Concurrent Updates with RCU: Search Tree as an Example*  

Maya Arbel  
Department of Computer Science  
Technion  
mayaarl@cs.technion.ac.il  

Hagit Attiya  
Department of Computer Science  
Technion  
hagit@cs.technion.ac.il  

ABSTRACT  
Read copy update (RCU) is a novel synchronization mechanism, in which the burden of synchronization falls completely on the updaters, by having them wait for all pre-existing readers to finish their read-side critical section. This paper presents CITRUS, a concurrent binary search tree (BST) with a wait-free contains operation, using RCU synchronization and fine-grained locking for synchronization among updaters. This is the first RCU-based data structure that allows concurrent updaters. While there are methodologies for using RCU to coordinate between readers and updaters, they do not address the issue of coordination among updaters, and indeed, all existing RCU-based data structures rely on coarse-grained synchronization between updaters.

Experimental evaluation shows that CITRUS beats previous RCU-based search trees, even under mild update contention, and contains a concurrent binary search tree (BST) with a wait-free contains operation, using RCU synchronization and fine-grained locking for coordination among updaters. This is the first RCU-based data structure that allows concurrent updaters. While there are methodologies for using RCU to coordinate between readers and updaters, they do not address the issue of coordination among updaters, and indeed, all existing RCU-based data structures rely on coarse-grained synchronization between updaters.

This observation led to the design of CITRUS, a binary search tree implementing a dictionary with concurrent updates (read/modify/write) and contains queries. Updates synchronize using fine-

Categories and Subject Descriptors  
D.1.3 [Programming Techniques]: Concurrent Programming—Concurrent programming; F.1.2 [Computation By Abstract Devices]: Modes of Computation—Parallelism and concurrency  

General Terms  
Algorithms, Architecture  

Keywords  
Shared memory; internal search tree; read-copy-update  

1. INTRODUCTION

The traditional approach to synchronization between readers and writers allows concurrency among readers, but excludes readers when a writer is executing, e.g., through a readers/writer lock. An alternative, contemporary approach is presented by read copy update (RCU) [22], a distinctive synchronization mechanism favoring read-only operations even further, by allowing them to proceed even while writers are modifying the data they are reading. Instead, the read-side critical section is wrapped by the \texttt{rcu\_read\_lock} and \texttt{rcu\_read\_unlock} functions; an additional synchronize\_rcu function can be used by a writer as a barrier ensuring that all preceding read-side critical sections complete. In typical RCU usage, the burden of synchronization is placed on updates, who can wait for all pre-existing readers to finish their read-side critical section.

RCU is used extensively within the Linux kernel [20], mostly to facilitate memory reclamation [12, 21]. However, its potential for concurrent programming remained unexploited. Although several RCU-based data structures have been proposed, for example, search trees and hash tables [6, 18, 25, 26], they all do not allow concurrent updates, either pessimistically, using a coarse-grained lock [6], or optimistically, using transactional memory [18]. At best, the data structure is partitioned into segments, e.g., buckets in a hash table [25, 26], each guarded by a single lock.

This leaves unanswered the question of coordination between concurrent updates to the data structure, while using RCU, which is the topic of this paper. A natural approach for supporting concurrent updates is to employ fine-grained locks, acquired and released according to some locking policy, e.g., two-phase locking or hand-over-hand locking. We found that these fine-grained approaches fail to ensure consistent views for read-only operations that access several items in the data structure, e.g., partial snapshots [2] or other iterators [24]. Figure 1 shows an example in which two readers, \(r_1\) and \(r_2\), attempt to collect the leaves, by in-order traversal, while two updates delete leaves 9 and 12. In (a), \(r_2\) already traversed the left sub-tree while \(r_1\) has not yet started its traversal. After 9 is deleted (b), \(r_1\) finishes traversing the tree while \(r_2\) did not take additional steps, and 12 is deleted before both readers complete (c). Since each reader may observe a different permutation of the writes to the data structure, the values returned by \(r_1\) and \(r_2\) are such that they observed the updates in different order. This means that without additional iterations or interaction, there is no consistent way to order the updates and the readers, complicating the task of designing a concurrent data structure.

Nevertheless, this example (and others that we found) hints that the difficulty is in having read-only operations that need to atomically access several locations. The counter-examples do not hold when the read-only operation only searches for a particular data item, e.g., in a dictionary.

This observation led to the design of CITRUS, a binary search tree implementing a dictionary with concurrent updates (\texttt{insert}, \texttt{delete}) and contains queries. Updates synchronize using fine-
a wait-free manner, in parallel with updates, relying on RCU to ensure correctness. The combination of RCU synchronization with fine-grained locking of modified nodes yields a simple design, greatly resembling the sequential algorithm, and leading to a (relatively) simple proof of correctness. The operation or at the beginning of an update, is done in a wait-free manner, as in the sequential data structure, but inside an RCU read-side critical section.

Updates start by searching for their key, in a manner similar to contains. An unsuccessful update (insert that finds its key or delete that does not find its key), returns immediately. Otherwise, the update is located where the change should be done.

A new node is inserted as a leaf, requiring little synchronization, but delete may need to remove an internal node. When the node has two children, this is done by replacing the node with its successor in the tree. This scenario requires coordination with concurrent searches that could miss the successor in both its previous location and in its new location, necessitating delicate synchronization in some concurrent search trees [4, 7, 9, 19]. CITRUS easily circumvents this pitfall by copying the successor to the new location, and using the RCU synchronization mechanism to wait for on-going searches, before removing the successor from its previous location.

Experimental evaluation of CITRUS shows that its performance beats previous RCU-based search trees [6, 18], even under mild update contention. It also compares well with the best-available concurrent dictionaries, e.g., [4, 15, 23].

The evaluation also reveals that the user-level RCU implementation [8] is ill-suited for workloads in which many updates concurrently synchronize through it. We sketch a new implementation that avoids these pitfalls and scales better with growing number of updates. In the new implementation, a thread indicates that it starts a read-side critical section by incrementing a counter and setting a flag to true; the flag is set to false at the end of the section. To synchronize, an update waits until every other thread either increases its counter or sets its flag to false.

