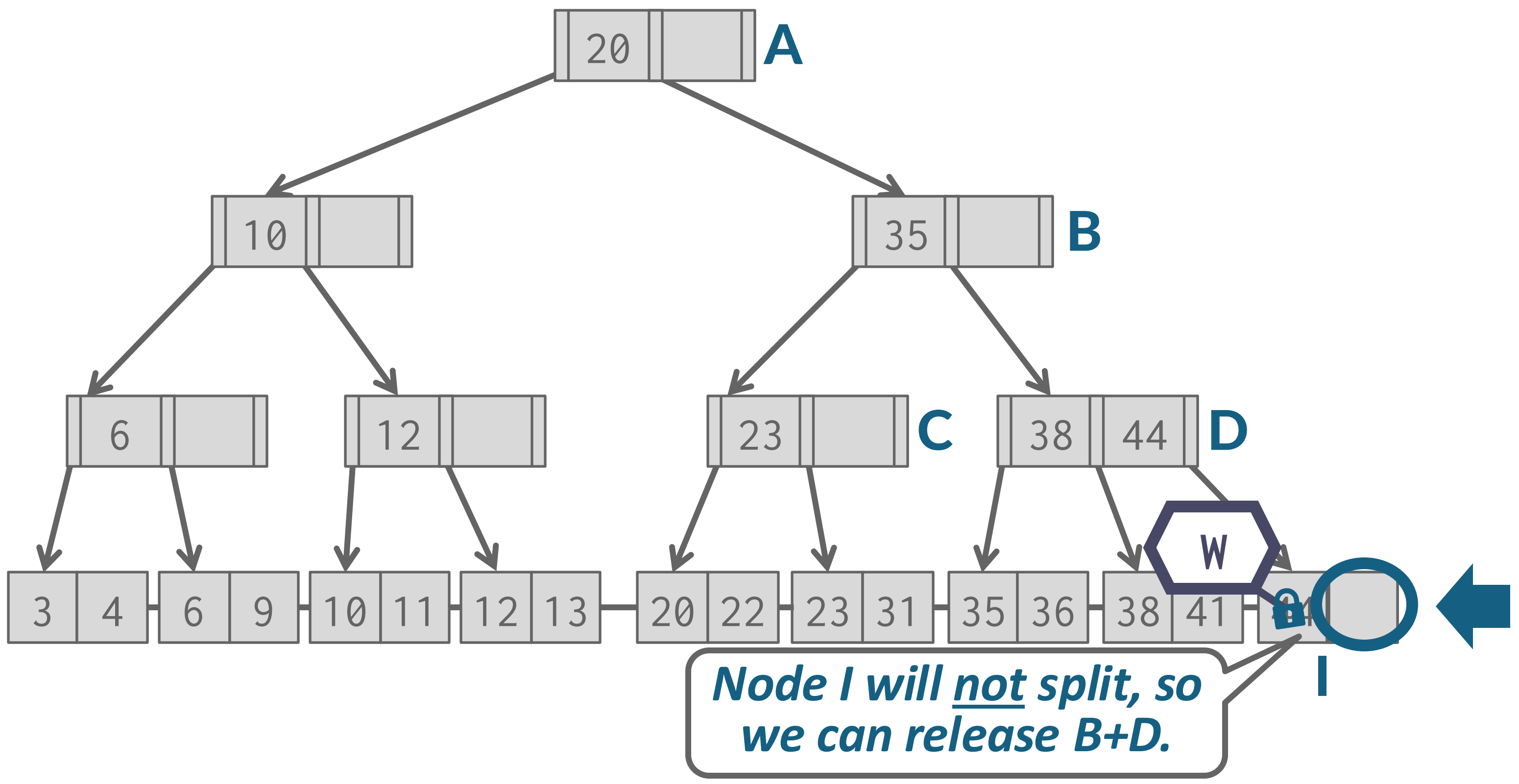
Example #1 – Insert 45



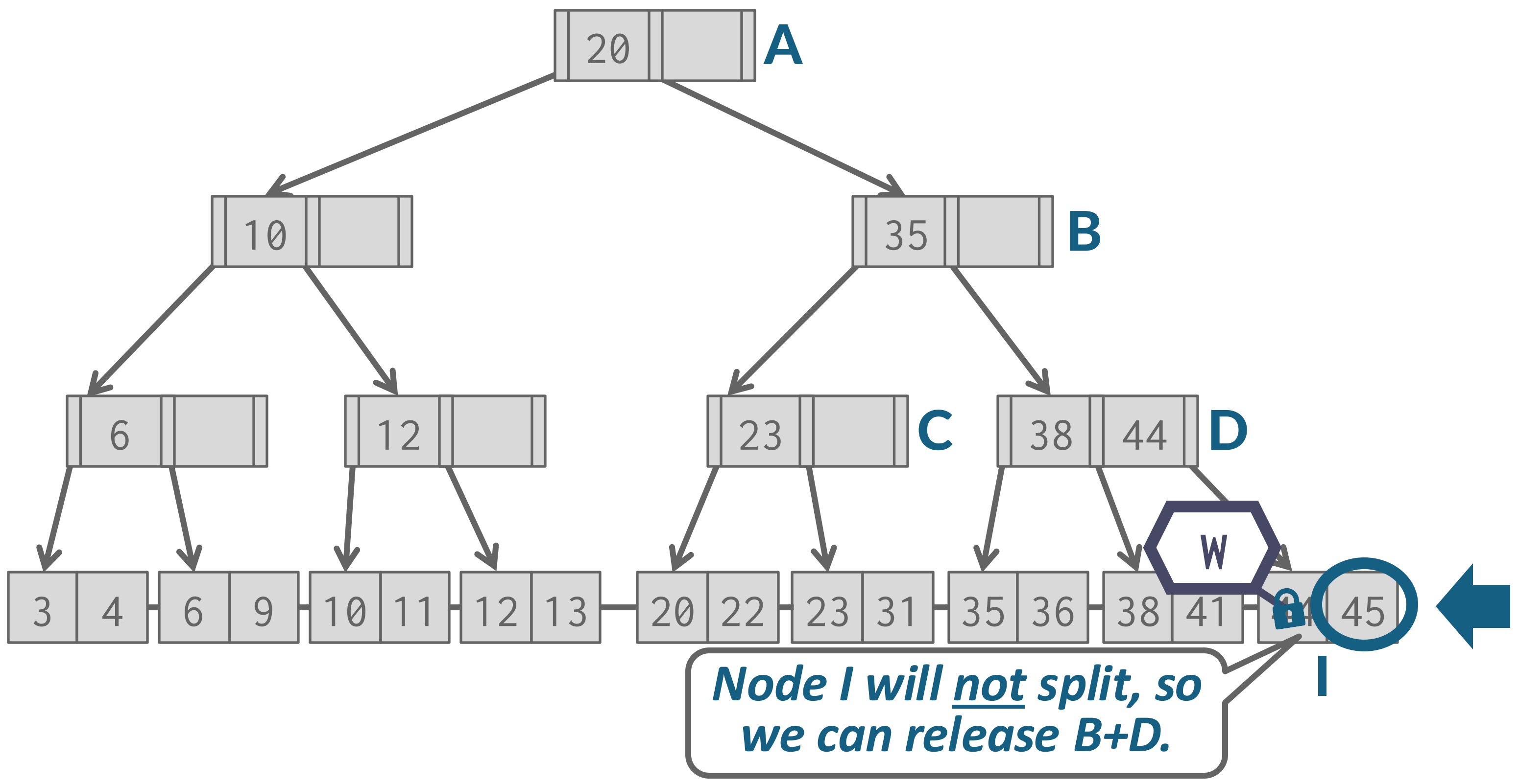
Example #1 – Insert 45



Example #1 – Insert 45



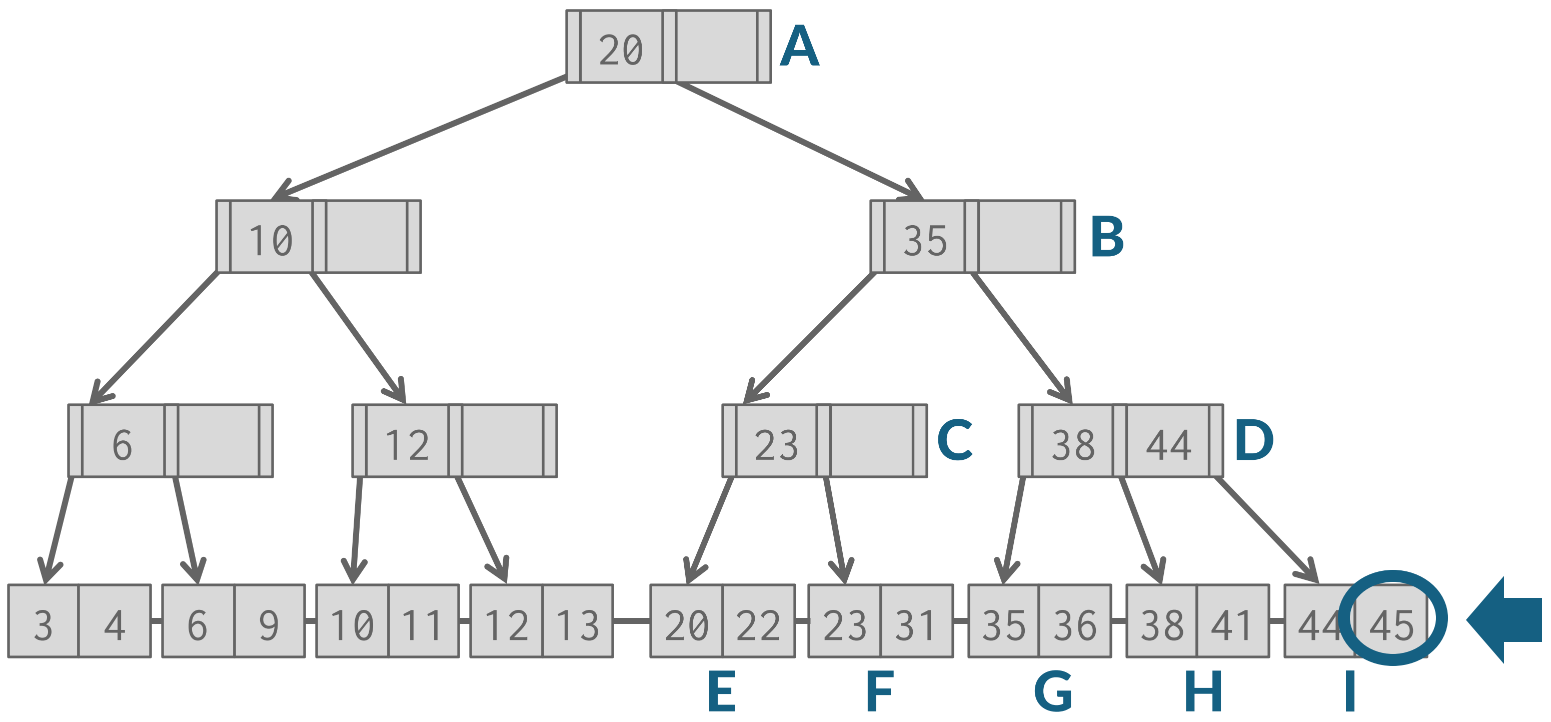
Example #1 – Insert 45



Node I will not split, so we can release B+D.

