

# CONCURRENT BINARY SEARCH TREES: DESIGN AND OPTIMIZATIONS

by

Arunmoezhi Ramachandran

APPROVED BY SUPERVISORY COMMITTEE:

---

Neeraj Mittal, Chair

---

Balaji Raghavachari

---

Venkatesan Subbarayan

---

Rob F Van Der Wijngaart

Copyright © 2015

Arunmoezhi Ramachandran

All rights reserved

*To my parents and sister*

CONCURRENT BINARY SEARCH TREES: DESIGN AND OPTIMIZATIONS

by

ARUNMOEZHI RAMACHANDRAN, BE, MS

DISSERTATION

Presented to the Faculty of  
The University of Texas at Dallas  
in Partial Fulfillment  
of the Requirements  
for the Degree of

DOCTOR OF PHILOSOPHY IN  
COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

May 2016

## ACKNOWLEDGMENTS

I would like to thank my PhD advisor, Professor Neeraj Mittal for guiding me throughout my PhD journey. I'm also grateful to Dr. Rob Van Der Wijngaart who was my mentor during my internship with Intel. I enjoyed working with him on my first research project.

I'm thankful to Professor R. Chandrasekaran who has always been kind and helpful to me. I also have to thank my PhD committee, Professors Balaji Raghavachari and Venkatesan Subbarayan for their advice and suggestions in general.

December 2015

## PREFACE

This dissertation was produced in accordance with guidelines which permit the inclusion as part of the dissertation the text of an original paper or papers submitted for publication. The dissertation must still conform to all other requirements explained in the “Guide for the Preparation of Master’s Theses and Doctoral Dissertations at The University of Texas at Dallas.” It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts which provide logical bridges between different manuscripts are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the student’s contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the dissertation attest to the accuracy of this statement.

# CONCURRENT BINARY SEARCH TREES: DESIGN AND OPTIMIZATIONS

Publication No. \_\_\_\_\_

Arunmoezhi Ramachandran, PhD  
The University of Texas at Dallas, 2016

Supervising Professor: Neeraj Mittal

Over the last decade processor clock speeds have hit a wall. But the demand for performance improvements continued to grow. So there has been a major shift towards multi-core and many-core processors. However, most of the software applications running on multi-core processors do not fully utilize all the cores. Rewriting the entire software stack for large applications seems impractical. On the other hand, improving the performance of common data structures which are the building blocks of any software application seemed feasible. This motivated the design of concurrent data structures.

Designing a concurrent data structure is far more challenging than its sequential counterpart because threads executing concurrently may interleave in exponential possible ways. A concurrent data structure should preserve its equivalent sequential specifications for all such interleavings.

In this work, we focus on concurrent binary search trees. We present a blocking and a non-blocking algorithm for concurrent manipulation of a binary search tree in an asynchronous shared memory system that supports search, insert and delete operations. We also provide a general technique to optimize them. Our technique is sufficiently general in the sense that

it can be applied to a variety of concurrent binary search trees based on both blocking and non-blocking approaches.

Moreover, we also present several techniques to make search operations on such binary search trees *wait-free*. Our techniques have the advantage that a search operation does not need to perform any write instruction on shared memory thereby minimizing the cache traffic.

Experiments indicate that our algorithms perform best in most cases. And our optimization technique improves performance of our algorithms and other existing algorithms.



## TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . .	v
PREFACE . . . . .	vi
ABSTRACT . . . . .	vii
LIST OF FIGURES . . . . .	xi
LIST OF TABLES . . . . .	xii
CHAPTER 1 INTRODUCTION . . . . .	1
1.1 Contributions . . . . .	4
1.2 Dissertation Roadmap . . . . .	5
CHAPTER 2 SYSTEM MODEL . . . . .	6
2.1 Binary Search Tree . . . . .	6
2.2 Synchronization Primitives . . . . .	6
2.3 Proof of correctness . . . . .	7
CHAPTER 3 LOCK BASED CONCURRENT BINARY SEARCH TREE . . . . .	8
3.1 The Lock-Based Algorithm . . . . .	8
3.1.1 Overview of the Algorithm . . . . .	8
3.1.2 Details of the Algorithm . . . . .	9
3.1.3 Formal Description . . . . .	16
3.1.4 Correctness Proof . . . . .	22
CHAPTER 4 LOCK FREE CONCURRENT BINARY SEARCH TREE . . . . .	25
4.1 The Lock-Free Algorithm . . . . .	25
4.1.1 Overview of the Algorithm . . . . .	25
4.1.2 Details of the Algorithm . . . . .	27
4.1.3 Formal Description . . . . .	45
4.1.4 Correctness Proof . . . . .	47
CHAPTER 5 LOCAL RECOVERY FOR CONCURRENT BINARY SEARCH TREES	51
5.1 The Local Recovery Algorithm . . . . .	51

5.1.1	Overview of the Algorithm . . . . .	51
5.1.2	Details of the Algorithm . . . . .	57
CHAPTER 6 EXPERIMENTAL EVALUATION . . . . .		68
6.1	Experimental Setup . . . . .	68
6.2	Lock based tree . . . . .	70
6.3	Lock free tree . . . . .	74
6.4	Impact of local recovery . . . . .	78
CHAPTER 7 CONCLUSION . . . . .		83
REFERENCES . . . . .		85
VITA		

## LIST OF FIGURES

3.1	CASTLE - Sentinel keys and nodes . . . . .	10
3.2	CASTLE - Nodes in the access path of seek . . . . .	11
3.3	CASTLE - A case when last right turn node is no longer part of the tree . . . .	11
3.4	CASTLE - An illustration of an insert operation . . . . .	13
3.5	CASTLE - An illustration of an simple delete operation . . . . .	14
3.6	CASTLE - An illustration of a complex delete operation . . . . .	15
4.1	ELFTREE - An illustration of an insert operation. . . . .	26
4.2	ELFTREE - An illustration of a simple delete operation. . . . .	26
4.3	ELFTREE - An illustration of a complex delete operation. . . . .	27
4.4	ELFTREE - Nodes in the access path of seek along with sentinel keys and nodes	28
4.5	ELFTREE - Sequence of steps in an insert operation . . . . .	31
4.6	ELFTREE - Sequence of steps in a delete operation . . . . .	34
5.1	An illustration of a key moving up the tree . . . . .	52
6.1	CASTLE - Comparison of throughput of different algorithms - thread sweep . .	71
6.2	CASTLE - Comparison of throughput of different algorithms - key sweep . . .	72
6.3	ELFTREE - Comparison of throughput of different algorithms - key sweep . . .	75
6.4	ELFTREE - Comparison of throughput of different algorithms - thread sweep . .	76
6.5	Effect of local recovery on system throughput . . . . .	80
6.6	Effect of local recovery on seek time . . . . .	81

## LIST OF TABLES

6.1	Comparison of different concurrent algorithms in the absence of contention. . . .	73
6.2	Comparison of different lock-free algorithms in the absence of contention. . . .	77
6.3	Effect of local recovery on system throughput . . . . .	82

# CHAPTER 1

## INTRODUCTION

With the growing prevalence of multi-core, multi-processor systems, concurrent data structures are becoming increasingly important. In such a data structure, multiple processes may need to operate on the data structure at the same time. Contention between different processes must be managed in such a way that all operations complete correctly and leave the data structure in a valid state.

Concurrency is often managed using locks. A lock can be used to achieve mutual exclusion, which can then be used to ensure that any updates to the data structure or a portion of it are performed by one process at a time. This makes it easier to design a lock-based concurrent data structure and reason about its correctness. Moreover, this also makes it easier to implement a lock-based data structure and debug it than its lock-free counterpart. Lock-based algorithms for concurrent versions of many important data structures for storing and managing shared data have been developed including linked lists, queues, priority queues, hash tables and skiplists (*e.g.*, (Michael and Scott, 1996; Michael, 2002; Lea, 2003; Heller et al., 2005; Lev et al., 2007; Herlihy and Shavit, 2012)).