2. PRELIMINARIES

A concurrent system consists of a set of threads, communicating by applying primitives to shared variables. A shared data structure implementation provides a set of operations, each invoked with possible parameters and returning with a response. The invocation of an operation is a local step of a thread, leading to the execution of an algorithm, directing the thread to execute a sequence of atomic steps. Each atomic step is either an instance of a shared memory primitive or computation on variables that are local to the thread. Returning from an operation is a local computation step.

A configuration is an instantaneous snapshot of the system describing the state of all local and shared variables. In the initial configuration, all variables hold an initial value.

An execution π is an alternating sequence of configurations and steps, C0, s1, ..., si, Ci, ..., where C0 is the initial configuration, and each configuration Ci is the result of executing the step si in configuration C(i−1). A prefix σ of π is a sub-sequence of π starting in C0 and ending with a configuration. An interval of π is a sub-sequence that starts with a step and ends with a configuration. The interval of an operation op starts with the invocation step of op and ends with the configuration following the response of op or the end of π, if there is no such response.

The API for read copy update (RCU) provides several functions, three of which are used in CITRUS to synchronize between readers and updates: rcu_read_lock, rcu_read_unlock and synchronize_rcu. A read-side critical section is the interval starting with the step returning from rcu_read_lock and ending with the configuration after the invocation of rcu_read_unlock. The implementations of rcu_read_lock and rcu_read_unlock must be wait-free. The RCU property (Figure 2) ensures that if a step of a read-side critical section precedes the invocation of synchronize_rcu, then all steps of this critical section precede the return from synchronize_rcu [12, 13].

A dictionary is a set of key-value pairs, with totally ordered keys, with the following operations:

insert(k,val) adds (k, val) to the set; returns true if k is not in the set and false otherwise.

delete(k) removes (k, val) from the set; returns true if k is in the set and false otherwise.

contains(k) returns val if (k, val) is in the set; otherwise, it returns false.

An update is either an insert or a delete.

A binary search tree implements a dictionary; it is internal if key-value pairs are stored in all nodes. A node v of the tree stores a key, Key(v), which never changes. Two dummy key values −1, ∞ are used to avoid corner cases, when the tree has fewer than two nodes; for every key k, −1 < k < ∞. The root of the tree always points to a node with key −1, this node has a right child with key ∞; all other nodes are in the left sub-tree of ∞. In a binary search tree (BST), all descendants in the left sub-tree of v have keys smaller than the key of v, and all descendants in the right sub-tree of v have keys larger than the key of v.

The successor of node v is the node u with the smallest key among the nodes with keys larger than or equal to Key(v). In a BST, u is the leftmost node in the right sub-tree of v.

3. THE CITRUS TREE ALGORITHM

Overview. A primary goal of CITRUS is to avoid locking when searching for a node, either in contains or at the beginning of
an update. This is implemented in an auxiliary procedure `get`, which starts at the root and searches down the tree in a manner similar to the sequential algorithm, except that it is performed inside a read-side critical section, wrapped with `rcu_read_lock` and `rcu_read_unlock`.

A `contains` simply invokes `get` to find the key; if the key is found, it returns the value stored in the node returned by `get`; otherwise, it returns `false`.

An `insert` invokes `get`, and returns `false` if `get` finds the key. Otherwise, a new node with the key is inserted as a leaf, added to the tree as the child of the last node in `get`’s search.

A `delete` invokes `get`, and returns `false` if `get` does not find the key. If the key is found in node `v`, then there are two cases. If `v` has at most one child, the node is removed by redirecting the child field of `v`’s parent to point to `v`’s child (see Figure 3(a) and (b)); later in the proof this is called `bypassing`. Otherwise, `v` has two children and it is replaced with its successor in the tree, which is stored in the leftmost node in the right sub-tree of `v`.

Moving the successor, when `v` has two children, requires coordination with a concurrent `get` searching for the successor, which may return a false negative response (Figure 4). To overcome this problem, `delete` first inserts a copy of the successor node in the deleted key’s location. Then the `delete` waits for concurrent searches by executing `synchronize_rcu`, and only then, removes the old successor node. Searches that start before `synchronize_rcu` starts, find the successor in its previous location, where it remains until they complete (Figure 3(c)-(e)). Searches that start after `synchronize_rcu` starts, find the new copy of the successor.

Note that during a `delete`, there might be two copies of the successor—in the original and in the new locations. This motivates the following weak BST (WBST) property (formally defined in Definition 1): all descendants in the left sub-tree of `v` have keys smaller than the key of `v`, and all descendants in the right sub-tree of `v` have keys larger than or equal to the key of `v`.

The WBST property allows multiple nodes with the same key. If all nodes with the same key hold the same value, preserving the WBST property ensures that `contains` is correct, as it may return the value of some duplicate node, and ignore the others.

Updates synchronize among themselves by acquiring locks on nodes returned by `get`. To avoid RCU deadlocks, locks are acquired outside the read-side critical section. This creates a region of uncertainty, for example, if an `insert` reaches the node with key `k`, but an overlapping `delete` operation removes this node from the tree before `insert` locks it (see Figure 5). If `insert` continues its operation, the new key will not be part of the tree.

We overcome this problem by validating the nodes after locking them, and restarting the operation if validation fails. An update restarts either because the locked nodes no longer have a parent-child relation or because one (or both) of the two nodes was deleted from the tree. Validating the parent-child relation is done locally by checking the child pointer of the parent node. When the operation validating is an `insert`, the child pointer of the parent should be `⊥` (null pointer). A `tag` is added to each child field, in order to avoid an ABA problem (a child pointer changed to non-`⊥` when a leaf is inserted, then back to `⊥`, when the leaf is moved by a `delete`). (A total of two tag fields in each node, one for each child.) A tag field is initialized to zero, and incremented every time the corresponding child field is set to `⊥`. Validating that a node was not removed is also done locally, using a `marked` field indicating that the node was deleted (in a manner similar to [14]). If validation fails, the operation releases all locks and restarts from the root.

**Detailed description.** All operations use procedure `get` to search for a key `k`, which starts from the root, and returns two nodes, `curr` and `prev`, which is the parent of `curr` (Line 15). The procedure also returns a `direction`, indicating whether `curr` is the left or right child of `prev`, and the `tag` of `prev` that is associated with direction. If the key `k` exists in the tree, then `curr` is the node containing `k`. If the key does not exist, `curr` = `⊥` and `prev` is the last node found in the search. All operations executed by `get` are wrapped with `rcu_read_lock` and `rcu_read_unlock` creating a read-side critical section (Lines 2 and 14).