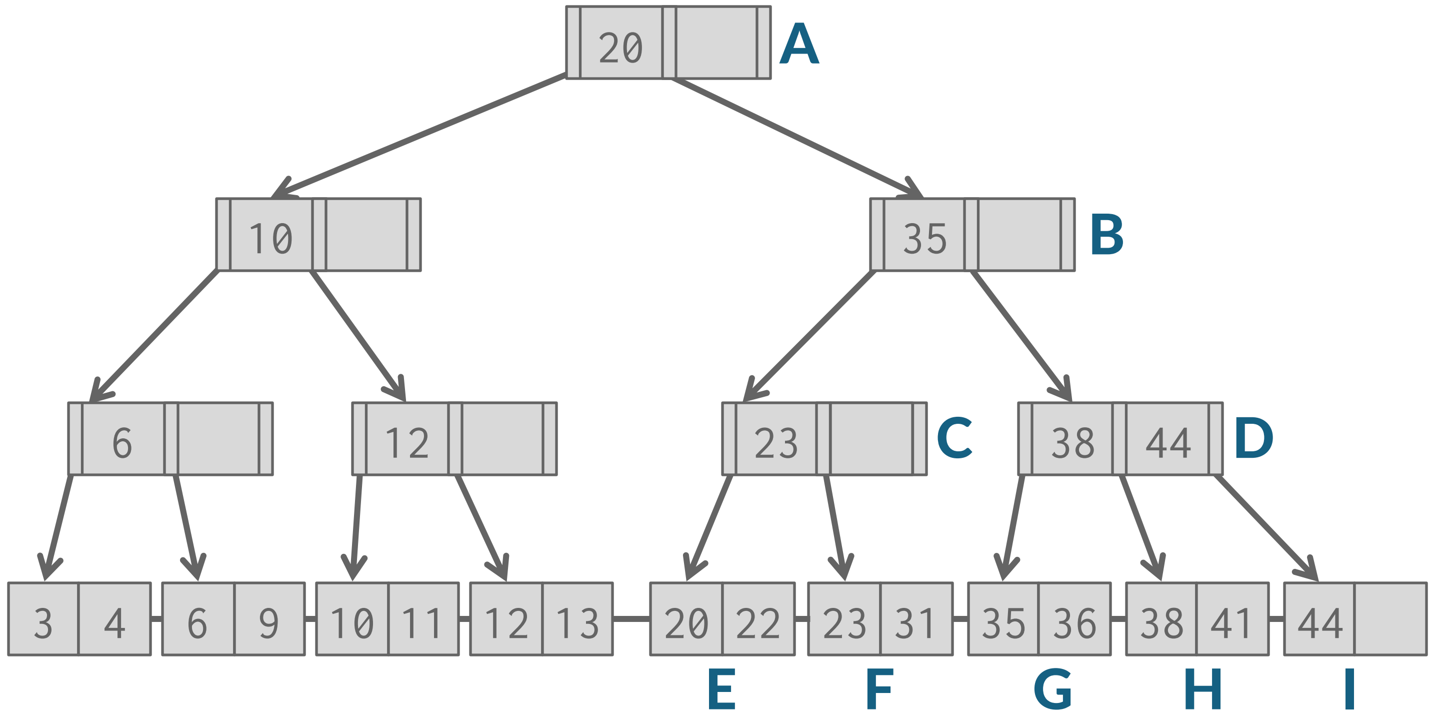
SCHOOL OF
DATA SCIENCE

Example #1 – Insert 45

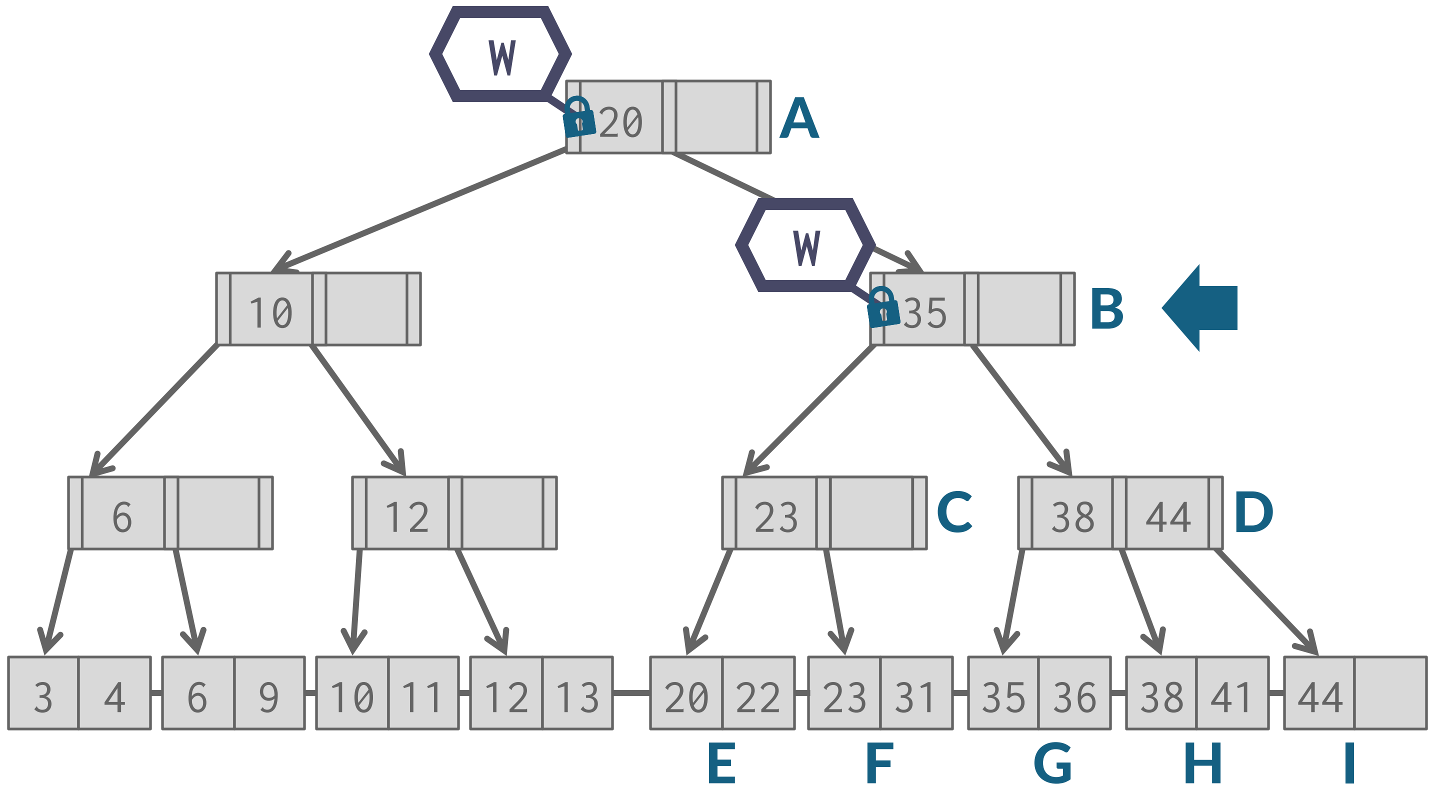


SCHOOL OF
DATA SCIENCE

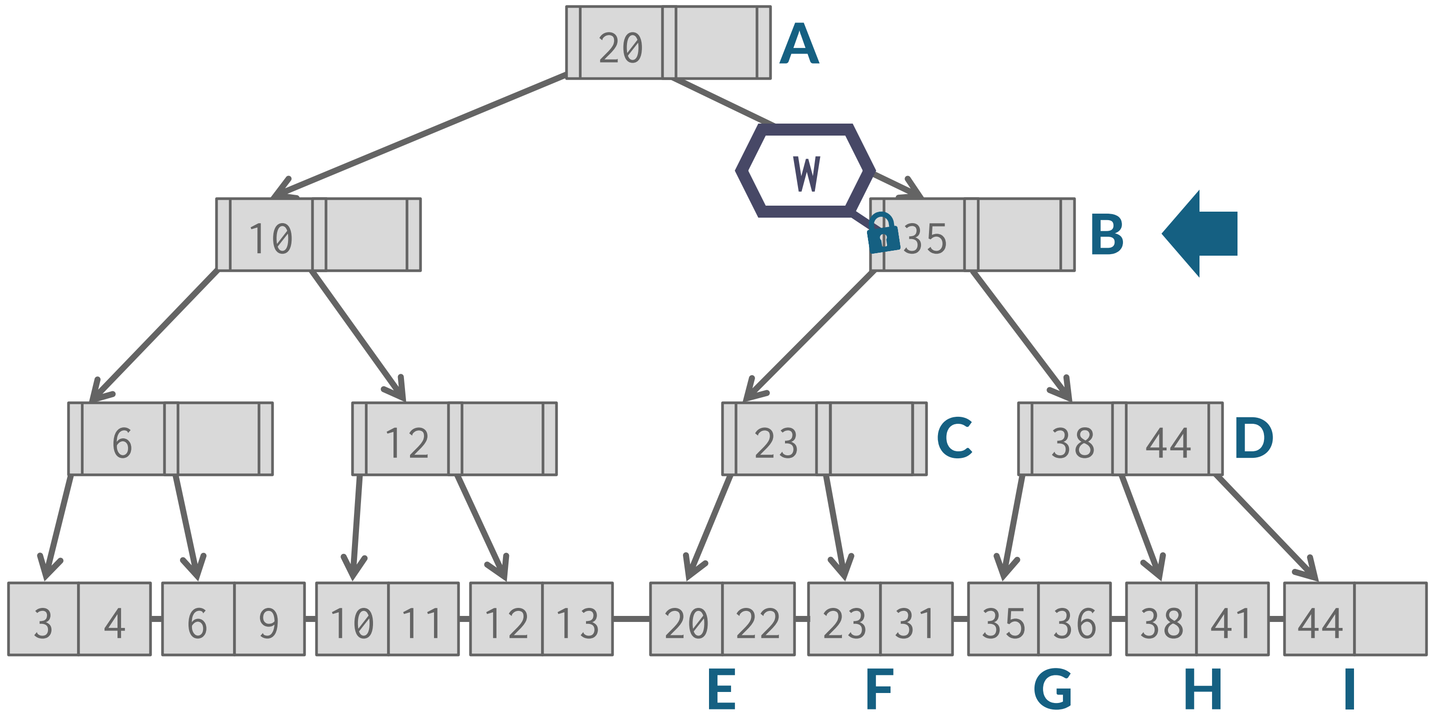
Example #1 – Insert 25



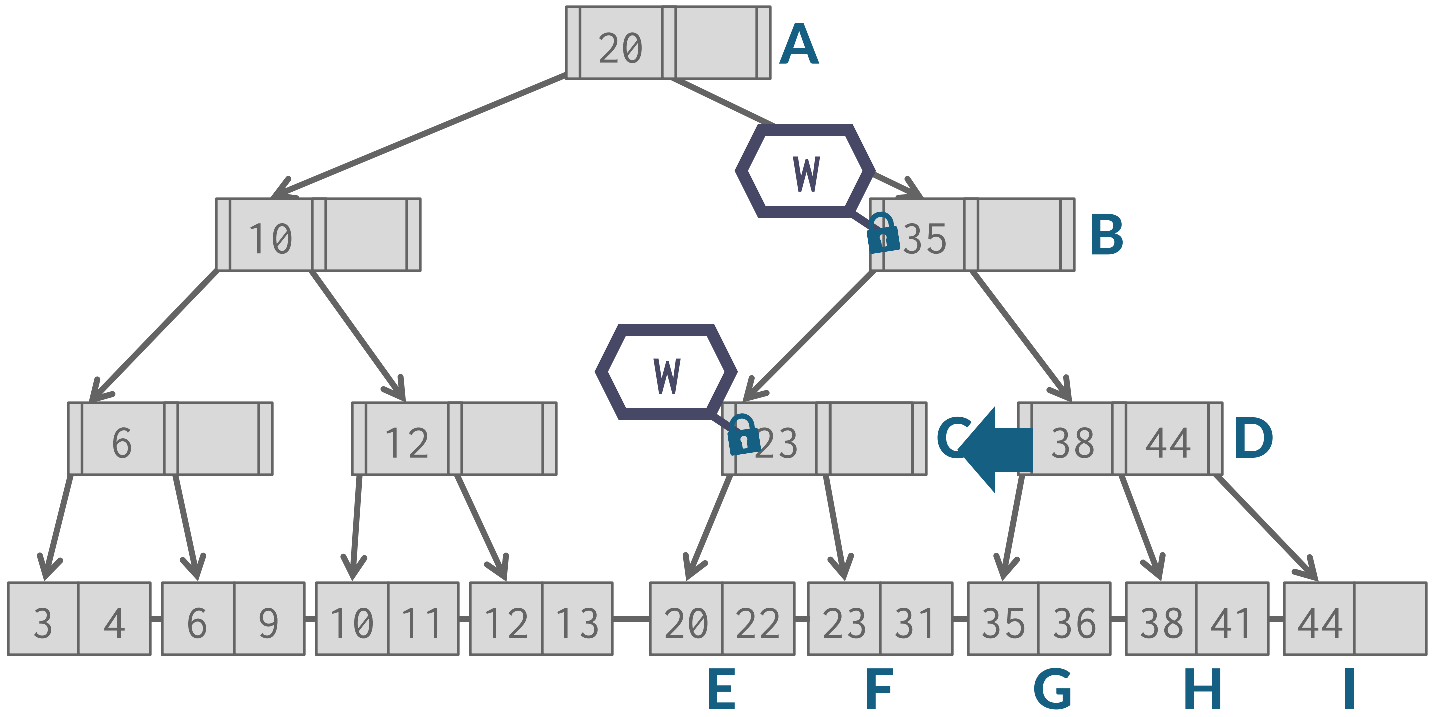
Example #1 – Insert 25



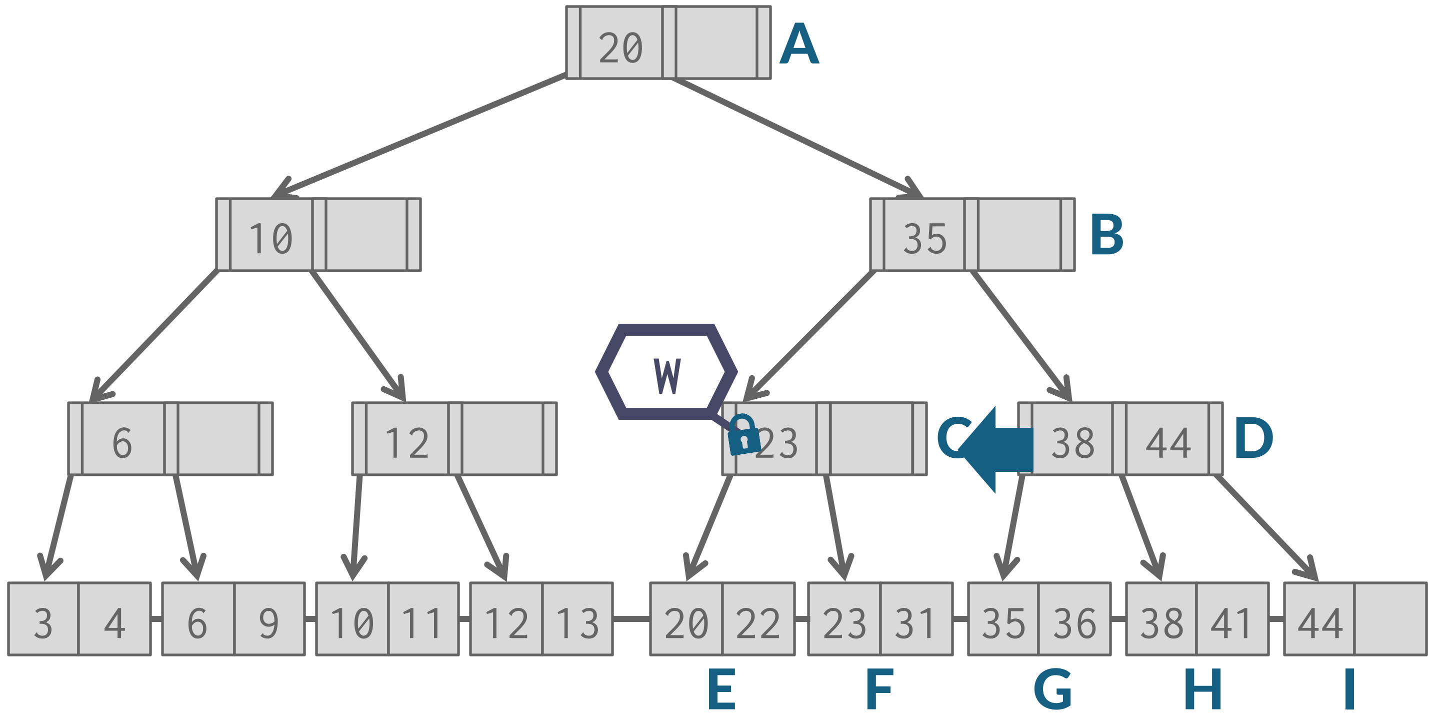
Example #1 – Insert 25



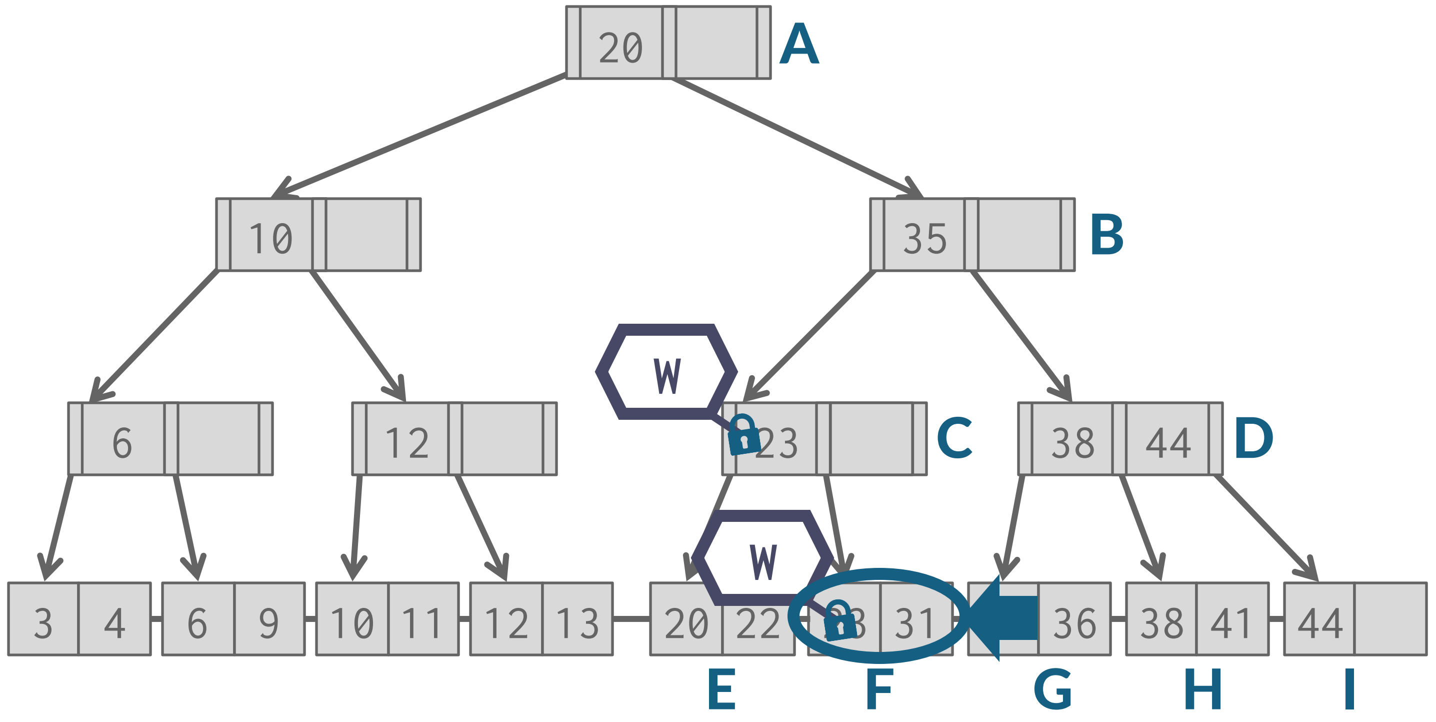
Example #1 – Insert 25



Example #1 – Insert 25

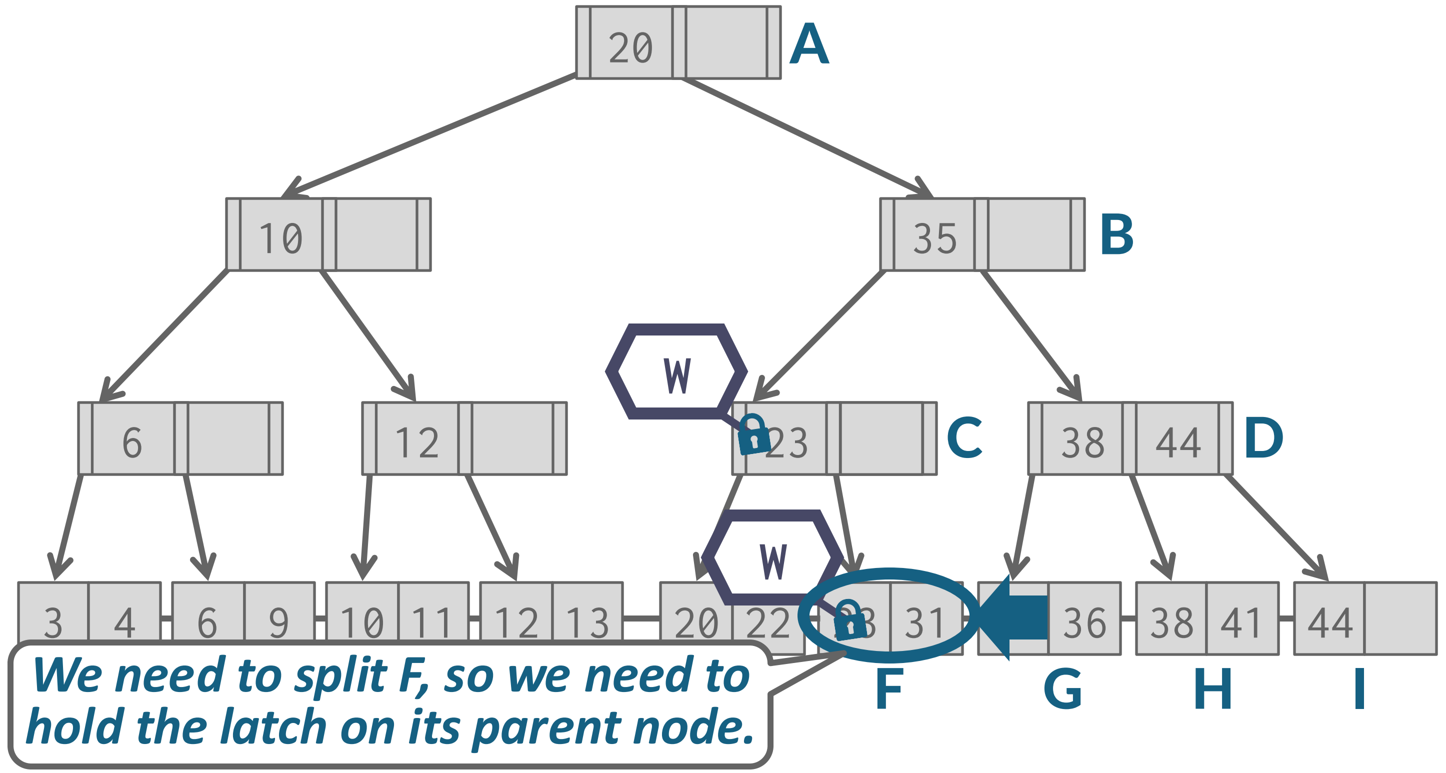


Example #1 – Insert 25



SCHOOL OF DATA SCIENCE

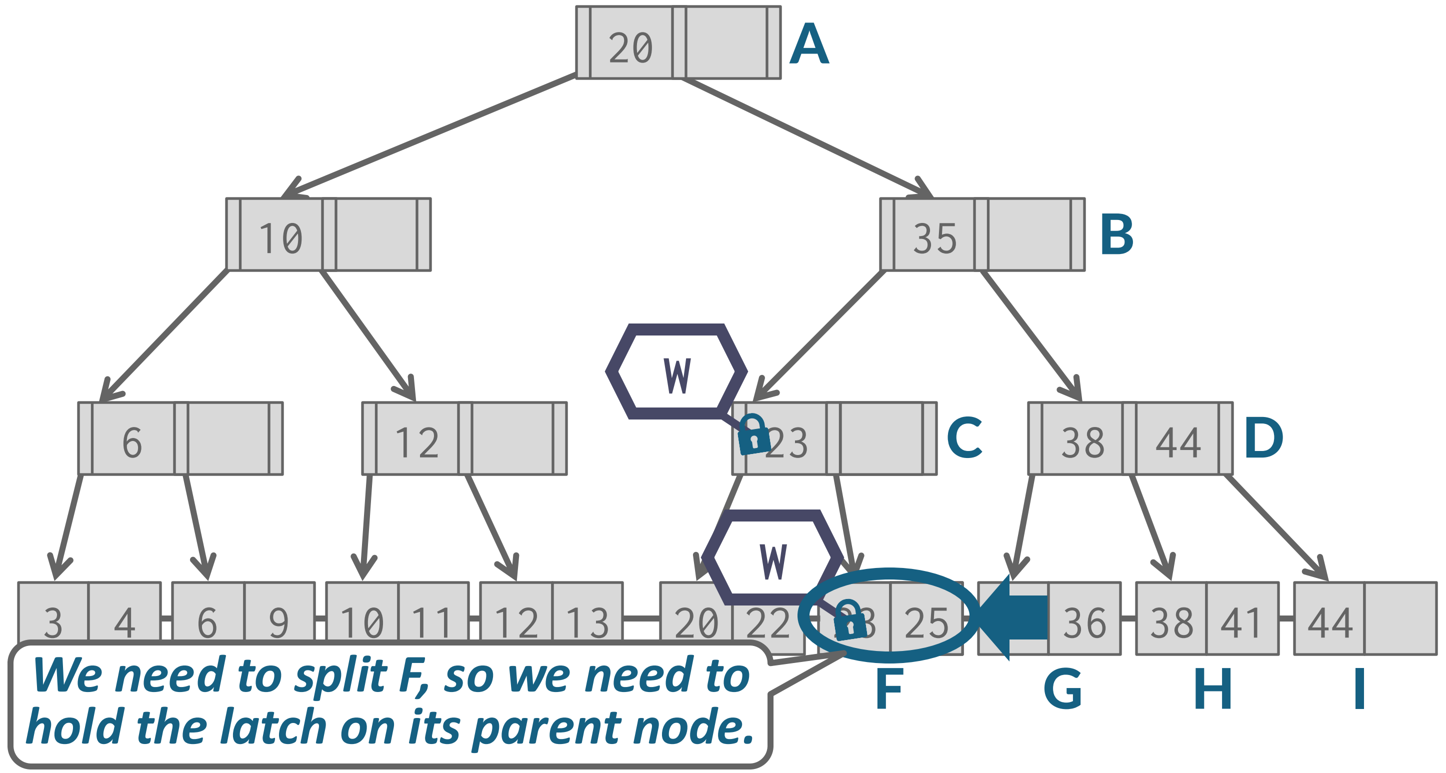
Example #1 – Insert 25



We need to split F, so we need to hold the latch on its parent node.

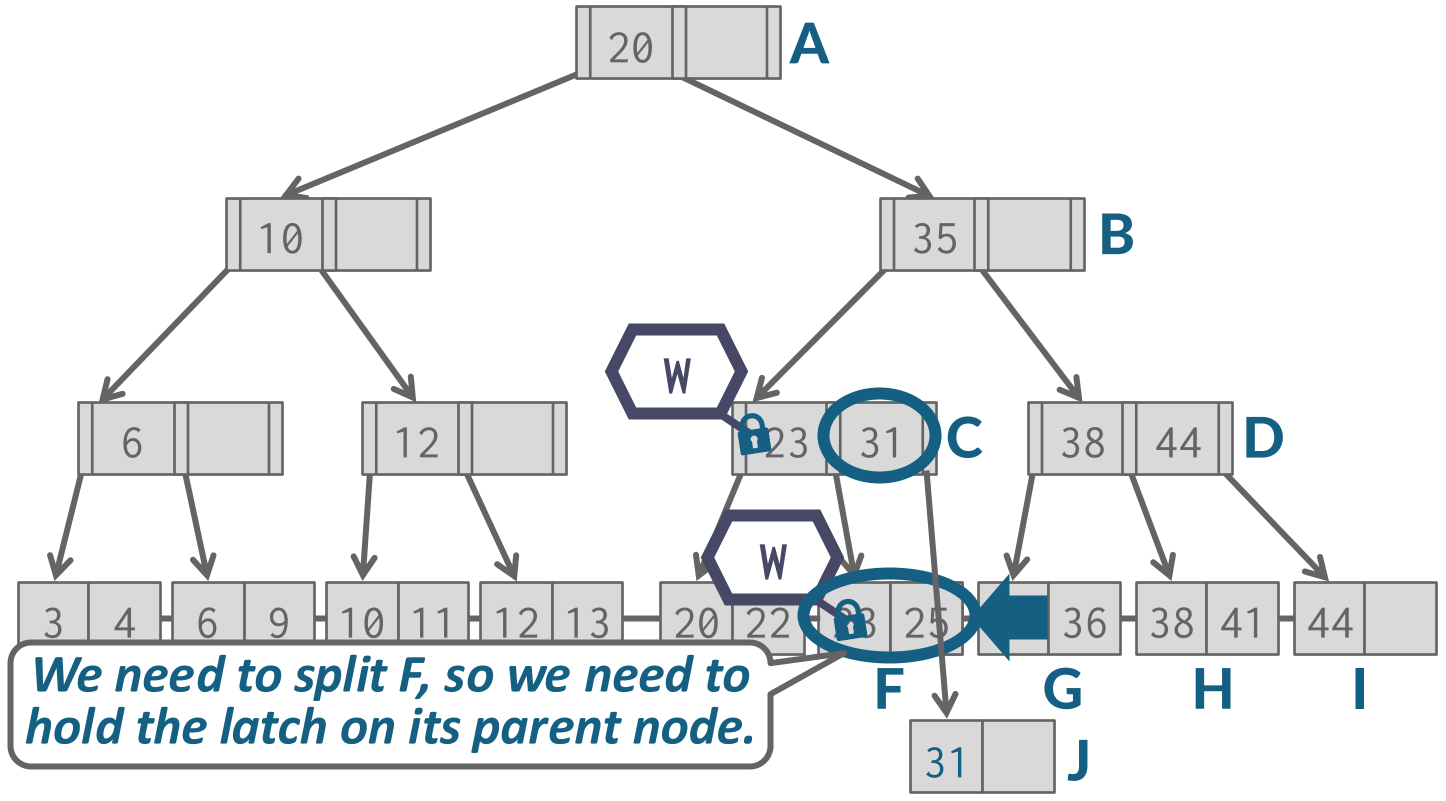
SCHOOL OF DATA SCIENCE

Example #1 – Insert 25



We need to split F, so we need to hold the latch on its parent node.

Example #1 – Insert 25



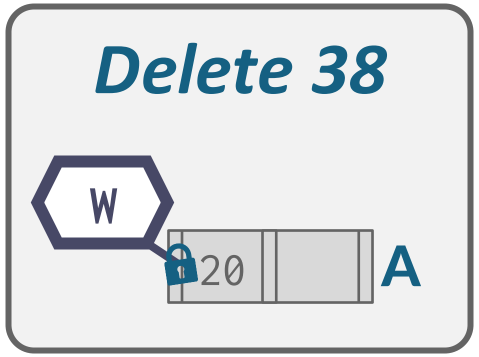
We need to split F, so we need to hold the latch on its parent node.

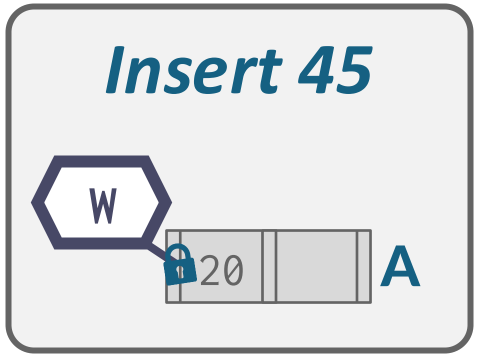
Observation

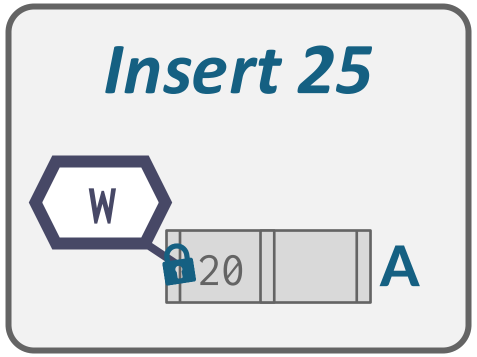
- What was the first step that all the update examples did on the B+Tree?

Observation

- What was the first step that all the update examples did on the B+Tree?

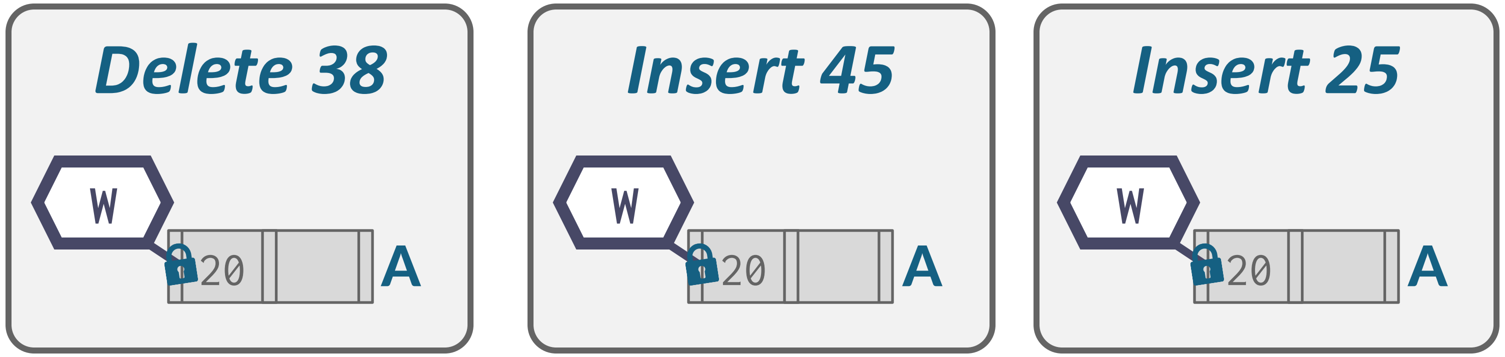






Observation

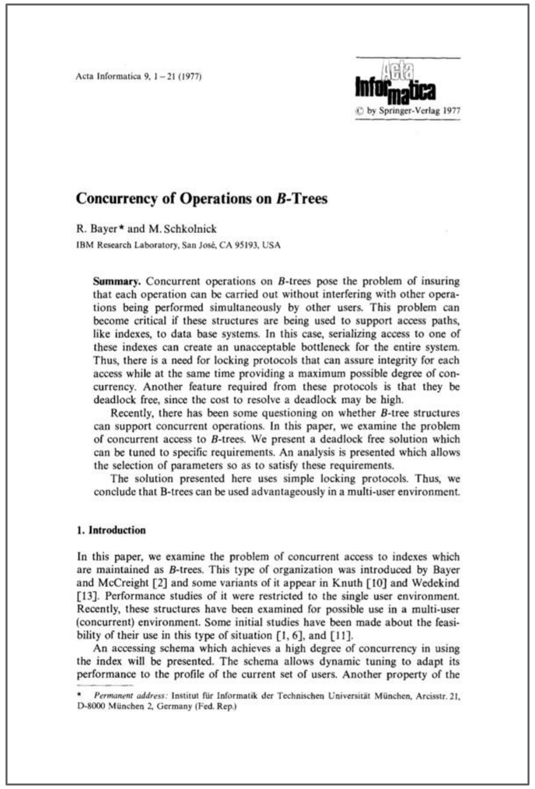
- What was the first step that all the update examples did on the B+Tree?



- 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.



Acta Informatica 9, 1–21 (1977)

Acta Informatica
© by Springer-Verlag 1977

Concurrency of Operations on *B*-Trees

R. Bayer* and M. Schkolnick
IBM Research Laboratory, San José, CA 95193, USA

Summary. Concurrent operations on *B*-trees pose the problem of insuring that each operation can be carried out without interfering with other operations being performed simultaneously by other users. This problem can become critical if these structures are being used to support access paths, like indexes, to data base systems. In this case, serializing access to one of these indexes can create an unacceptable bottleneck for the entire system. Thus, there is a need for locking protocols that can assure integrity for each access while at the same time providing a maximum possible degree of concurrency. Another feature required from these protocols is that they be deadlock free, since the cost to resolve a deadlock may be high.

Recently, there has been some questioning on whether *B*-tree structures can support concurrent operations. In this paper, we examine the problem of concurrent access to *B*-trees. We present a deadlock free solution which can be tuned to specific requirements. An analysis is presented which allows the selection of parameters so as to satisfy these requirements.

The solution presented here uses simple locking protocols. Thus, we conclude that B-trees can be used advantageously in a multi-user environment.