Binary search tree is one of the fundamental data structures for organizing *ordered* data that supports search, insert and delete operations (Cormen et al., 1991). A binary search tree may be unbalanced (different leaf nodes may be at very different depths) or balanced (all leaf nodes are at roughly the same depth). A balanced binary search tree provides better worst-case guarantees about the cost of performing an operation on the tree. However, in many cases, the overhead of keeping the tree balanced, especially in a concurrent environment, may incur significant overhead. As a result, in many cases, an unbalanced binary search

tree outperforms a balanced binary search tree in practice. In this work, our focus is on developing an efficient concurrent algorithms for an *unbalanced* binary search tree.

Concurrent algorithms for unbalanced binary search trees have been proposed in (Ellen et al., 2010; Howley and Jones, 2012; Natarajan and Mittal, 2014; Drachsler et al., 2014; Ellen et al., 2014; Chatterjee et al., 2014; Arbel and Attiya, 2014). Algorithms in (Drachsler et al., 2014; Arbel and Attiya, 2014) are blocking (or lock-based), whereas those in (Ellen et al., 2010; Howley and Jones, 2012; Natarajan and Mittal, 2014; Ellen et al., 2014; Chatterjee et al., 2014) are non-blocking (or lock-free). Also, algorithms in (Howley and Jones, 2012; Drachsler et al., 2014; Arbel and Attiya, 2014; Chatterjee et al., 2014) use internal representation of a search tree in which all nodes store data, where as those in (Ellen et al., 2010; Natarajan and Mittal, 2014; Ellen et al., 2014) use an external representation of a search tree in which only leaf nodes store data (data stored in internal nodes is used for routing purposes only).

Algorithms that use internal representation of a search tree have to address the problem that arises due to a key moving from one location in the tree to another. This occurs when the key undergoing deletion resides in a binary node, which requires it to be either replaced with its predecessor (next smallest key) or its successor (next largest key). As a result, an operation traversing the tree may fail to find the target key both at its old location and at its new location, even though the target key was continuously present in the tree. Different algorithms use different approaches to handle the problem arising due to key movement. The algorithm by Drachsler *et al.* in (Drachsler et al., 2014) maintains a *sorted* linked list of all the keys in the tree. If the traversal of the tree fails to find a given key, then an operation traverses the linked list to look for the key. The algorithm by Arbel and Hattiya (Arbel and Attiya, 2014) uses the RCU (Read-Copy-Update) framework (first employed in Linux kernels) to allow reads to occur concurrently with updates.

Most of the concurrent algorithms (for BSTs) that have proposed so far use a naïve approach and simply restart the traversal from the root of the tree (Ellen et al., 2010; Howley

and Jones, 2012; Natarajan and Mittal, 2014; Drachsler et al., 2014; Arbel and Attiya, 2014). This is especially undesirable if the tree has large height, and the overhead of repeatedly traversing the tree may dominate all other overheads of performing an operation.

Recently, a few algorithms have been proposed in which an operation attempts to recover from a failure *locally* (Ellen et al., 2014; Chatterjee et al., 2014). The algorithm in (Ellen et al., 2014), which is based on external representation and builds upon the algorithm in (Ellen et al., 2010), maintains a stack of the nodes visited during the traversal of the tree and simply restarts from the last “unmarked” node (a node is marked before it is removed from the tree). Intuitively, this works because, in an external search tree, keys *do not* move from one location in the tree to another. However, in a search tree based on internal representation, keys may move from one location (in the tree) to another. This occurs when the key undergoing deletion resides in a binary node, which requires it to be either replaced with its predecessor (next smallest key) or its successor (next largest key). This causes two problems. First, an operation traversing the tree may fail to find the target key both at its old location and at its new location, even though the target key was continuously present in the tree. Second, it is not always safe to simply restart from the last unmarked node in the stack since the key may have moved to an ancestor of such a node. Their restart approach is, however, sufficiently general that it can be applied to other concurrent search trees based on external representation (Natarajan and Mittal, 2014).