![Figure 3: (a) and (b) show delete(k) when a node has one child. (c)-(e) show delete(k) that replaces a node with two children with its successor; synchronize_rcu is invoked between (d) and (e).](attachment:image.png)

![Figure 4: Search for key k' returns a false negative response due to an overlapping delete.](attachment:image.png)

![Figure 5: insert can add a node to an incorrect location due to an overlapping delete.](attachment:image.png)
Function &get:

```c
function get(key)
rcu_read_lock
prev ← root
curr ← prev.child[right] // root is never ⊥
currentKey ← curr.key // root’s right child is never ⊥
direction ← right
while curr ≠ ⊥ and currentKey ≠ key do
  prev ← curr
  direction ← (currentKey > key ? left : right)
  curr ← prev.child[direction]
if curr ≠ ⊥ then
  currentKey ← curr.key
tag = prev.tag[direction]
rcu_read_unlock
return (prev.tag,curr,direction)
```

Function &contains:

```c
function contains(key)
(prev,tag,curr,direction) ← get(key)
if curr = ⊥ then // The key was not found
  return false
return curr.value
```

Function &insert:

```c
function insert(key,value)
loop
  (prev,tag,curr,direction) ← get(key)
  if curr ≠ ⊥ then // The key was found
    return false
  lock(prev)
  if validate(prev,tag,⊥,direction) then
    node ← new(key,value,⊥,⊥) // Create a new leaf node
    prev.child[direction] ← node
    unlock(prev)
    return true
  unlock(prev) // Validation failed, release locks and retry
```

Function &validate:

```c
function validate(prev,tag,curr,direction)
if prev.marked or prev.child[direction] ≠ curr then
  return false
if curr ≠ ⊥ then // If curr ≠ ⊥, validate curr’s marked bit
  return curr.marked
return prev.tag[direction] = tag // Otherwise validate tag
```

The &incrementTag\( procedure receives a node and a direction, if the child of node in this direction is ⊥, it increments the tag associated with this direction.

```c
function incrementTag(node,direction)
if node.child[direction] = ⊥ then
  node.tag[direction] ← node.tag[direction]+1
```

The &delete\( operation invokes &get\( if the key is not found, it returns false (Line 46). If the key is found, &prev and &curr are locked (Lines 47, 48) and validated (Line 49). If validation fails, the operation unlocks &prev and &curr and starts over. If &curr has only one child, &delete does not use the successor (Line 50), &curr is marked (Line 51) and deleted (Line 53).

If &curr has two children (Line 57), the operation tries to delete &curr by using a successor. It finds the successor &suc and its parent &prevSucc, by traversing the leftmost branch of the sub-tree rooted at &curr (Lines 58-64). This traversal does not need a read-side critical section since the keys of nodes traversed do not impact the direction. Once found, &prevSucc and &suc are locked (Lines 67, 68) and validated (Line 69). If validation fails, all nodes are unlocked and the operation starts over; otherwise, a deletion using a successor is executed (Lines 72-83). A new node with &suc’s key and &curr’s children is created (Line 70) and locked (Line 71). Then, &curr is marked (Line 72) and replaced by &node (Line 73). Next, the operation waits for all pre-existing readers, by invoking &synchronize_rcu (Line 74). When the operation returns, it removes the old successor from the tree (Lines 75-80); note that &suc might be the right child of &curr (Line 76). The operation completes by unlocking all the nodes and returning true (Line 83). After removing a node from the tree (in one of Lines 53, 77 or 80), &incrementTag\( is called in case the tag should be updated.

### 4. LINEARIZABILITY OF CITRUS

Fix an execution \( \pi \) of the CITRUS algorithm. Roughly speaking, an implementation is linearizable [17] if it is possible to identify, for each operation, a linearization point, inside its interval, so that the responses of the operations are consistent: that is, they are the same as if the operations were performed sequentially in their linearization points.

The following notation is used in the proof. Let \( u, v \) be two nodes in the tree, \( v \rightarrow u \) indicates that \( u \) is a child of \( v \) \( \rightarrow u \) when \( u \) is the left child of \( v \) \( \rightarrow u \) when \( u \) is the right child of \( v \); generally, \( v \rightarrow u, d = \ell \) means that \( u \) is the left child of \( v \), and analogously if \( d = \textit{right} \). There is a path from node \( v \) to node \( u \) in configuration \( C \); if there is a sequence of nodes \( v = v_0, v_1, ..., v_m = u, m > 0 \), such that for every \( i, 0 \leq i < m, v_i \rightarrow v_{i+1} \) in \( C \). We denote \( PC(v,u)=v_0, ..., v_m \). A node \( v \) is reachable in configuration \( C \) if there is a path from the root to \( v \) in \( C \). A key \( k \) is reachable in configuration \( C \) if there is a reachable node \( v \) such that \( k = \text{Key}(v) \); recall that \( \text{Key}(v) \) never changes. The set of reachable keys stored in the sub-tree rooted at \( v \) in a configuration \( C \), is denoted \( \text{Set}_k(v) \); \( \text{Set}_k(\text{root}) \) is the set of reachable keys in the whole tree. We define the WBST property:

**Definition 1.** The weak BST (WBST) property holds in configuration \( C \) if for every internal node \( v \), if \( u \) is a node such that \( v \rightarrow u \) and \( k \in \text{Set}(u) \) then \( k < \text{Key}(v) \) and if \( v \rightarrow u \) and \( k \in \text{Set}(u) \) then \( k \geq \text{Key}(v) \). The WBST property holds in an execution prefix \( \sigma \) if it holds in every configuration \( C \) in \( \sigma \).
If the WBST property holds in an execution prefix \( \sigma \) that ends in configuration \( C \), then the range of keys in the sub-tree of node \( v \), denoted \( \text{Range}_C(v) \), can be defined by induction on \( \rho_C(\text{root}, v)\): 
\[
\text{Range}_C(\text{root}) = (-\infty, \infty) ; \text{if the parent of} \ v \text{ is node } u \text{ with } k = \text{Key}(u) \text{ and } \text{Range}_C(u) = [\text{min}_u, \text{max}_u], \text{ then } \text{Range}_C(v) = [\text{min}_u, k] \text{ or } \text{Range}_C(v) = (\text{min}_u, k) \text{ if } u \xrightarrow{\ell} v \text{, and } \text{Range}_C(v) = [k, \text{max}_u] \text{ if } u \xrightarrow{\ell} v. 
\]