1. Introduction

In this paper, we examine the problem of concurrent access to indexes which are maintained as *B*-trees. This type of organization was introduced by Bayer and McCreight [2] and some variants of it appear in Knuth [10] and Wedekind [13]. Performance studies of it were restricted to the single user environment. Recently, these structures have been examined for possible use in a multi-user (concurrent) environment. Some initial studies have been made about the feasibility of their use in this type of situation [1, 6], and [11].

An accessing schema which achieves a high degree of concurrency in using the index will be presented. The schema allows dynamic tuning to adapt its performance to the profile of the current set of users. Another property of the

* *Permanent address:* Institut für Informatik der Technischen Universität München, Arcisstr. 21, D-8000 München 2, Germany (Fed. Rep.)

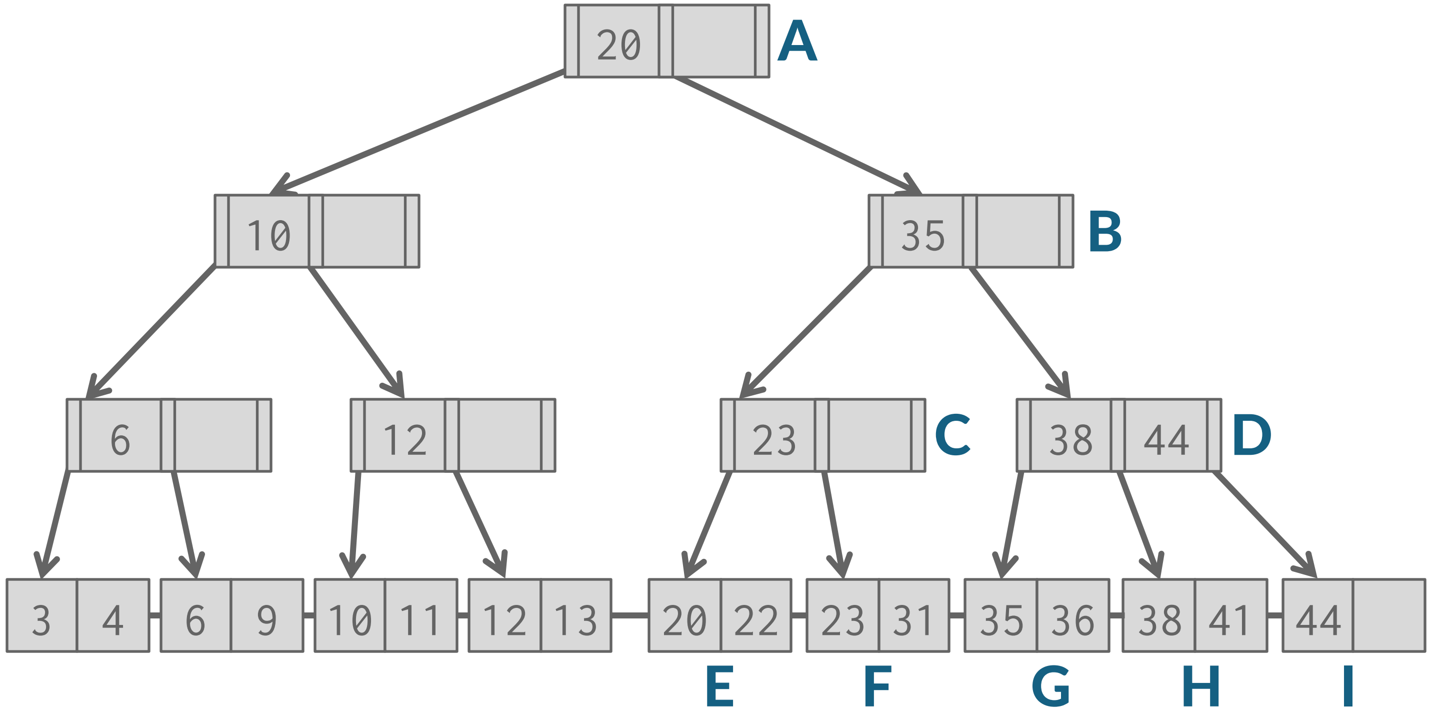
Better Latching Algorithm

- **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.

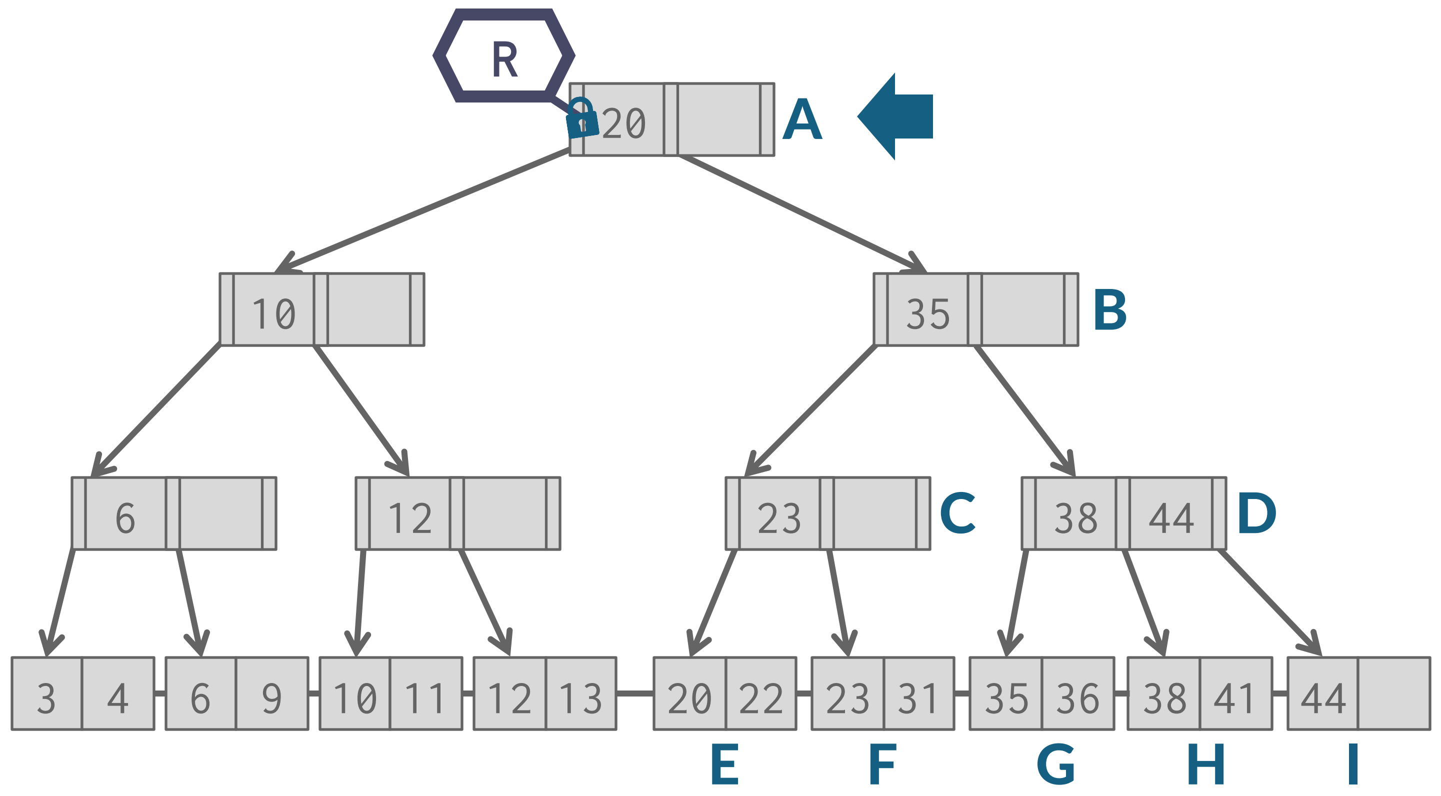
- This approach optimistically assumes that only leaf node will be modified; if not, `R` latches set on the first pass to leaf are wasteful.