The algorithm in (Chatterjee et al., 2014), which is based on internal representation, uses *backlink pointers* to find its way and recover from a failure while executing an operation. Their restart approach appears to be customized for their concurrent search tree and is not clear how it can be extended to other concurrent search trees based on internal representation.

## 1.1 Contributions

First we present a new *lock-based* algorithm for concurrent manipulation of a binary search tree in an asynchronous shared memory system that supports search, insert and delete operations. Our algorithm is based on an internal representation of a search tree as in (Drachsler et al., 2014; Arbel and Attiya, 2014). However, as in (Natarajan and Mittal, 2014), it operates at edge-level (locks edges) rather than at node-level (locks nodes); this minimizes the contention window of a write operation and improves the system throughput. Further, in our algorithm, (i) a search operation uses only read and write instructions, (ii) an insert operation does not acquire any locks, and (iii) a delete operation only needs to lock up to four edges in the absence of contention. Our experiments indicate that our lock-based algorithm outperforms existing algorithms for a concurrent binary search tree—blocking as well as non-blocking—for medium-sized and larger trees, achieving up to 59% higher throughput than the next best algorithm.

Second we extend the previous algorithm to develop a new *lock-free* algorithm. It combines ideas from two existing lock-free algorithms, namely those by Howley and Jones (Howley and Jones, 2012) and Natarajan and Mittal (Natarajan and Mittal, 2014), and is especially *optimized for the conflict-free scenario*. Like Howley and Jones’ algorithm, it uses internal representation of a search tree in which all nodes store keys. Also, like Natarajan and Mittal’s algorithm, it operates at edge-level rather than node-level and does not use a separate explicit object for enabling coordination among conflicting operations. As a result, it inherits benefits of both the lock-free algorithms. Specifically, when compared to modify operations of Howley and Jones’ internal binary search tree, its modify operations (a) have a smaller contention window, (b) allocate fewer objects, (c) execute fewer atomic instructions, and (d) have a smaller memory footprint. Our experiments indicate that our new lock-free algorithm outperforms other lock-free algorithms in most cases, providing up to 35% improvement in some cases over the next best algorithm.



Third, we present a general approach for local recovery that enables a process to quickly recover from a failure while performing an operation by restarting the traversal from a point “close” to the operation’s window rather than the root of the tree. Our approach can be applied to many existing concurrent algorithms for maintaining binary search trees using internal representation—blocking as well as non-blocking—such as those in (Howley and Jones, 2012; Drachsler et al., 2014; Arbel and Attiya, 2014; Ramachandran and Mittal, 2015b). Our local recovery approach uses only local variables and does not require modifying a tree node (of the original algorithm) to store any additional information. Using experimental evaluation, we demonstrate that our local recovery approach can yield significant speed-ups for many concurrent algorithms.

Finally, we present two light-weight techniques to make search operations for concurrent binary search trees based on internal representation, such as those in those in (Howley and Jones, 2012; Drachsler et al., 2014; Arbel and Attiya, 2014; Ramachandran and Mittal, 2015a,b), *wait-free* with low additional overhead. Both of our techniques have the desirable feature that a search operation does not need to perform any write instructions on the share memory thereby minimizing the cache coherence traffic.

## 1.2 Dissertation Roadmap

This dissertation organized as follows. We describe our system model in Chapter 2. Our lock-based algorithm for a binary search tree is described in Chapter 3. Our lock-free algorithm for a binary search tree is described in Chapter 4. Our general technique for local recovery is described in Chapter 5. The experimental evaluation of different concurrent algorithms for a binary search tree is described in Chapter 6. Finally Chapter 7 concludes the dissertation and outlines directions for future research.