Let \( u, v, w \) be nodes such that \( u \xrightarrow{d} v \xrightarrow{d} w \), \( d \in \{\ell, \text{right}\} \), and node \( v \) has only one child, in configuration \( C \in \pi \). A primitive write operation \( s \) that immediately follows \( C \) is a bypass of node \( v \) if \( u \xrightarrow{\ell} w \) in the configuration that follows \( s \) in \( \pi \) (Figure 3(a) and (b)). The write in Line 53 is a bypass of the node \( \text{curr} \), and the writes in Line 77 and Line 80 are a bypass of the node \( \text{succ} \).

In order to distinguish between variables of different operations, we denote by \( \text{var}_\text{op} \) the variable \( \text{var} \) of operation \( \text{op} \); the operation is omitted when it is clear from the context. An operation \( \text{op} \) accesses a node \( v \) when \( \text{op} \) reads one of \( v \)'s fields. Only updates execute primitive writes, and when an update \( \text{op} \) writes to a field of node \( v \), \( v \) is locked by \( \text{op} \). Therefore, the fields of node \( v \) cannot be modified by another operation until the lock on \( v \) is released.

Updates use \( \text{validate} \) to ensure that they operate on nodes in the tree. The next lemma argues that if a node is not reachable in a configuration, then its marked bit is \( \text{true} \). Note that a node is unmarked when it is added to the tree, and marked before it is removed. Since updates operate on locked nodes, this suffices to prove the lemma, even though marking and removing a node are not performed atomically.

**Lemma 1.** If a node \( v \) is not reachable in configuration \( C \in \pi \) and \( v \) is unmarked in configuration \( C' \in \pi \) that follows \( C \), then \( v \) is reachable in \( C' \).

**Proof.** Assume, by way of contradiction, that the property does not hold and let \( C' \) be the last configuration in \( \pi \) in which the property holds. Let \( s \) be the step by operation \( \text{op} \) immediately following \( C \); \( s \) must be a primitive write to a child field of a locked node, in an update.

Case 1: \text{insert}. Line 29: By the validation in Line 27, \( \text{prev} \) has no child in direction \( d \) (where \( d \) is the value of direction), hence, \( s \) does not make an unmarked node unreachable. Since \( \text{node} \) was created by \( \text{op} \), it is unreachable in the configuration immediately following the step in Line 28. If \( s \) makes \( \text{node} \) reachable, then \( \text{node} \) is unmarked, as required.

Case 2: \text{delete}. By the validation in Line 49, \( \text{prev} \xrightarrow{\ell} \text{curr} \) (where \( d \) is the value of direction).

Bypassing Lines 53, 77, 80 makes only \( \text{curr} \) or \( \text{succ} \) unreachable. The lemma holds since \( \text{curr} \) is marked before Line 53 and \( \text{succ} \) is marked before Lines 77 and 80.

Line 73: By Line 72, \( \text{curr} \) is marked and can become unreachable. Both children of \( \text{curr} \) become the children of \( \text{node} \) in Line 70. Since \( s \) adds \( \text{node} \) to the tree as a left or right child of \( \text{prev} \) (depending on \( d \)), \( \text{curr} \) is the only node that becomes unreachable. Since \( \text{node} \) is created by \( \text{op} \), it is unreachable in the configuration immediately following the step in Line 70. If \( \text{node} \) becomes reachable by \( s \), then it is unmarked, as required. \( \square \)

To ensure that \( \text{get} \) with key \( k \) is correct, we prove that all nodes accessed by \( \text{get} \) are in the tree at some point during its interval (Lemma 2) and that if \( k \) is in the tree throughout a search from the root, then it is found (Lemma 8). Recall that an operation \( \text{op} \) accesses a node \( v \) when \( \text{op} \) reads one of \( v \)'s fields.

**Lemma 2.** Let \( \pi' \) be the interval of operation \( \text{op} \). If \( \text{op} \) accesses a node \( v \) in step \( s \), then \( v \) is reachable in some configuration \( C \in \pi' \) that precedes \( s \).

**Proof.** By induction on the length of the path taken by \( \text{op} \) from the root to node \( v \). In the base case, the root is always reachable.

For the induction step, assume that the lemma holds for paths of length \( \ell < \ell \). Let \( v \) be the \( \ell \)th node in the path taken by \( \text{op} \) and let \( u \) be the \( (\ell - 1) \)th node in this path. Suppose, without loss of generality, that \( \text{op} \) read \( u \xrightarrow{\ell} v \) in step \( s' \).

By the induction hypothesis, \( u \) is reachable in some configuration during \( \pi' \); let \( C' \) be the last such configuration. If \( s' \) precedes \( C' \), then \( v \) is reachable in the configuration \( C' \) that immediately follows \( s' \). Since the pointer read in step \( s' \) precedes any access to \( v \), \( C' \) precedes \( s \) (Figure 6(a)).
If $s'$ does not precede $C$, we argue that $v$ is the right child of $u$ when $u$ is removed from the tree. Suppose, by way of contradiction, that $u$ has some other right child $w \neq v$ in $C$ and $s''$ is the step removing $u$ from the tree, i.e., the step that immediately follows $C$.

Since $\rho$ read $u \rightarrow v$, there is an update that sets the right child of $u$ to $v$ before $s'$. Since $u$ is reachable in $C$ and unreachable in the configuration following $s''$, Lemma 1 implies that $u$ is marked before $s''$. The update removing $u$ does not change $u$'s children fields (by the code, no operation writes to child fields of $C$ or $\rhou$). Any other update validates all locked nodes to check that they are not marked. Since $v$ is marked, any update that tries to write to $u$ restarts without changing $u$, which is a contradiction.