SCHOOL OF
DATA SCIENCE

Example #2 – Delete 38

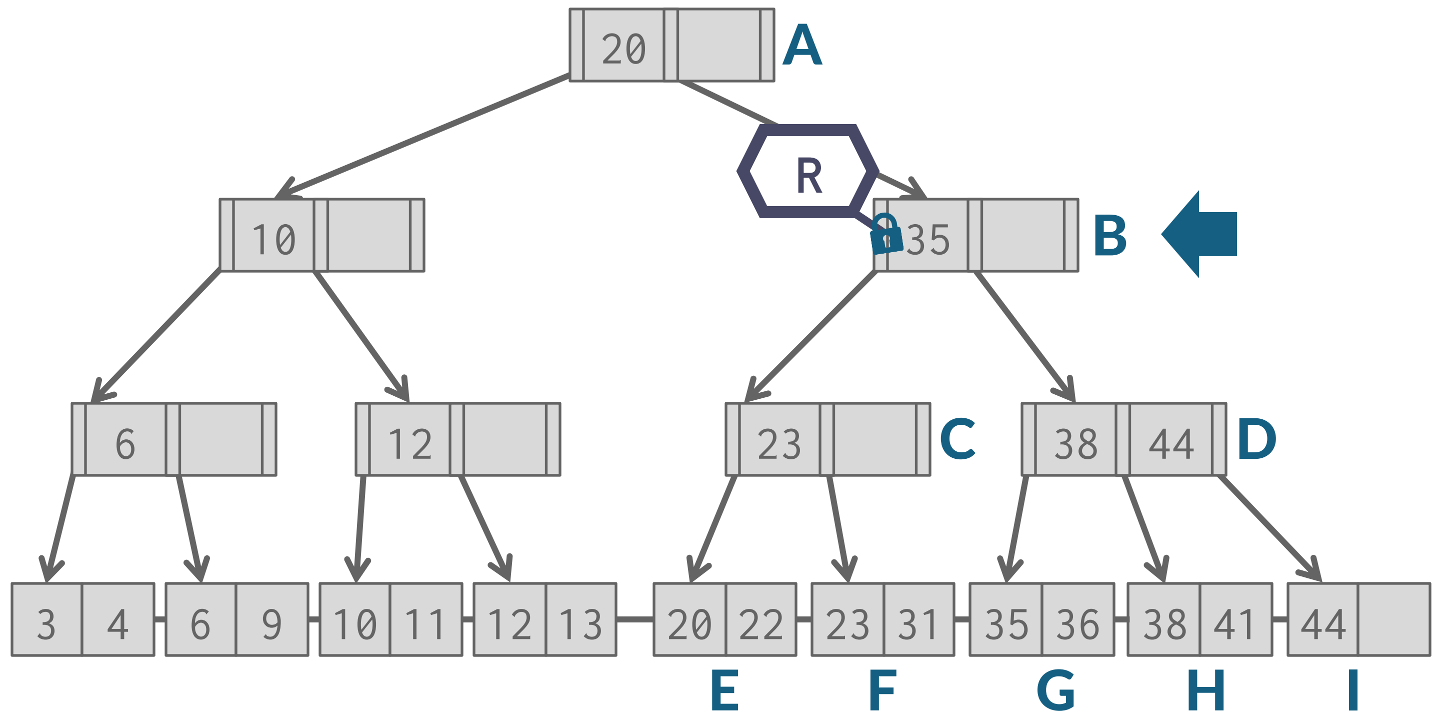


Example #2 – Delete 38

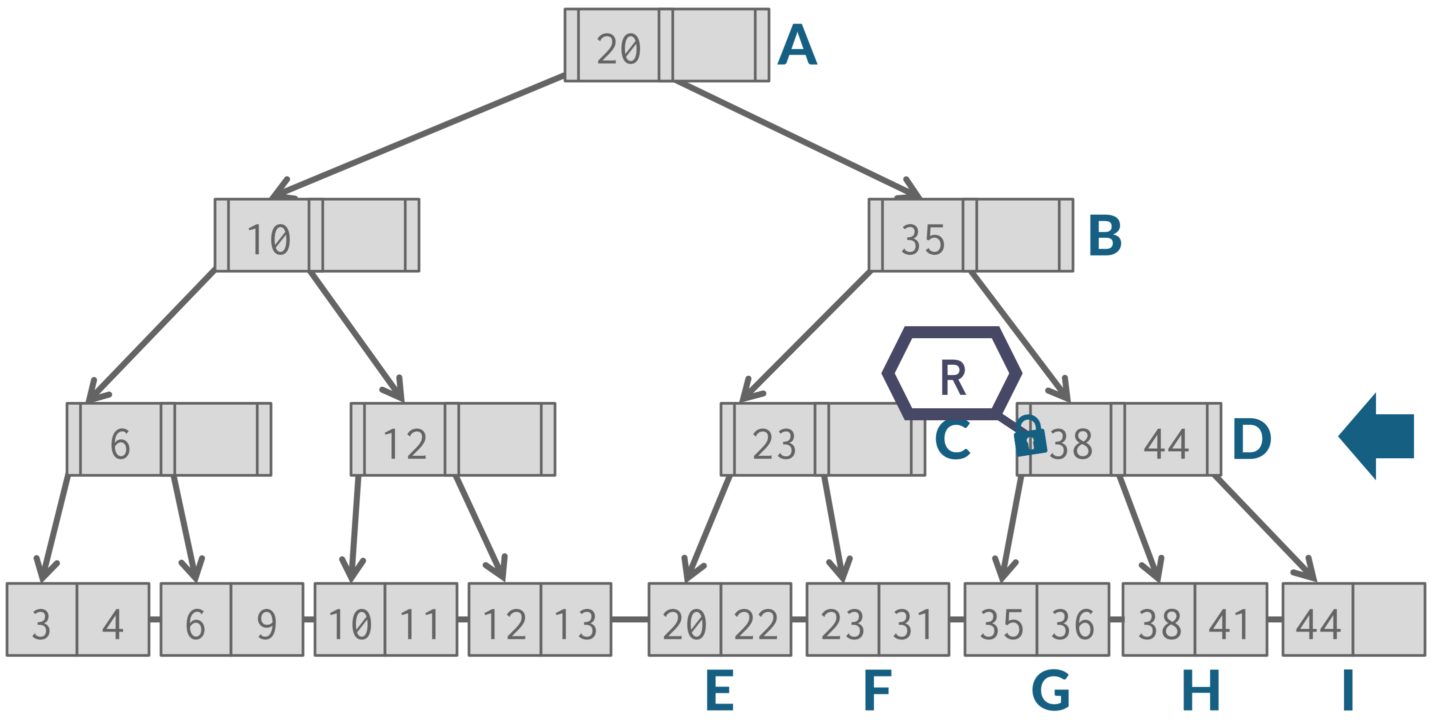


SCHOOL OF DATA SCIENCE

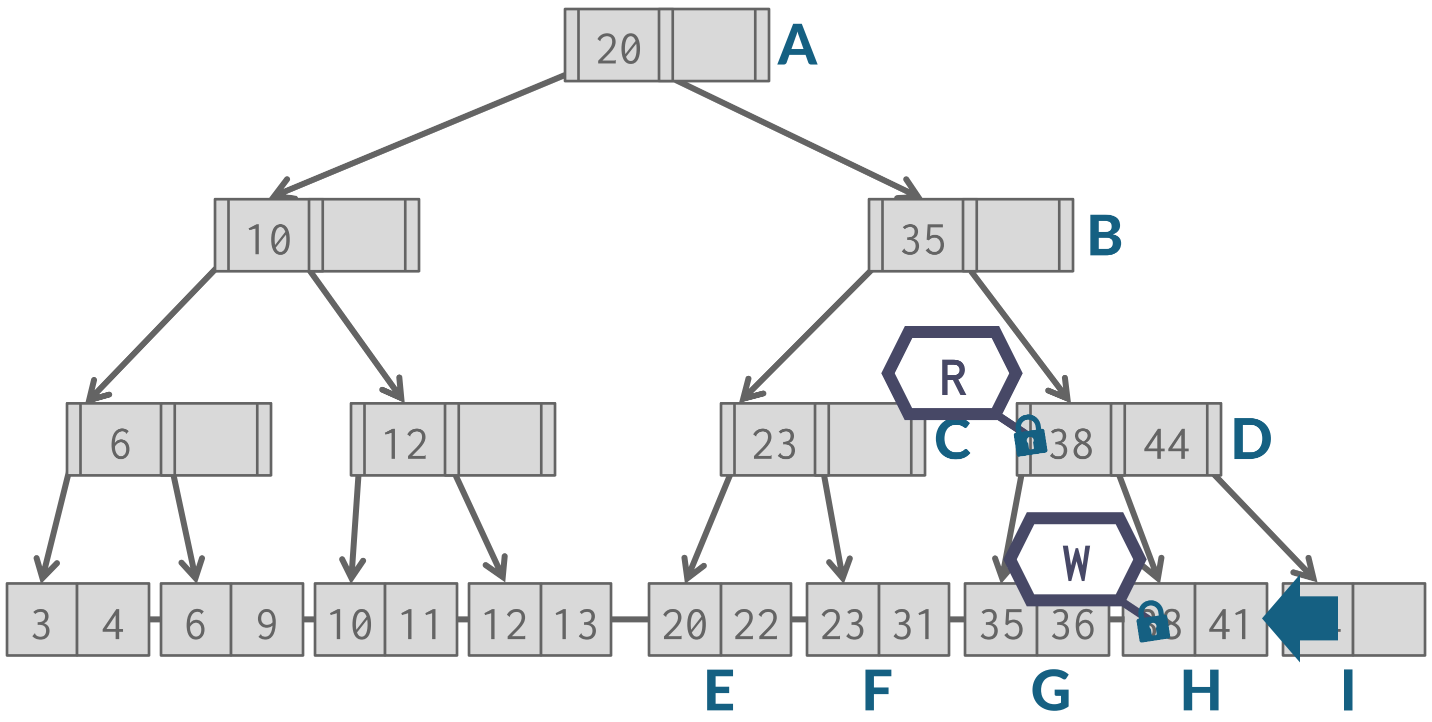
Example #2 – Delete 38



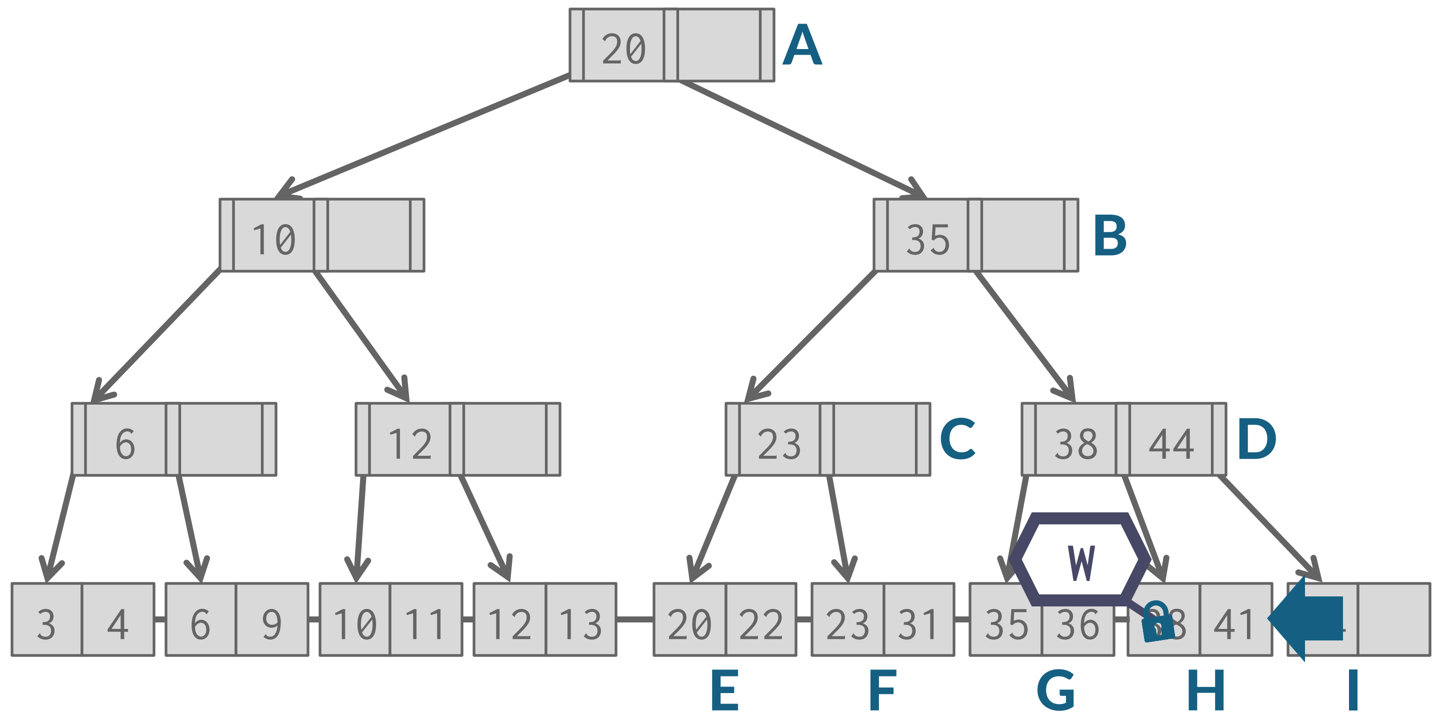
Example #2 – Delete 38



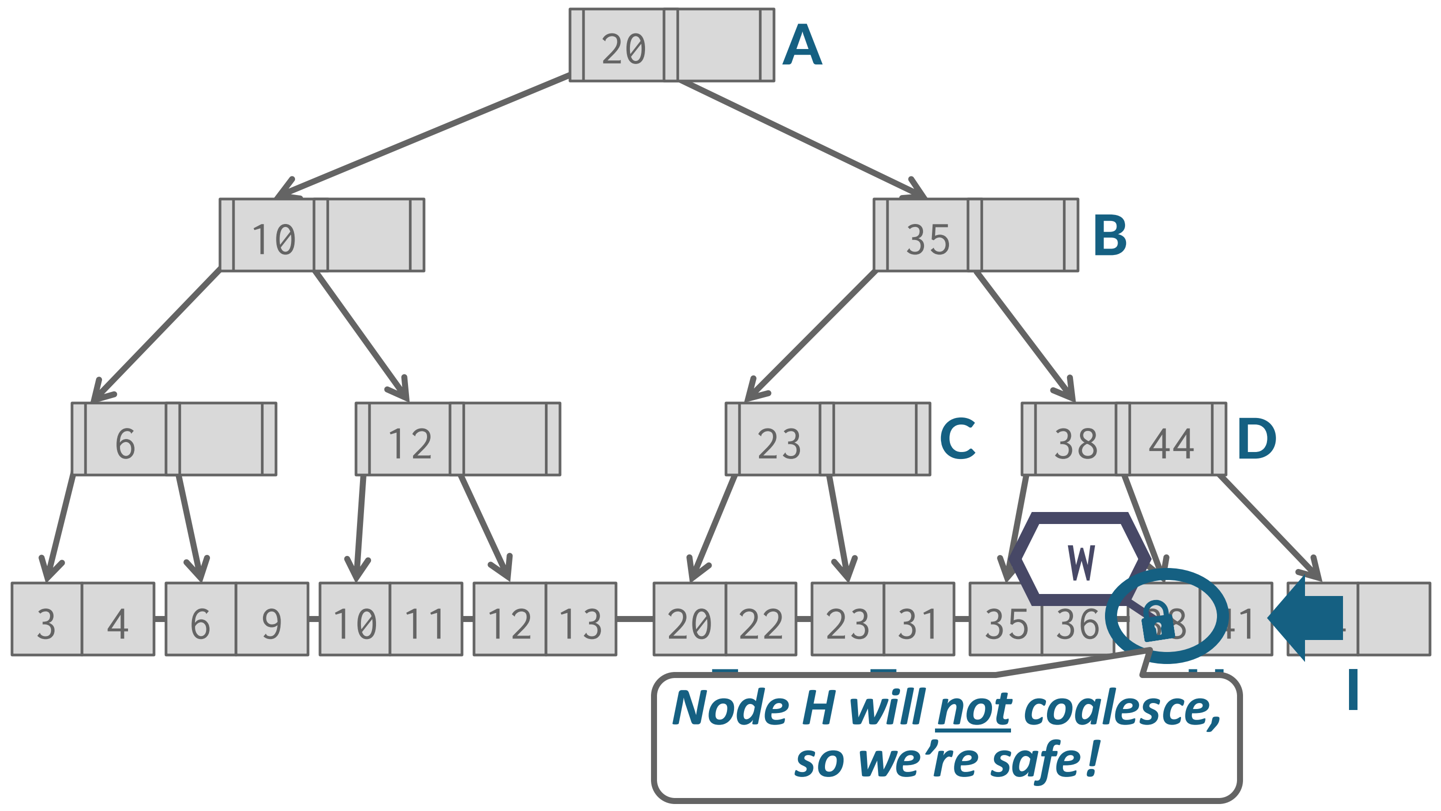
Example #2 – Delete 38



Example #2 – Delete 38

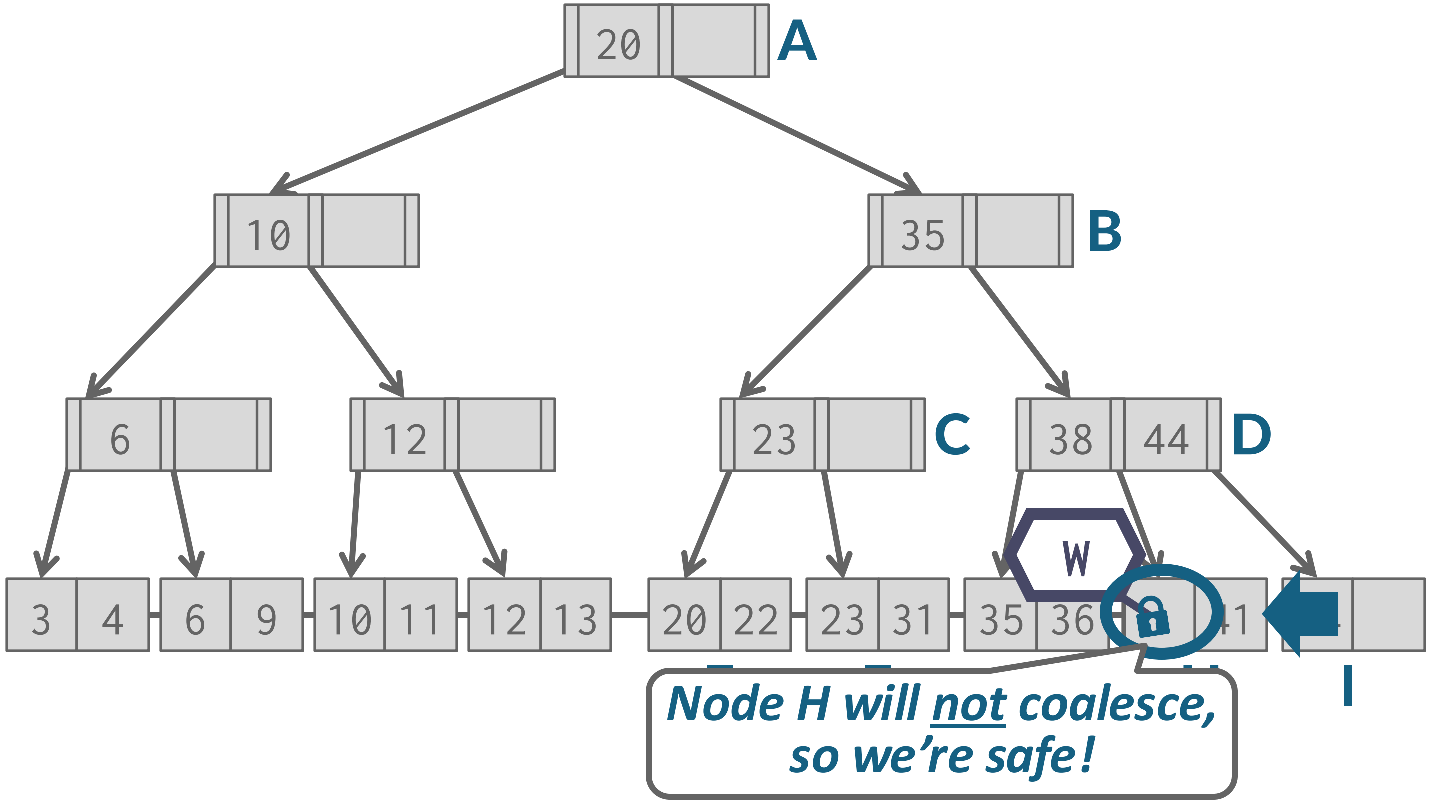


Example #2 – Delete 38

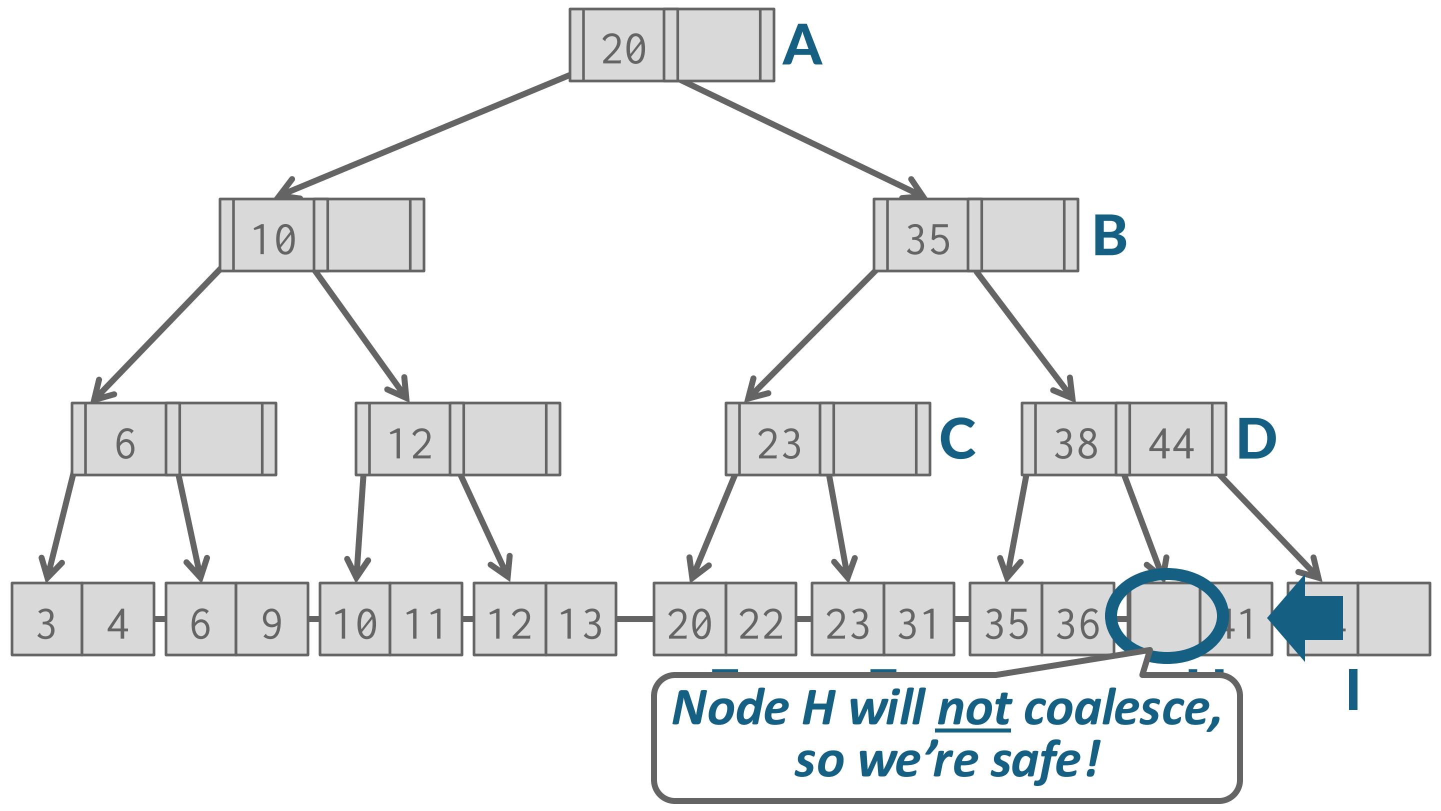


Node H will <u>not</u> coalesce, so we're safe!