## CHAPTER 2

### SYSTEM MODEL

#### 2.1 Binary Search Tree

We assume that a binary search tree (BST) implements a dictionary abstract data type and supports *search*, *insert* and *delete* operations. For convenience, we refer to the insert and delete operations as *modify* operations. A search operation explores the tree for a given key and returns **true** if the key is present in the tree and **false** otherwise. An insert operation adds a given key to the tree if the key is not already present in the tree. Duplicate keys are not allowed in our model. A delete operation removes a key from the tree if the key is indeed present in the tree. In both cases, a modify operation returns **true** if it changed the set of keys present in the tree (added or removed a key) and **false** otherwise.

A binary search tree satisfies the following properties:

- (a) the left subtree of a node contains only nodes with keys less than the node's key,
- (b) the right subtree of a node contains only nodes with keys greater than or equal to the node's key, and
- (c) the left and right subtrees of a node are also binary search trees.

#### 2.2 Synchronization Primitives

We assume an asynchronous shared memory system that, in addition to read and write instructions, also supports compare-and-swap (CAS) atomic instruction. A compare-and-swap instruction takes three arguments: *address*, *old* and *new*; it compares the contents of

a memory location (*address*) to a given value (*old*) and, only if they are the same, modifies the contents of that location to a given new value (*new*). The **CAS** instruction is commonly available in many modern processors such as Intel 64 and AMD64.

We also use locks and assume that the following properties hold true about the locks

- (a) safe: it satisfies the mutual exclusion property, *i.e.*, at most one process can hold the lock at any time, and
- (b) live: it satisfies the deadlock freedom property, *i.e.*, if the lock is free and one or more processes attempt to acquire the lock, then some process is eventually able to acquire the lock.

### 2.3 Proof of correctness

To demonstrate the correctness of our algorithm, we use *linearizability* (Herlihy and Wing, 1990) for the safety property and *deadlock-freedom* (Herlihy and Shavit, 2012) for the liveness property. Broadly speaking, linearizability requires that an operation should appear to take effect instantaneously at some point during its execution. Deadlock-freedom requires that some process with a pending operation be able to complete its operation eventually.

## CHAPTER 3

### LOCK BASED CONCURRENT BINARY SEARCH TREE

#### 3.1 The Lock-Based Algorithm

We first provide an overview of our algorithm. We then describe the algorithm in more detail and also give its pseudo-code. For ease of exposition, we describe our algorithm assuming no memory reclamation, which can be performed using the well-known technique of hazard pointers (Michael, 2004).

##### 3.1.1 Overview of the Algorithm

Every operation in our algorithm uses *seek* function as a subroutine. The seek function traverses the tree from the root node until it either finds the target key or reaches a non-binary node whose next edge to be followed points to a null node. We refer to the path traversed by the operation during the seek as the *access-path*, and the last node in the access-path as the *terminal node*. The operation then compares the target key with the stored key (the key present in the terminal node). Depending on the result of the comparison and the type of the operation, the operation either terminates or moves to the execution phase. In certain cases in which a key may have moved upward along the access-path, the seek function may have to restart and traverse the tree again; details about restarting are provided later. We now describe the next steps for each of the type of operation one-by-one.

**Search:** A search operation starts by invoking seek operation. It returns **true** if the stored key matches the target key and **false** otherwise.

**Insert:** An insert operation starts by invoking seek operation. It returns **false** if the target key matches the stored key; otherwise, it moves to the execution phase. In the execution phase, it attempts to insert the key into the tree as a child node of the last node in the access-path using a CAS instruction. If the instruction succeeds, then the operation returns **true**; otherwise, it restarts by invoking the seek function again.

**Delete:** A delete operation starts by invoking seek function. It returns **false** if the stored key does not match the target key; otherwise, it moves to the execution phase. In the execution phase, it attempts to remove the key stored in the terminal node of the access-path. There are two cases depending on whether the terminal node is a binary node (has two children) or not (has at most one child). In the first case, the operation is referred to as *complex delete operation*. In the second case, it is referred to as *simple delete operation*. In the case of simple delete, the terminal node is removed by changing the pointer at the parent node of the terminal node. In the case of complex delete, the key to be deleted is replaced with the *next largest* key in the tree, which will be stored in the *leftmost node* of the *right subtree* of the terminal node.