When $u$ is removed from the tree, $v$ is the right child of $u$. Since $C$ is the last configuration in $\pi'$ in which $v$ is reachable, both $v$ and $u$ are reachable in $C \in \pi'$. $C$ precedes $s'$ that precedes $s$, as required.

The next lemma shows the correctness of tag validation. It shows that if a child pointer is set to $\perp$, by another operation, after the tag is read and before the node is locked, then the tag is incremented during this interval. Due to space constraints, the proof appears in the full version.

**Lemma 3.** Let $\pi'$ be the interval starting with the step of Line 13 by $\rho$ and ending with the configuration that immediately follows the lock acquisition by $\rho$. If $\rho' \leftarrow \rho \leftarrow v$ is the interval of $\mu \in \pi'$ then $\rho' \leftarrow \rho \leftarrow v$ in every configuration in the interval of $\rho$ that follows $\pi'$.

Searches, implemented in $\text{get}$, follow the sequential algorithm, so it is critical to show that the tree maintains the WBST property. This is proved by induction on the prefixes of the execution $\pi$. A critical step in the proof is to show that an insert adds a node in the right location. This relies on the following lemma (Lemma 4), showing that the RCU mechanism of waiting for all pre-existing readers guarantees that if the search of an insert ends up in the wrong location, due to overlapping delete, then its validation fails.

**Lemma 4.** Let $\rho$ be an insert operation with key $k$ that successfully validates $\rho' \leftarrow \rho$ and $\rho''$, and let $\rho$ be the execution prefix that ends in configuration $C$ that immediately follows the return from validate. If the WBST property holds in $\pi$, then $\rho \in \text{Range}(\rho')$.

**Proof.** If the path traversed by $\rho$ is the same as $\rho' \leftarrow \rho \leftarrow v$, then, by the WBST property, $k \in \text{Range}(\rho')$. The path traversed by $\rho$ is different from $\rho' \leftarrow \rho \leftarrow v'$, if it was changed by an overlapping update $\rho'$. Consider every possible write to a child field, during $\pi'$, the interval of $\rho$.

Case 1: insert. Line 29: By the validation in Line 27, $\rho' \leftarrow \rho$ has no child in direction $d$ (where $d$ is the value of direction). Thus, $\rho''$ adds a leaf to the tree, without changing the range of $\rho' \leftarrow \rho$.

Case 2: delete. By the validation in Line 49, $\rho' \leftarrow \rho \leftarrow v$ (where $d$ is the value of direction).

Bypassing (Lines 53, 77, 80) only increases the ranges of other nodes and $k \in \text{Range}(\rho')$ is maintained.

Line 73: Both children of $\text{curroc}$ become the children of $\text{nodeoc}$ in Line 70. Hence, $\text{curroc}$ is replaced with $\text{nodeoc}$ that has the key of $\text{succoc}$. The node $\text{succoc}$ is found by traversing the leftmost branch of the sub-tree rooted at $\text{curroc}$. Since the WBST property holds in $\pi$, this traversal starting at $\text{curroc}$ finishes at the location of the successor of $\text{curroc}$. Since $\rho'$ accesses $\text{succoc}$, Lemma 2 implies that $\text{succoc}$ is reachable after some prefix of $\pi$ ending before Line 69. By the validation in Line 69 and Lemma 1, $\text{succoc}$ is reachable and it has no left child, implying that $\text{succoc}$ is the successor of $\text{curroc}$. Therefore, $\text{Key}(\text{succoc}) = k'' \geq k'$.

If $k < k'' < k''$, then $k \in \text{Range}(\rho')$ still holds.

Suppose otherwise that $k' < k < k''$ and $k \in \text{Range}(\rho')$. If the step of Line 73 by $\rho'$ precedes or is in the read-side critical section of $\rho$ then before removing the successor $\text{succoc}$ from the tree, $\rho'$ invokes $\text{sync}_\text{rooc}_\text{wbsst}$ and waits until the end of $\rho'$'s read-side critical section, by the RCU property. Since the WBST property holds during the get of $\rho'$, its search ends in $\text{succoc}$ (where $\text{Key}(\text{succoc}) = k'$ and $\text{succoc}$ equals $\text{prevo}$ and $\text{curroc}$ equals $\perp$). Since $\rho'$ is an insert, it invokes $\text{read}_\text{read}_\text{lock}$ and locks $\rho' \leftarrow \rho$. Prior to validation, the other operation $\rho''$ unlocks $\text{succoc}$ only after marking it in Line 75. Therefore, the validation of $\rho' \leftarrow \rho$ finds that it is marked, and fails (Figure 7).

Otherwise, the step of Line 73 by $\rho'$ is executed after the read-side critical section of $\rho$ and before $\rho' \leftarrow \rho$ locks $\rho'' \leftarrow \rho$. If $\rho'' \leftarrow \rho$ equals $\text{succoc}$ then $\rho'$ marks $\rho''$ before releasing its lock, and the validation of $\rho''$ fails. If not then by the WBST property the change was in the left sub-tree of $\text{prevo}$. In this case, validation fails, because either $\rho'' \leftarrow \rho \leftarrow v'$ and $\rho'' \leftarrow \rho \leftarrow v''$ (Lemma 3).

The WBST property now follows by a simple induction. Lemmas 1 and 2 show that the correct successor replaces a deleted node with two children (bypassing a node with one child clearly preserves the WBST property), while Lemma 4 shows that new nodes are correctly inserted. Due to space constraints, the proof appears in the full version.

**Lemma 5.** The WBST property holds after every prefix $\sigma$ of $\pi$.

The next lemma shows that each node has a single parent, and thus, delete makes a node unreachable in a single write to a child field. It is proved by way of contradiction, considering the first time when the property is violated. Due to space constraints, the proof appears in the full version.

**Lemma 6.** A node that is reachable in a configuration $C \in \pi$ has one reachable parent in $C$.

The following lemma is used to prove that if a node $v$ with key $k$ remains in the tree during a get, then $v$ can be found. The reason is that the path to $v$ never gets longer, and that even if a node $u$, that used to be on the path from the root to $v$, is read, there is still a path from $u$ to $v$.

**Lemma 7.** If there is a path from $u$ to $v$ in configuration $C \in \pi$ and $v$ is reachable in a configuration $C' \in \pi$ that follows $C$, then there is a path from $u$ to $v$ in $C'$ and $|\text{PC}(u, v)| \leq |\text{PC}(u, v)|$.