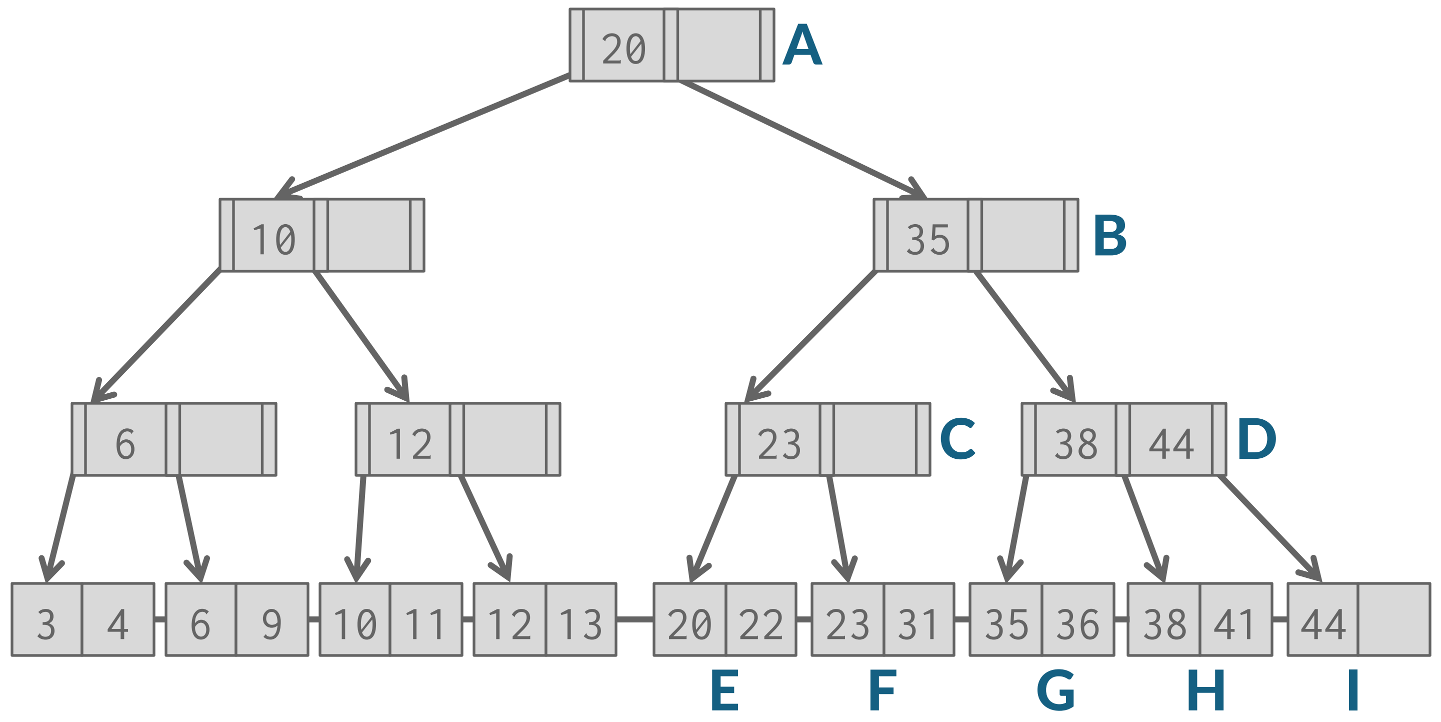
Example #2 – Delete 38



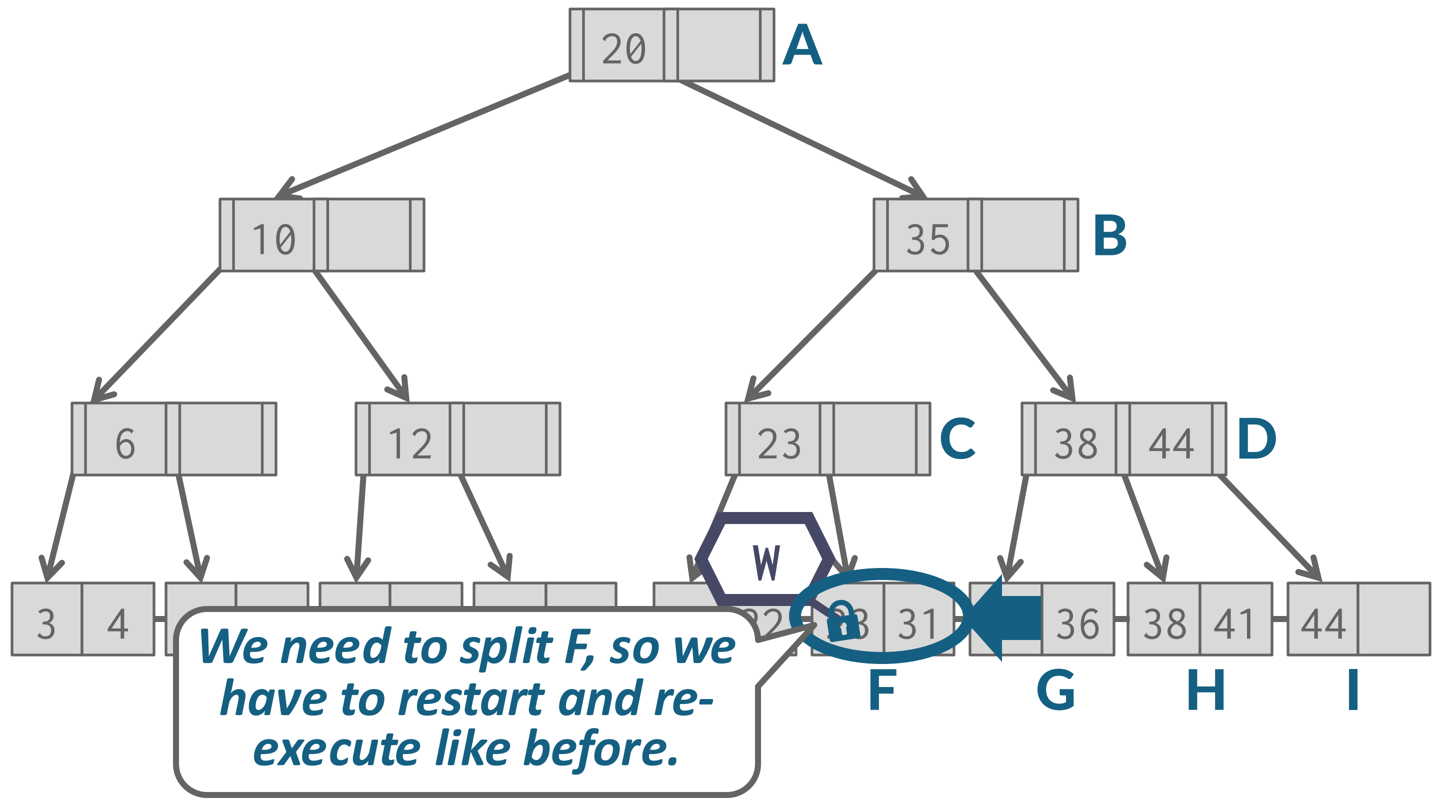
Example #2 – Delete 38



Example #2 – Insert 25



Example #2 – Insert 25

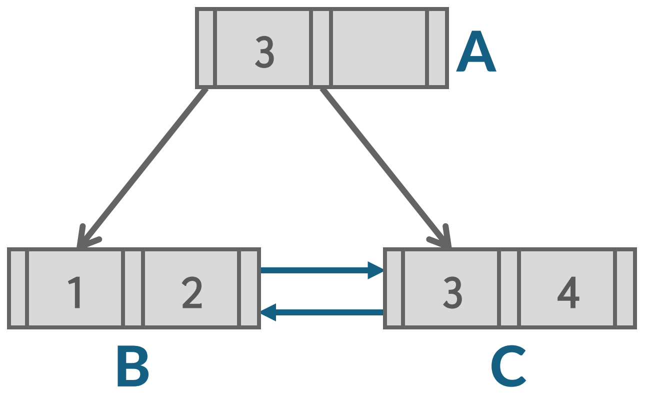

20
A
10
35
B
6
12
23
C
38
44
D
W
3
4
22
23
31
36
38
41
44
F
G
H
I
We need to split F, so we have to restart and re-execute like before.

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.

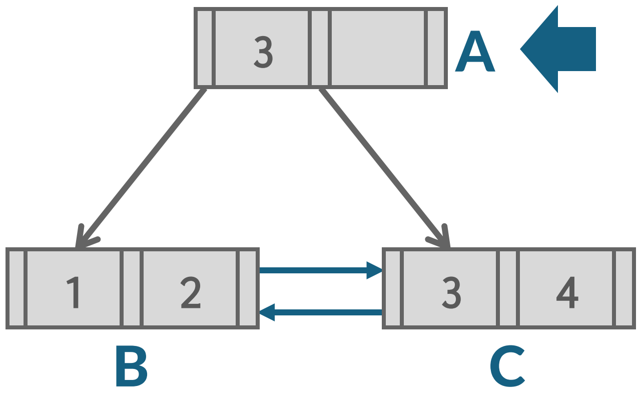
- But what if threads want to move from one leaf node to another leaf node?