### 3.1.2 Details of the Algorithm

As in most algorithms, to make it easier to handle special cases, we use sentinel keys and sentinel nodes. The structure of an empty tree with only sentinel keys (denoted by  $\infty_1$  and  $\infty_2$  with  $\infty_1 < \infty_2$ ) and sentinel nodes (denoted by  $\mathbb{R}$  and  $\mathbb{S}$ ) is shown in Figure 3.1.

Our algorithm, like the one in (Natarajan and Mittal, 2014), operates at edge level. A delete operation obtains ownership of the edges it needs to work on by locking them. To enable locking of an edge, we steal a bit from the child addresses of a node referred to as *lock-flag*. We also steal another bit from the child addresses of a node to indicate that the node is undergoing deletion and will be removed from the tree. We denote this bit by

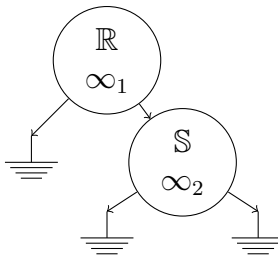


Figure 3.1: Sentinel keys and nodes ( $\infty_1 < \infty_2$ )

*mark-flag*. Finally, to avoid the ABA problem, as in Howley and Jones (Howley and Jones, 2012), we use *unique* null pointers. To that end, we steal yet another bit from the child address, referred to as *null-flag*, and use it to indicate whether the address points to a null or a non-null address. So, when an address changes from a non-null value to a null value, we only set the null-flag and the contents of the address are not otherwise modified. This ensures that all null pointers are unique.

We next describe the details of the seek operation, which is executed by all operations (search as well as modify) after which we describe the details of the execution phase of insert and delete operations.

### The Seek Phase

A seek function keeps track of the node in the access-path at which it took the last “right turn” (*i.e.*, it last followed a right edge). Let this “right turn” node be referred to as *anchor node* when the traversal reaches the terminal node. Note that the terminal node is the node whose key matched the target key or whose next child edge is set to a null address. For an illustration, please see Figure 3.2. In the latter case (stored key does not match the target key), it is possible that the key may have moved up in the tree. To ascertain that the seek function did not miss the key because it may have moved up during the traversal, we use the following set of conditions that are *sufficient* (but not necessary) to guarantee that the seek function did not miss the key. First, the anchor node is still part of the tree. (For an

illustration, see Figure 3.3) Second, the key stored in the anchor node has not changed since it first encountered the anchor node during the (current) traversal. To check for the above two conditions, we determine whether the anchor node is undergoing deletion by examining its right child edge. We discuss the two cases one-by-one.