**Proof.** Assume, by way of contradiction, that $v$ is reachable in some configuration that follows $C$, but the properties are not maintained. Let $C'$ be the last configuration in which both properties
hold. Let s be the step immediately following C′; s must be a write to a child field of a locked node, by an update op. We consider all possible cases:

Case 1: insert. Line 29: By the validation in Line 27, prev has no child in direction d (where d is the value of direction), thus, op adds a leaf to the tree, which does not change the reachability of v from u nor the length of existing paths in the tree.

Case 2: delete. By the validation in Line 49, prev \( \xrightarrow{d} \) curr (where d is the value of direction).

Bypassing (Lines 53, 77, 80): By the conditions of the lemma and Lemma 6, the node being bypassed is not v (bypassing a node with a single reachable parent makes it unreachable). If the node being bypassed is u, then s does not change any child field on the path from u to v, and the path and its length are maintained. If the node being bypassed is any other node in \( p_{C'}(u, v) \), then s only shortens the path and the lemma holds.

Line 73: Both children of curr become the children of node, in Line 70, and curr is replaced with node. Therefore, only curr may become unreachable. By Lemma 6 curr becomes unreachable, and by the conditions of the lemma v \( \neq \) curr. If curr = u, then s does not change any child field on the path from u to v, and the path and its length are maintained. If curr is any other node in \( p_{C'}(u, v) \), and C′′ is the configuration immediately following s, then node \( \in p_{C'}(u, v) \), the path exists and its length remains the same.

We linearize the successful updates, as follows:

- A successful insert is linearized with its primitive write in Line 29.
- A successful delete is linearized with its primitive write in Line 53 or 73. (At most one of these lines is executed.)

Let \( s_1, s_2, \ldots \) be the primitive writes in \( \pi \), in which successful updates are linearized, in their order in \( \pi \). For every \( i \geq 1 \), let \( C_i \) be the configuration immediately preceding \( s_i \), and let \( C_i' \) be the configuration immediately following \( s_i \). A successful insert linearized at \( s_i \) is consistent if \( k \in \text{Set}_{C_i'}(\text{root}) \) and \( k \in \text{Set}_{C_i}(\text{root}) \), and a successful delete linearized at \( s_i \) is consistent if \( k \in \text{Set}_{C_i}(\text{root}) \) and \( k \notin \text{Set}_{C_i'}(\text{root}) \).

The next lemma is important for proving the consistency of successful inserts. It shows that if the key was inserted, by another operation, during the get of an insert then either (i) get finds the key or (ii) the key is inserted as a child of the last node in the search. Later, the first case of this lemma is used to linearize unsuccessful contains and delete.

**Lemma 8.** Let \( \pi' \) be read-side critical section of get \( (k) \), and suppose that every successful update linearized before the last configuration of \( \pi' \) is consistent. Let \( C \) be the first configuration in \( \pi' \) such that \( k \in \text{Set}_C(\text{root}) \) and assume that \( k \in \text{Set}_{C'}(\text{root}) \), for every configuration \( C' \in \pi' \) that follows \( C \). If \( C \) precedes the last access to a child field by get then, (i) get returns curr such that \( k = \text{Key}(\text{curr}) \). Otherwise, (ii) get returns \( \text{prev} \) such that \( \text{prev} \xrightarrow{d} v \) and \( k = \text{Key}(v) \) where d is the value of direction.

**Proof.** Let v be the closest reachable node to the root in C such that \( k = \text{Key}(v) \), and let \( v' \) be the value of \( \text{prev}_{\op} \) in C or the root if C precedes Line 3 of get. Since the WBST property holds in \( \pi^1 \) and insert adds leaves, \( v' \in \text{PC}(\text{root}, v) \).

If C does not precedes the last access to a child field by get, then \( \text{prev} = v' \xrightarrow{d} v \) as required.

If C precedes the last access to a child field by get, then get must reach a node with key k. Suppose otherwise that get does not reach a node with key k. If v is reachable in the last configuration of \( \pi^1 \), then the path from v′ to v does not get longer (Lemma 7).

Since the WBST property holds in \( \pi \), get traverses this path, and hence, if a node with key k is not found, then v is unreachable in the last configuration of \( \pi^1 \).

Let \( s \in \pi^1 \) be the primitive write by operation \( \op' \) that makes v unreachable. Let \( C_s \) be the configuration immediately preceding \( s \) in \( \pi' \) and \( C_s' \) be the configuration immediately following \( s \) in \( \pi' \); \( v \) is reachable in \( C_s \) and unreachable in \( C_s' \). By the validation in Line 27, insert does not make a node unreachable, and hence, only delete can make v unreachable. Since \( k \in \text{Set}_{C_s'}(\text{root}) \), \( \op' \) has key \( k' \neq k \). The only other way a delete makes a node unreachable is when \( v = \text{succ}_{\op'} \) and s is the primitive write of either Line 77 or Line 80.

Let \( s' \in \pi' \) be the primitive write by operation \( \op' \) that adds \( \text{node}_{\op'} \) to the tree. Let \( C_{s'} \) be the configuration immediately preceding \( s' \) in \( \pi' \) and let \( C_{s'}' \) be the configuration immediately following \( s' \) in \( \pi' \); \( v \) is reachable in \( C_{s'} \) and unreachable in \( C_{s'}' \). By the validation in Line 27, \( \text{succ}_{\op} \) does not make a node unreachable, and hence, only delete can make v unreachable. Since \( k \in \text{Set}_{C_{s'}}(\text{root}) \), \( \op' \) has key \( k' \neq k \). The only other way a delete makes a node unreachable is when \( v = \text{succ}_{\op'} \) and s is the primitive write of either Line 77 or Line 80.

Let \( s \in \pi' \) be the primitive write by operation \( \op' \) that adds \( \text{node}_{\op'} \) to the tree. Let \( C_{s} \) be the configuration immediately preceding \( s \) in \( \pi' \) and let \( C_{s}' \) be the configuration immediately following \( s \) in \( \pi' \); \( v \) is reachable in \( C_{s} \) and unreachable in \( C_{s}' \). By the validation in Line 27, insert does not make a node unreachable, and hence, only delete can make v unreachable. Since \( k \in \text{Set}_{C_{s}'}(\text{root}) \), \( \op' \) has key \( k' \neq k \). The only other way a delete makes a node unreachable is when \( v = \text{succ}_{\op'} \) and s is the primitive write of either Line 77 or Line 80.