Leaf Node Scan Example #1

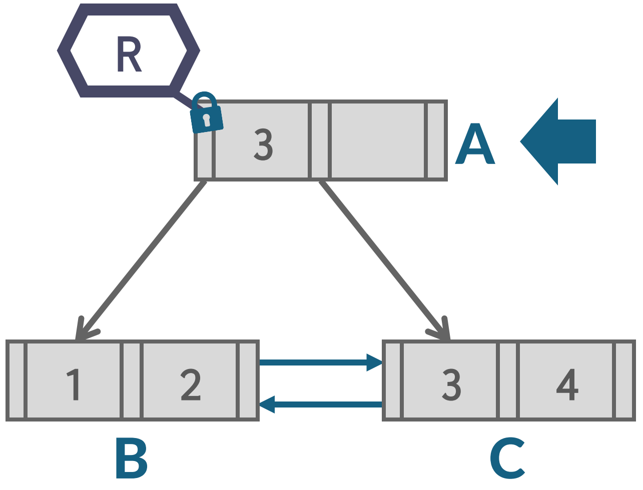


Leaf Node Scan Example #1

T_1: Find Keys < 4



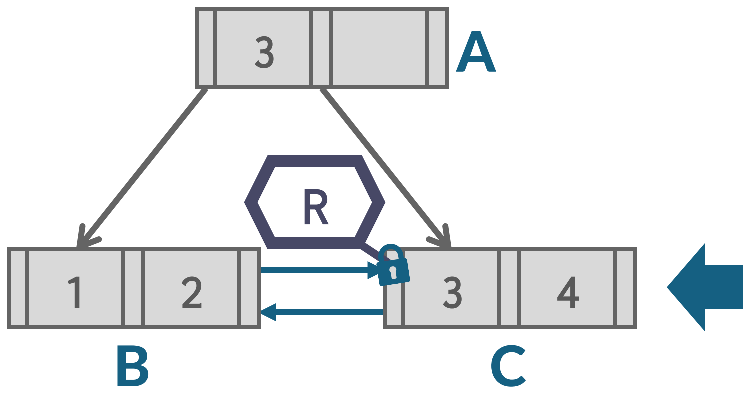
Leaf Node Scan Example #1



T_1: Find Keys < 4

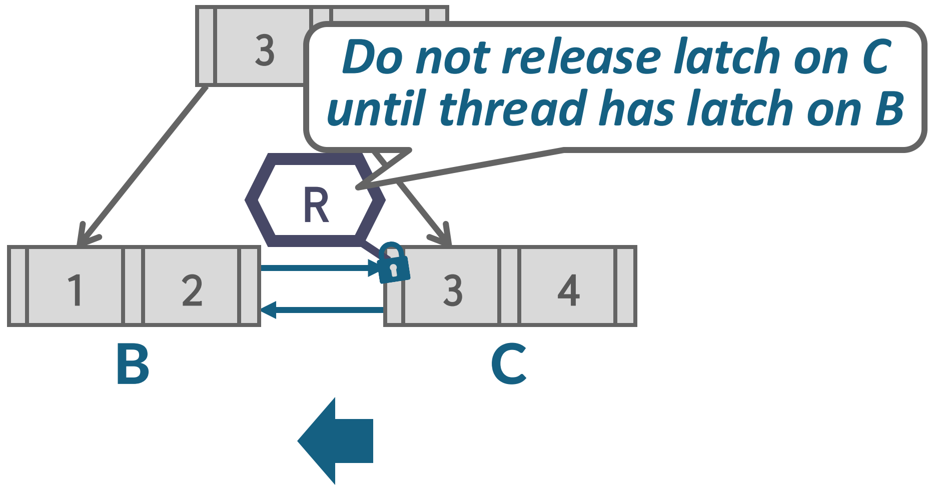
Leaf Node Scan Example #1

T$_1$: Find Keys < 4



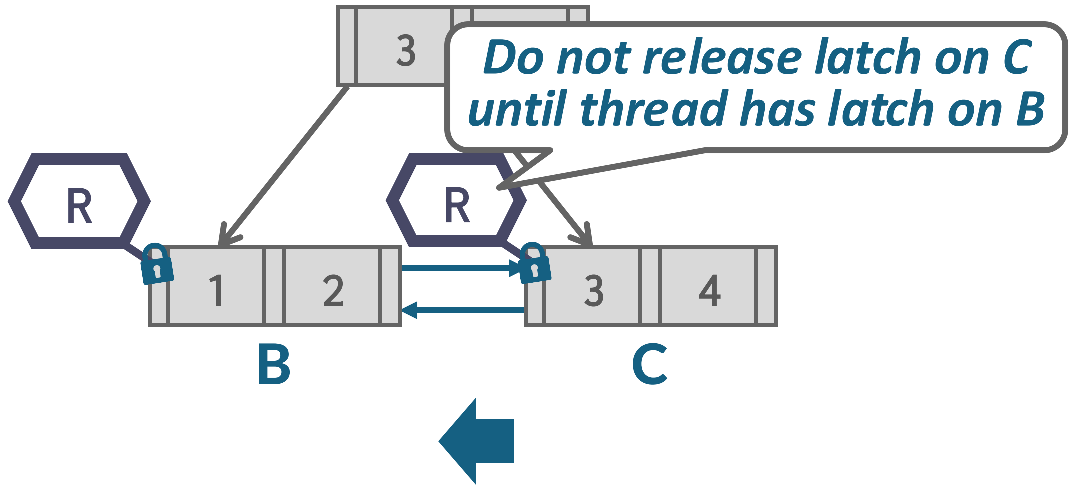
Leaf Node Scan Example #1

T_1: Find Keys < 4



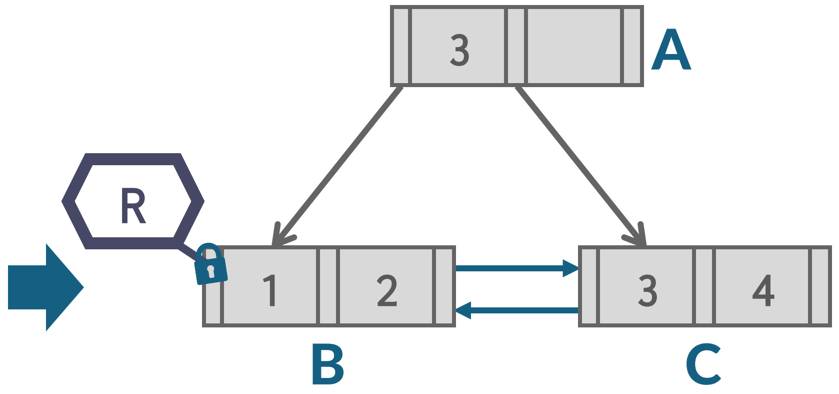
Leaf Node Scan Example #1

T_1: Find Keys < 4

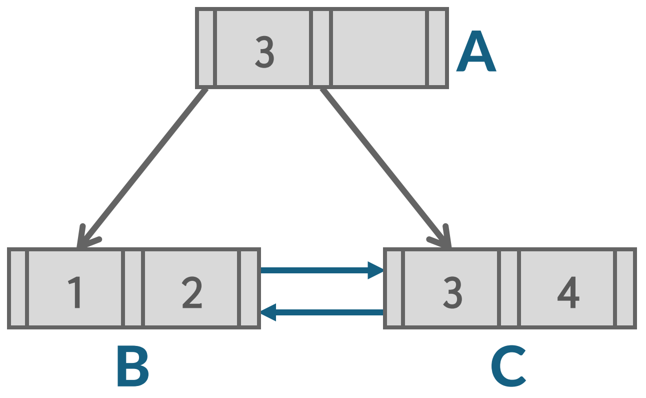


Leaf Node Scan Example #1

T_1: Find Keys < 4



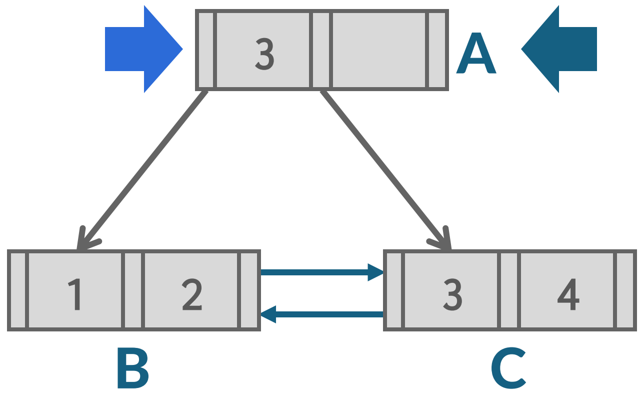
Leaf Node Scan Example #2



T_1: Find Keys < 4

T_2: Find Keys > 1

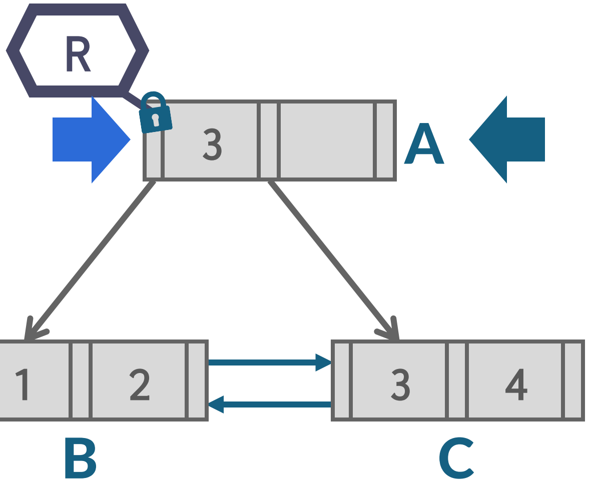
Leaf Node Scan Example #2



T_1: Find Keys < 4

T_2: Find Keys > 1

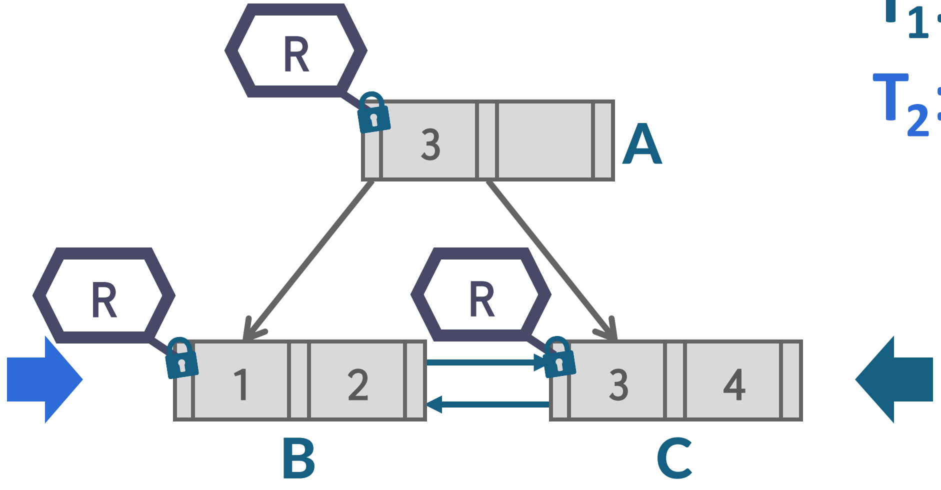
Leaf Node Scan Example #2



T_1: Find Keys < 4

T_2: Find Keys > 1

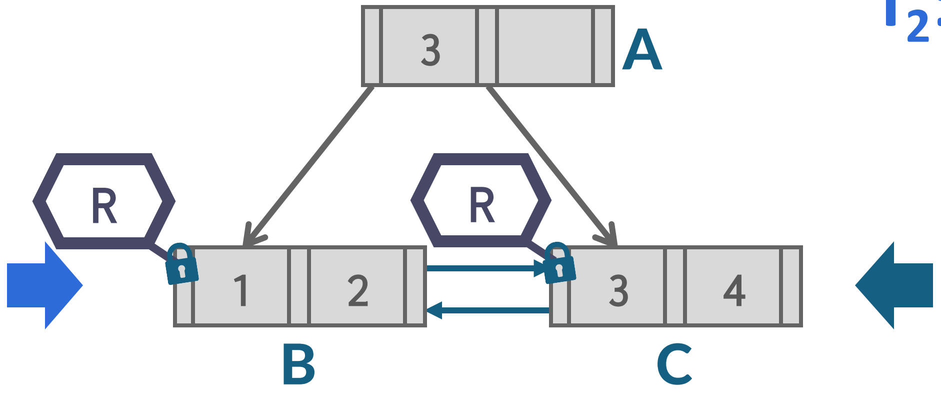
Leaf Node Scan Example #2



T_1: Find Keys < 4

T_2: Find Keys > 1

Leaf Node Scan Example #2

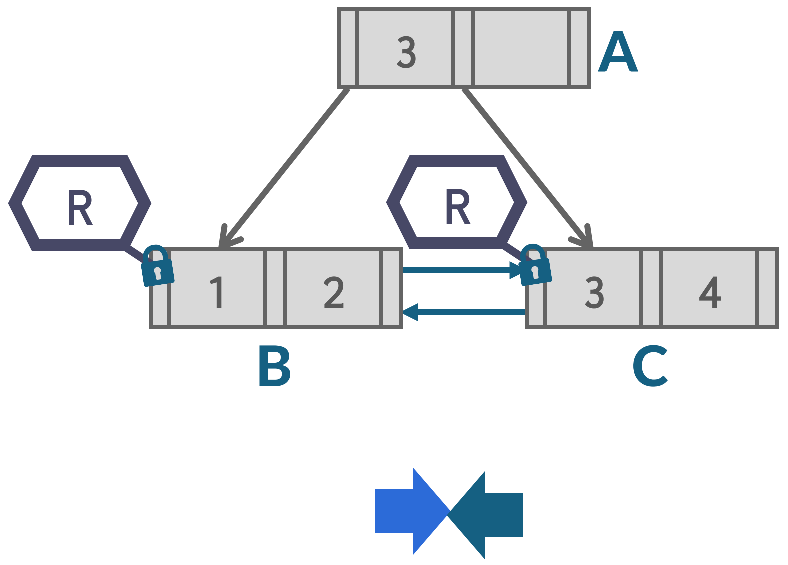


T_1: Find Keys < 4

T_2: Find Keys > 1

SCHOOL OF DATA SCIENCE

Leaf Node Scan Example #2



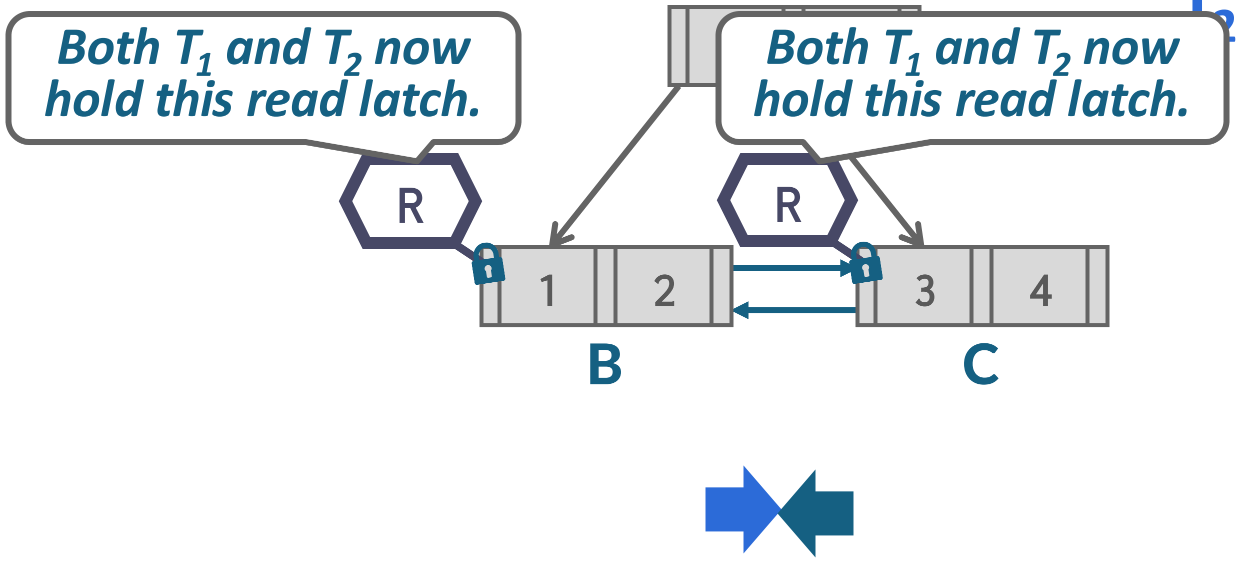
T_1: Find Keys < 4

T_2: Find Keys > 1

Leaf Node Scan Example #2

T_1: Find Keys < 4

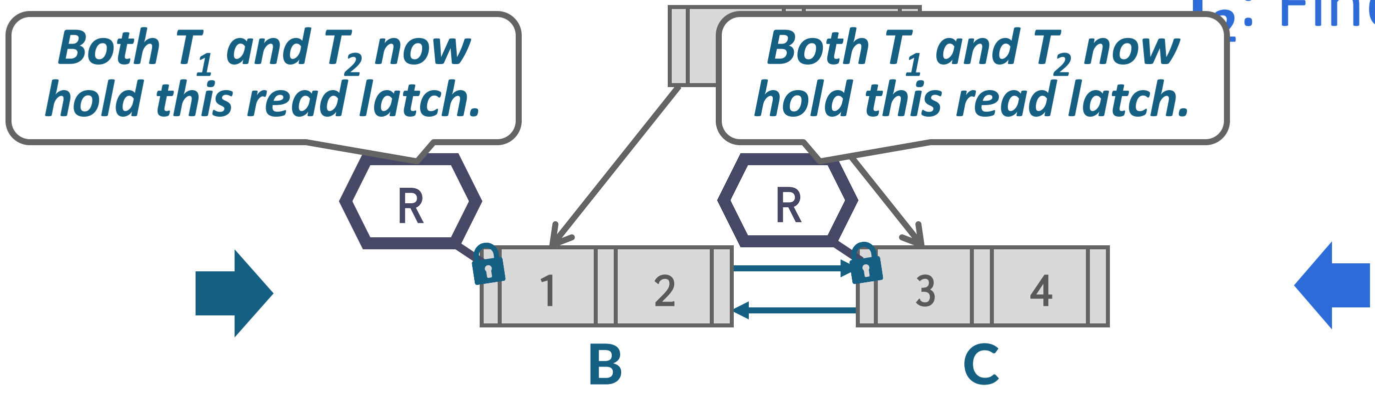
T_2: Find Keys > 1


Both T_1 and T_2 now hold this read latch.
R
1
2
B
Both T_1 and T_2 now hold this read latch.
R
3
4
C

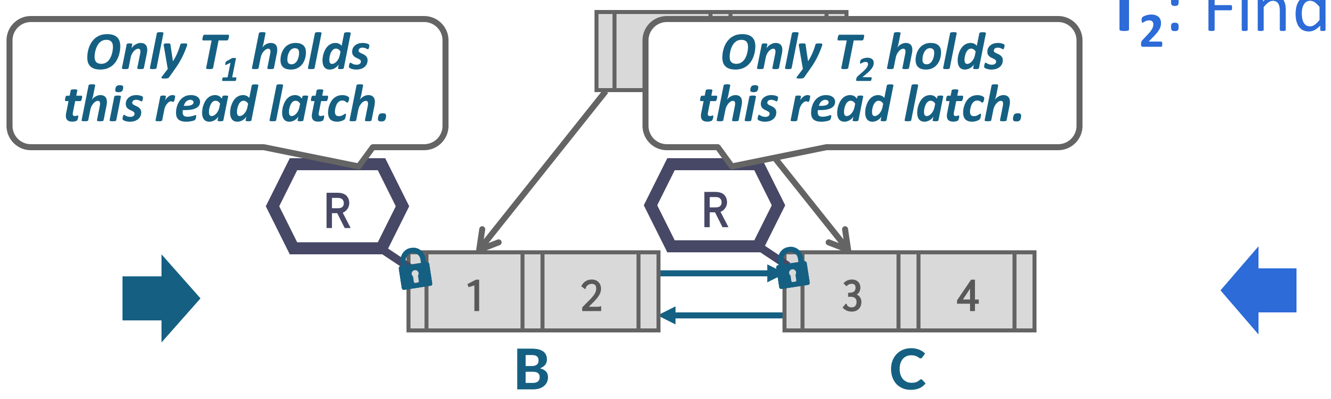
Leaf Node Scan Example #2

T_1: Find Keys < 4

T_2: Find Keys > 1


Both T_1 and T_2 now hold this read latch.
R
Both T_1 and T_2 now hold this read latch.
R
1
2
3
4
B
C