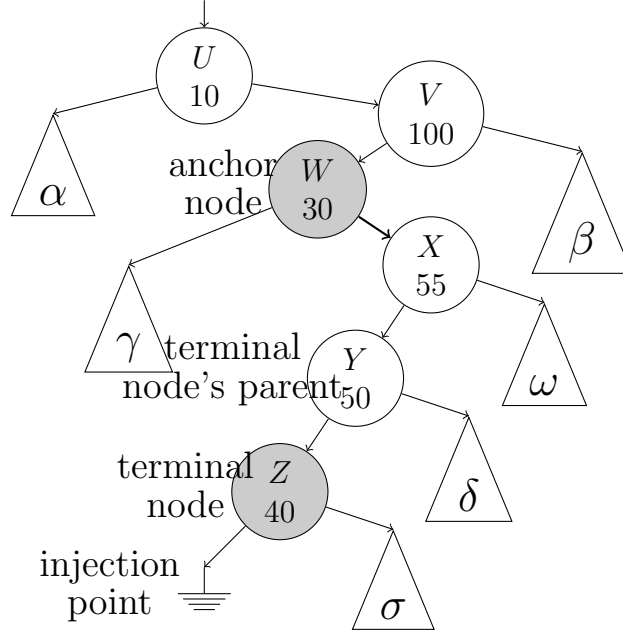


Figure 3.2: Nodes in the access path of seek

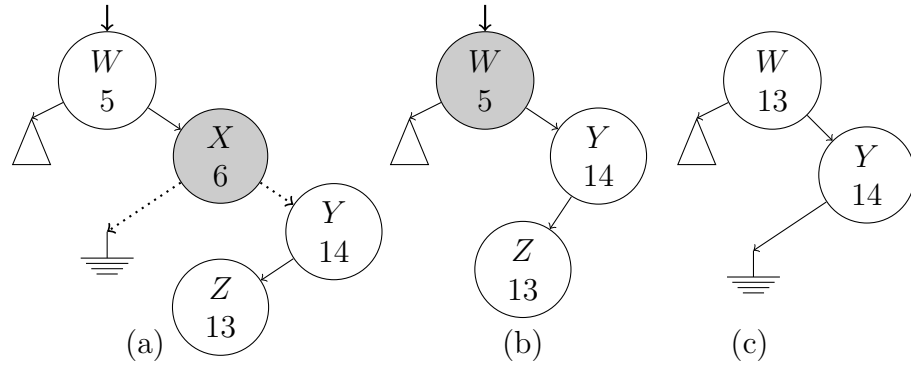


Figure 3.3: A case when last right turn node is no longer part of the tree

(a) *Right child edge not marked:* In this case, the anchor node is still part of the tree.

We next check whether the key stored in the anchor node has changed. If the key

has not changed, then the seek function returns the results of the (current) traversal, which consists of three addresses: (i) the address of the terminal node, (ii) the address of its parent, and (iii) the null address stored in the child field of the terminal node that caused the traversal to terminate. The last address is required to ensure that an insert operation works correctly (specifically to ascertain that the child field of the terminal node has not undergone any change since the completion of the traversal). We refer to it as the *injection point* of the insert operation. On the other hand, if the key has changed, then the seek function restarts from the root of the tree. A possible optimization is that the seek function restarts only if the target key is now less than the anchor node's key.

- (b) *Right child edge marked:* In this case, we compare the information gathered in the current traversal about the anchor node with that in the previous traversal, if one exists. Specifically, if the anchor node of the previous traversal is same as that of the current traversal and the keys found in the anchor node in the two traversals also match, then the seek function terminates, but returns the results of the previous traversal (instead of that of the current traversal). This is because the anchor node was definitely part of the tree during the previous traversal since it was reachable from the root of the tree at the beginning of the current traversal. Otherwise, the seek function restarts from the root of the tree.

For insert and delete operations, we refer to the terminal node as the *target node*.

### The Execution Phase of an Insert Operation

In the execution phase, an insert operation creates a new node containing the target key. It then adds the new node to the tree at the injection point using a CAS instruction. For an illustration, see Figure 3.4. If the CAS instruction succeeds, then (the new node becomes a



part of the tree and) the operation terminates; otherwise, the operation restarts from the seek phase. Note that the insert operations are lock-free.

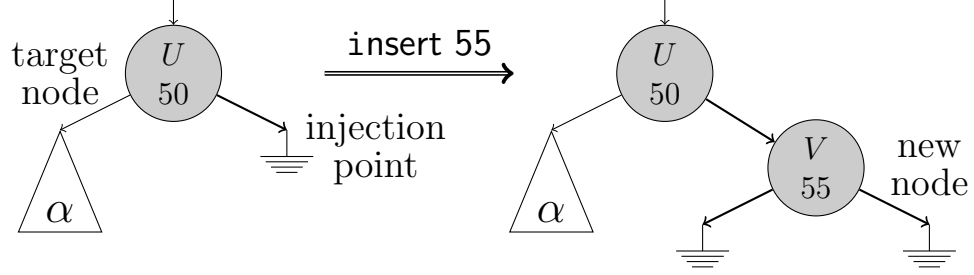


Figure 3.4: An illustration of an insert operation

### The Execution Phase of a Delete Operation

The execution of a delete operation starts by checking if the target node is a binary node or not. If it is a binary node, then the delete operation is classified as complex; otherwise it is classified as simple.