The next lemma proves that there is only one node with key k, and thus, only one node should be made unreachable in order for delete to be consistent. Due to space constraints, the proof appears in the full version.

**Lemma 9.** Let v be a node that is successfully validated by an operation \( \op \). Let \( \sigma \) be the prefix ending with the configuration \( C \) that immediately follows the return from validate. If all successful updates linearized in \( \sigma \) are consistent, then there is no other reachable node \( v' \) such that \( \text{Key}(v') = \text{Key}(v) \).

**Lemma 10.** All successful updates are consistent.

**Proof.** By induction on the linearization points, \( s_1, s_2, \ldots \). For every \( j \geq 1 \), let \( C_j \) be the configuration immediately preceding \( s_j \) and \( C_j' \) be the configuration immediately following \( s_j \).

Base: A successful update \( \op \) is linearized at its first primitive write, hence, there is no write before \( s_1 \). Let k be the key of \( \op \), and
note that \( k \notin \text{Set}_C(\text{root}) \) for every configuration \( C \) preceding \( s_1 \), and by Lemma 2, \( \text{op} \) cannot access a node with key \( k \), hence, \( \text{op} \) is a successful \text{insert}. Since \( \text{op} \) accesses \text{prev}, by Lemma 2, \text{prev} is reachable in some configuration preceding the step of Line 27. Since \( \text{prev} \) is validated in Line 27 and is unmarked, Lemma 1 implies that \( \text{prev} \) is reachable in \( C_1 \), and \( k \in \text{Set}_{C_1}(\text{root}) \).

For the induction step, assume that the successful updates linearized at \( s_1, s_2, \ldots, s_{i-1} \) are consistent, and consider the successful update \( \text{op} \) linearized at \( s_i \).

Case 1: \( \text{op} \) is \text{delete}(\( k \)). If the statement in Line 45 implies that \( k = \text{Key}(\text{curr}) \), the validation in Line 49, \( \text{curr} \) and \( \text{prev} \) are both unmarked and \( \text{prev} \rightarrow \text{curr} \) (where \( d \) is the value of direction).

Since \( \text{op} \) accesses \( \text{prev} \) and \( \text{curr} \), Lemma 2 implies that they are reachable in some configuration preceding the step of Line 49. Furthermore, Lemma 1 implies that \( \text{prev} \) and \( \text{curr} \) are reachable in \( C_i \). Since \( \text{prev} \) is the only reachable parent of \( \text{curr} \) (Lemma 6), \( \text{curr} \) is unreachable in \( C_\pi \).

By the induction hypothesis, all successful updates linearized at \( s_1, s_2, \ldots, s_{i-1} \) are consistent. Together with the fact that \( \text{curr} \) validates \( \text{prev} \), this implies that \( \text{curr} \) is the only reachable node with key \( k \) in \( C_i \) (Lemma 9). Therefore, \( k \notin \text{Set}_{C_\pi}(\text{root}) \).

Case 2: \( \text{op} \) is \text{insert}(\( k \)). Since \( \text{op} \) accesses \( \text{prev}_{\text{op}} \), Lemma 2 implies that \( \text{prev}_{\text{op}} \) is reachable in some configuration preceding the step of Line 27. Since \( \text{prev}_{\text{op}} \) is validated in Line 27, it is unmarked, and Lemma 1 implies that it is reachable in \( C_i \). Since \( k = \text{Key}(\text{node}_{\text{op}}), k \in \text{Set}_{C_\pi}(\text{root}) \).

Assume, by way of contradiction, that \( k \in \text{Set}_{C_\pi}(\text{root}) \). Let \( s_j, j < i \) be the linearization point of a successful insert with key \( k \). By the induction hypothesis, and since \( k \in \text{Set}_{C_\pi}(\text{root}) \), no \text{delete}(\( k \)) is linearized between \( s_j \) and \( s_i \), and for every configuration \( C \) that follows \( C_j \) and precedes \( C_i \), \( k \in \text{Set}_{C}(\text{root}) \).

If \( C_\pi \) precedes or is in the read-side critical section of \( \text{op} \), then Lemma 8 implies that either \( k = \text{Key}(\text{curr}_{\text{op}}) \) or \( \text{prev}_{\text{op}} \rightarrow \text{curr} \) and \( k = \text{Key}(\text{v}) \) (where \( d \) is the value of direction). If \( k = \text{Key}(\text{curr}_{\text{op}}) \) then \( \text{curr}_{\text{op}} \neq \bot \) and \( \text{op} \) returns \text{false} in contradiction to \( \text{op} \) being a successful update. If \( \text{prev}_{\text{op}} \rightarrow \text{v} \), then the validation in Line 27 fails, in contradiction to the linearization of \( \text{op} \).

Otherwise, since the WBST property holds in \( \pi \), \( s_j \) inserts the new node as the child of \( \text{prev}_{\text{op}} \) in direction \( d \), and the validation of \( \text{op} \) fails, because either \( \text{prev}_{\text{op}}.\text{child}[d] \neq \bot \) or \( \text{prev}_{\text{op}}.\text{tag}[d] \neq \text{tag}_{\text{op}} \) (Lemma 3).

An operation \( \text{op} \) with interval \( \pi', \) of the remaining types, is linearized as follows:

- If \( \text{op} \) is a failed contains or a failed delete, then \( \text{curr} = \bot \).
- If \( k \in \text{Set}_{C}(\text{root}) \) for every configuration \( C \in \pi' \), then \( k \in \text{Set}_{C}(\text{root}) \) for every configuration \( C \in \pi'' \), where \( \pi'' \) is the read-side critical section of \( \text{op} \). By Lemma 8(ii), \( \text{op} \) reaches a node with key \( k \), which is a contradiction. Therefore, \( k \notin \text{Set}_{C}(\text{root}) \), for some configuration \( C \in \pi' \). The linearization point of \( \text{op} \) is after last \( C' \in \pi' \) such that \( k \notin \text{Set}_{C'}(\text{root}) \), and before the successful insert with the same linearization point, if such insert exists. This linearization point is clearly consistent.