Leaf Node Scan Example #2

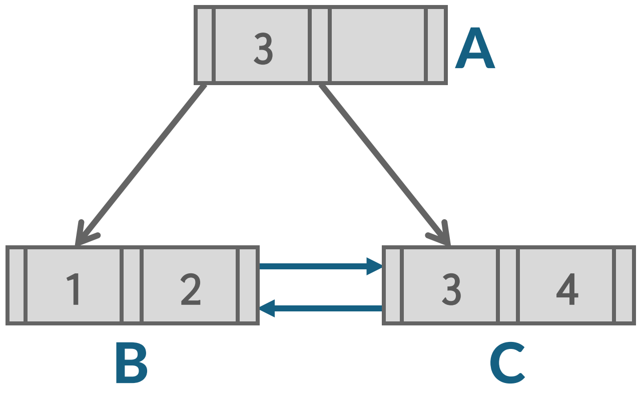

Only T_1 holds this read latch.
Only T_2 holds this read latch.
R
R
1
2
3
4
B
C

T_1: Find Keys < 4

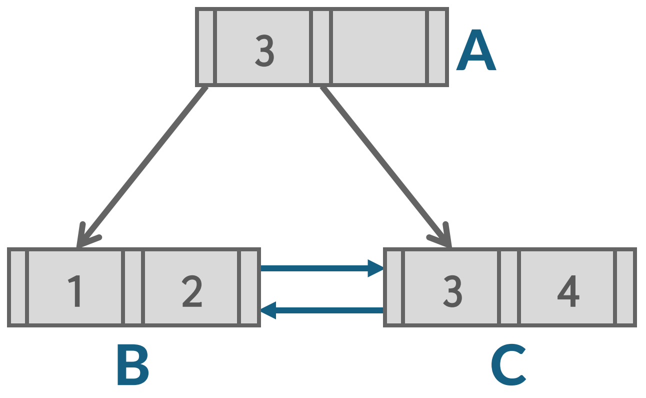
T_2: Find Keys > 1

SCHOOL OF
DATA SCIENCE

Leaf Node Scan Example #3



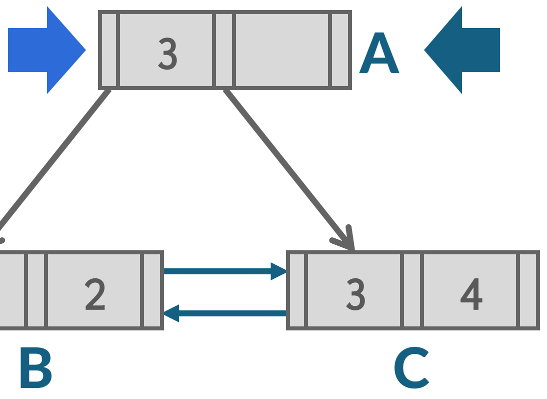
Leaf Node Scan Example #3



T_1: Delete 4

T_2: Find Keys > 1

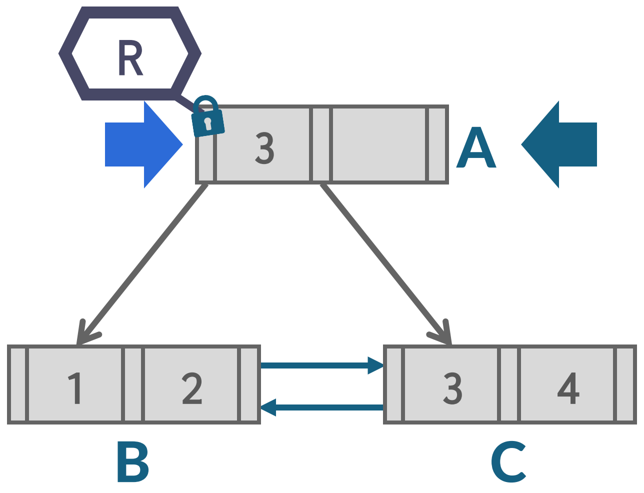
Leaf Node Scan Example #3



T_1: Delete 4

T_2: Find Keys > 1

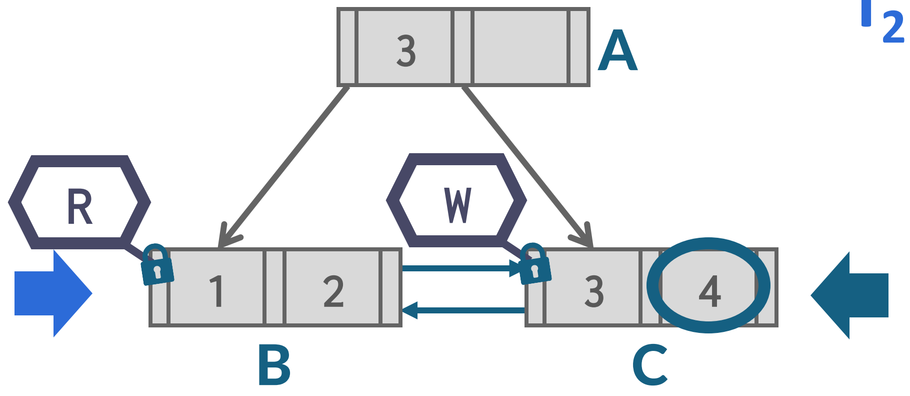
Leaf Node Scan Example #3



T_1: Delete 4

T_2: Find Keys > 1

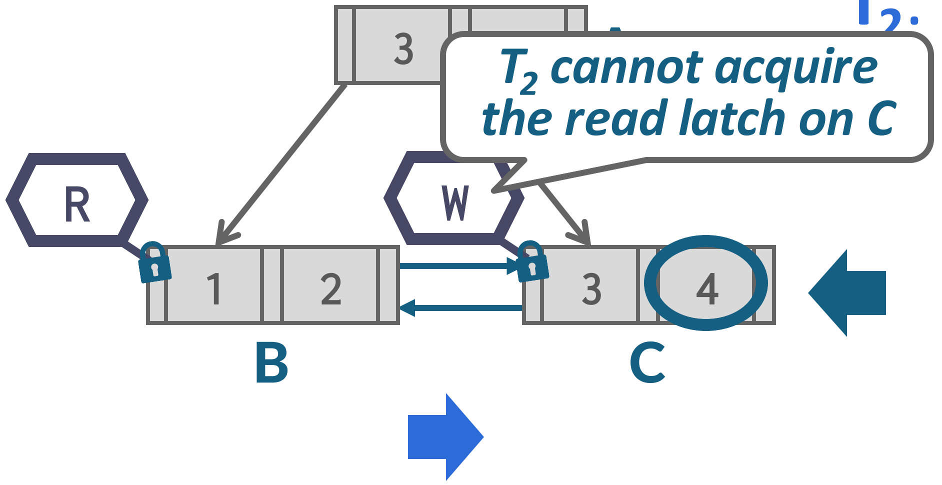
Leaf Node Scan Example #3



T$_1$: Delete 4

T$_2$: Find Keys > 1

Leaf Node Scan Example #3



T_1: Delete 4

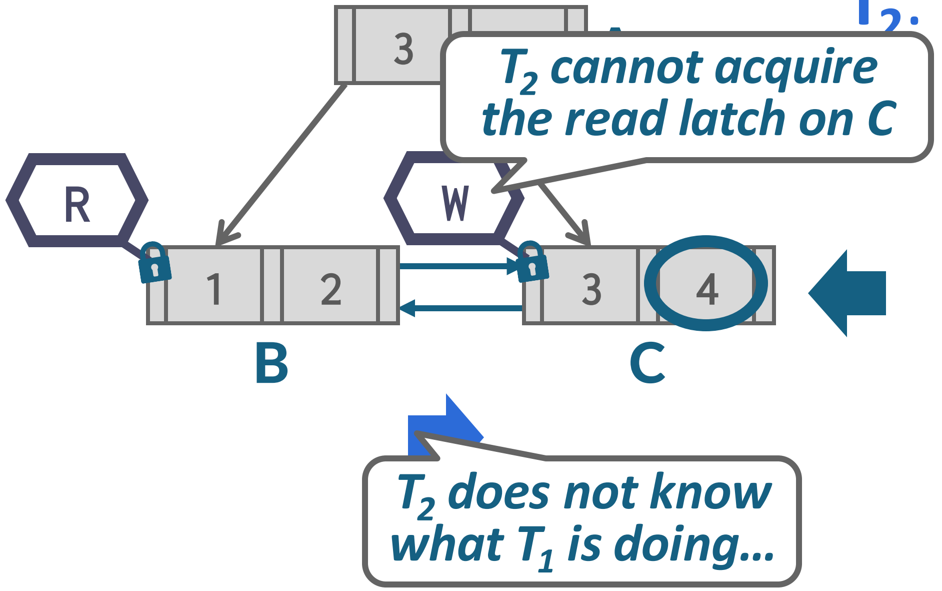
T_2: Find Keys > 1

SCHOOL OF DATA SCIENCE

Leaf Node Scan Example #3

T_1: Delete 4

T_2: Find Keys > 1

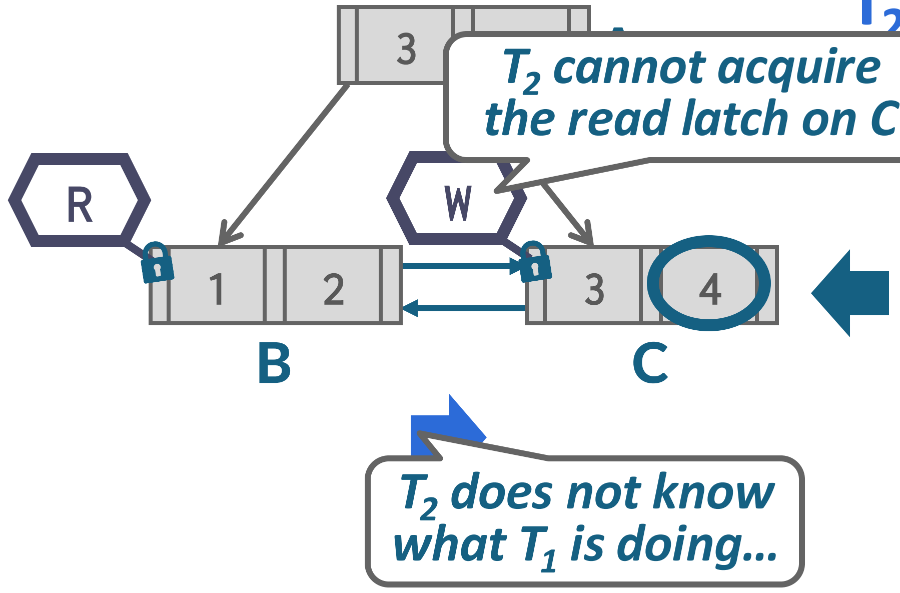

3
R
W
1
2
3
4
B
C
T_2 cannot acquire the read latch on C
T_2 does not know what T_1 is doing...

Leaf Node Scan Example #3

T_1: Delete 4

T_2: Find Keys > 1

T_2 Choices?



3
R
W
1
2
3
4
B
C
T_2 cannot acquire the read latch on C
T_2 does not know what T_1 is doing…

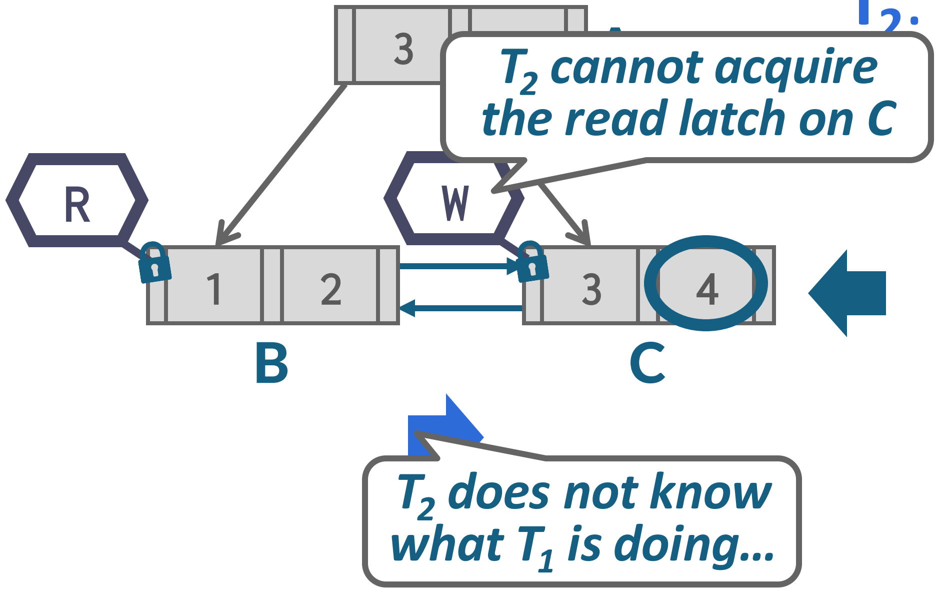
Leaf Node Scan Example #3

T_1: Delete 4

T_2: Find Keys > 1

T_2 Choices?

⌛ *Wait*


3
R
W
1
2
3
4
B
C
T_2 cannot acquire the read latch on C
T_2 does not know what T_1 is doing…

SCHOOL OF DATA SCIENCE

Leaf Node Scan Example #3

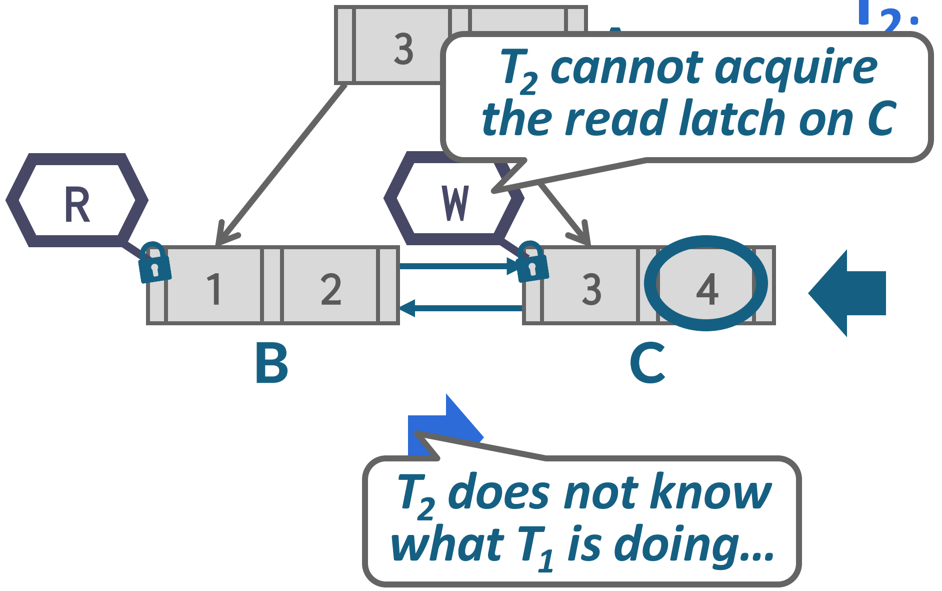
T$_1$: Delete 4

T$_2$: Find Keys > 1

T$_2$ Choices?

Wait

Kill Ourself

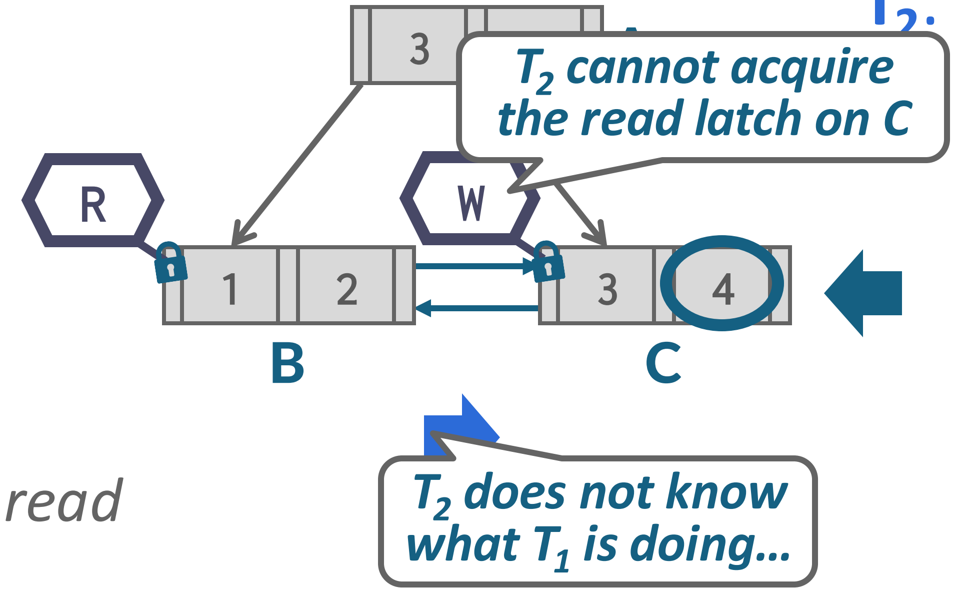


3
R
W
1
2
3
4
B
C
T$_2$ cannot acquire the read latch on C
T$_2$ does not know what T$_1$ is doing…

Leaf Node Scan Example #3

T_1: Delete 4

T_2: Find Keys > 1



3
R
W
1
2
3
4
B
C
T_2 cannot acquire the read latch on C
T_2 does not know what T_1 is doing…

T_2 Choices?

- *Wait*
- *Kill Ourself*
- *Kill Other Thread*

Leaf Node Scan Example #3

T_1: Delete 4

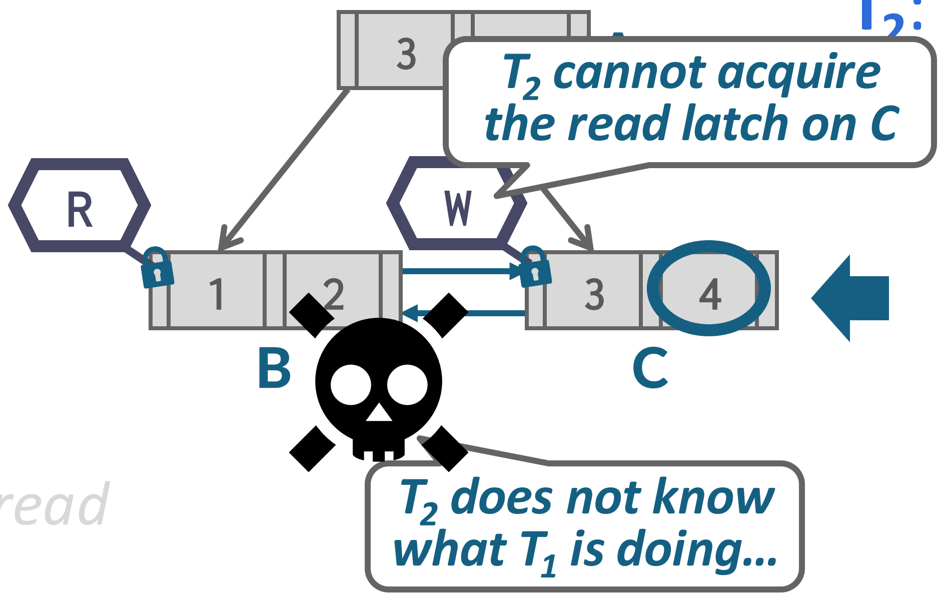
T_2: Find Keys > 1

T_2 Choices?

Wait

☠ *Kill Ourself*

Kill Other Thread


3
R
W
1
2
3
4
B
C
T_2 cannot acquire the read latch on C
T_2 does not know what T_1 is doing...

Leaf Node Scans

- 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.

SCHOOL OF
DATA SCIENCE

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 Lecture

- We are finally going to discuss how to execute some queries…