- If \( \text{op} \) is a successful contains that returns a node \( v \)’s value or a failed insert that reached node \( v \) with key \( k \), then it is linearized after the last configuration \( C \in \pi' \) such that \( v \) is reachable in \( C \), and before the successful delete with the same linearization point, if such delete exists.

In \( \text{op} \), \( \text{curr} = v \), and by condition of the while loop in \text{get} (Line 7), \( k = \text{Key}(v) \). By Lemma 2, \( \text{curr} \) is reachable in some configuration \( C' \in \pi' \), and the linearization point exists. This linearization point is consistent since \( k \in \text{Set}_{C}(\text{root}) \) and if \( \text{op} \) is a successful contains, it returns \( v.\text{value} \).

Together with Lemma 10, this proves the linearizability of \text{CITRUS}. Furthermore, if there is only a finite number of keys, then every path from the root is finite, and \text{contains} is wait-free.

**Theorem 11.** The \text{CITRUS} algorithm is a linearizable implementation of a binary search tree, supporting wait-free \text{contains}.

### 5. Evaluation

**Setup.** We implemented \text{CITRUS} in C since both the RCU implementation and the RCU-based trees are implemented in C. We considered two RCU-based trees, the red-black tree [18] and Bonsai, a balanced search tree [6], both using a global lock to synchronize among updates.\(^1\) We also compared \text{CITRUS} to three concurrent dictionary data structures, which have C implementations: The optimistic AVL tree\(^2\) [4], the lock-free search tree [23] and the lock-based lazy skiplist\(^3\) [15].

The experiments were run on a machine with four AMD Opteron 6376 Processors, each with 16 cores, for a total of 64 cores. All memory allocation used the jemalloc library, to avoid synchronization bottlenecks during memory allocation.

Experiment sets were run with two key ranges, \([0, 2 \cdot 10^6]\) and \([0, 2 \cdot 10^6] \); in both sets, the tree was pre-filled to the size of half the key range. During pre-filling memory was reclaimed. For every test, each thread ran for five seconds, continuously executing randomly chosen operations with a randomly chosen key, without performing any memory reclamation. We show the overall throughput (total number of executed operations divided by the running time). Each experiment was run five times for each configuration of operation distribution, key range and thread count; we report the arithmetic average as the final result.

Running experiments in the unmanaged C environment revealed that memory management, and in particular cache usage, has a significant impact on performance. The size of nodes, order of fields, and their alignment inside cache lines, often influences the results much more than the algorithmic aspects of the implementation.

**New RCU.** During our initial evaluation of \text{CITRUS}, we identified that the user-space RCU implementation [8] does not scale for workloads with many concurrent updates, due to expensive synchronization among them, which include acquiring a global lock. The left side of Figure 8 shows representative results; similar behaviour was observed under different update contention and key ranges. To show that the drop in throughput is just an implementation issue, we re-implemented the subset of the RCU API used in \text{CITRUS}, in a manner similar to epoch-based reclamation [11]. In our implementation, each thread has a counter and flag, the counter counts the number of critical sections executed by the thread and a flag indicates if the thread is currently inside its read-side critical section. The \text{rcu_read_lock} operation increments the counter and sets the flag to true, while the \text{rcu_read_unlock} operation sets the flag to false. When a thread executes a \text{synchronize_rcu} operation, it waits for every other thread, until one of two things occurs:

\(^1\)We were unable to run the red-black tree variant [18] using transactional memory to optimistically handle conflicting updates.


6. RELATED WORK

Read copy update (RCU) was introduced [22] as a solution to lock contention and synchronization overhead in read-intensive workloads. RCU can be used for explicit memory reclamation [21]. A formal semantics for RCU appears in [12, 13].

Relativistic programming is a methodology for concurrent programming using RCU, which does not assume sequentially consistent memory. It instructs readers to access items in the data structure in an order that is reverse to the order that updates modify them, but it does not deal with concurrent updates. Relativistic programming was employed in a concurrent hash table [25, 26] in which a lock protects each bucket. Relativistic programming was also used in a red-black tree [18], which allows only one concurrent writer to the tree, optimistically enforced by transactional memory [16]. RCU was used in a different way in Bonsai, a balanced tree algorithm [6]. Inspired by functional programming, Bonsai never modifies the tree in place, creating instead a new instance for the changed data structure.

Many concurrent search trees were presented in recent years, several of them using fine-grained locks, e.g., [1, 4, 7, 9]. The AVL tree of Bronson et al. [4] uses fine-grained locks, and it is partially external and relaxed balanced. Other trees are nonblocking, e.g., [3, 5, 10, 11, 19, 23]. These algorithms typically use compare&swap primitives, and in some cases, even stronger primitives on several shared variables, like multi-word compare&swap [11] or the customized LLX, SCX and VLX primitives [5].

Several of these implementations [5, 3, 7, 9, 10, 23] also provides wait-free contains.

7. DISCUSSION

We have shown that it is possible, and even relatively simple, to design RCU-based concurrent search trees with concurrent updates. This opens up many interesting directions for future research. The obvious question is to extend CITRUS to a balanced search tree. It is also important to integrate into CITRUS two primary aspects of RCU usage, namely, efficient memory reclamation and out-of-order execution of memory instructions. A broader topic is to employ RCU in lock-free algorithms, using primitives such as compare-and-swap instead of locks. This will necessitate more refined mechanisms for synchronization among readers and updates, since synchronize_rcu is inherently blocking.

Acknowledgements. We would like to thank Hongseok Yang for suggesting that we look at concurrent algorithms using RCU, Noam Rinetzky for discussions of the correctness proof, Adam Morrison for his help with understanding the performance of the implementation and many insightful discussions, and finally, Yehonatan Rubin, Dana Drachsler, Alexander Libov and Eran Yahav for their help with the evaluation.

8. REFERENCES

Figure 10: Throughput of the different algorithms with key range $[0, 2 \cdot 10^5]$ and $[0, 2 \cdot 10^6]$ under different operation distribution; y-axis show the throughput (operations/sec), and x-axis show the number of threads.