For a tree node  $X$ , let  $X.parent$  denote its parent node, and  $X.left$  and  $X.right$  denote its left and right child node, respectively. Also, hereafter in this section, let  $T$  denote the target node of the delete operation under consideration.

- (a) *Simple Delete*: In this case, either  $T.left$  or  $T.right$  is pointing to a null node. Note that both  $T.left$  and  $T.right$  may be pointing to null nodes in which case  $T$  will be a leaf node. Without loss of generality, assume that  $T.right$  is a null node. The removal of  $T$  involves locking the following three edges:  $\langle T.parent, T \rangle$ ,  $\langle T, T.left \rangle$  and  $\langle T, T.right \rangle$ . For an illustration, see Figure 3.5.

A lock on an edge is obtained by setting the lock-flag in the appropriate child field of the parent node using a CAS instruction. For example, to lock the edge  $\langle X, Y \rangle$ , where  $Y$  is the left child of  $X$ , the lock-flag in the left child of  $X$  is set to one. If all the edges are locked successfully, then the operation validates that the key stored in the

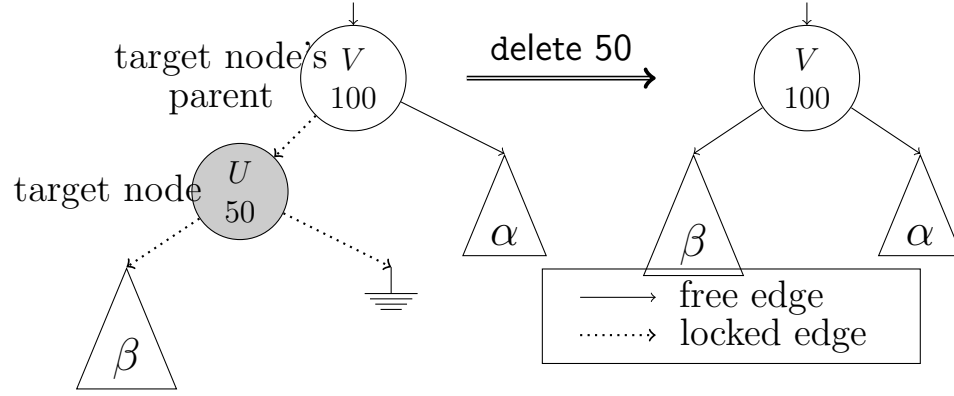


Figure 3.5: An illustration of a simple delete operation

target node still matches the target key. If the validation succeeds, then the operation marks both the children edges of  $T$  to indicate that  $T$  is going to be removed from the tree. Next, it changes the child pointer at  $T.parent$  that is pointing to  $T$  to point to  $T.left$  using a simple write instruction. Finally, the operation releases all the locks and returns **true**.

- (b) *Complex Delete*: In this case, both  $T.left$  and  $T.right$  are pointing to non-null nodes. The operation locates the next largest key in the tree, which is the smallest key in the subtree rooted at the right child of  $T$ . We refer to this key as the *successor key* and the node storing this key as the *successor node*. Hereafter in this section, let  $S$  denote the successor node. Deletion of the key stored in  $T$  involves copying the key stored in  $S$  to  $T$  and then removing  $S$  from the tree. To that end, the following edges are locked by setting the lock-flag on the edge using a CAS instruction:  $\langle T, T.right \rangle$ ,  $\langle S.parent, S \rangle$ ,  $\langle S, S.left \rangle$  and  $\langle S, S.right \rangle$ . For an illustration, see Figure 3.6. Note that the first two edges may be same which happens if the successor node is the right child of the target node. Also, since we do not lock the left edge of the target node, the left edge may change and may possibly start pointing to a null address. But, that does not impact the correctness of the complex delete operation.

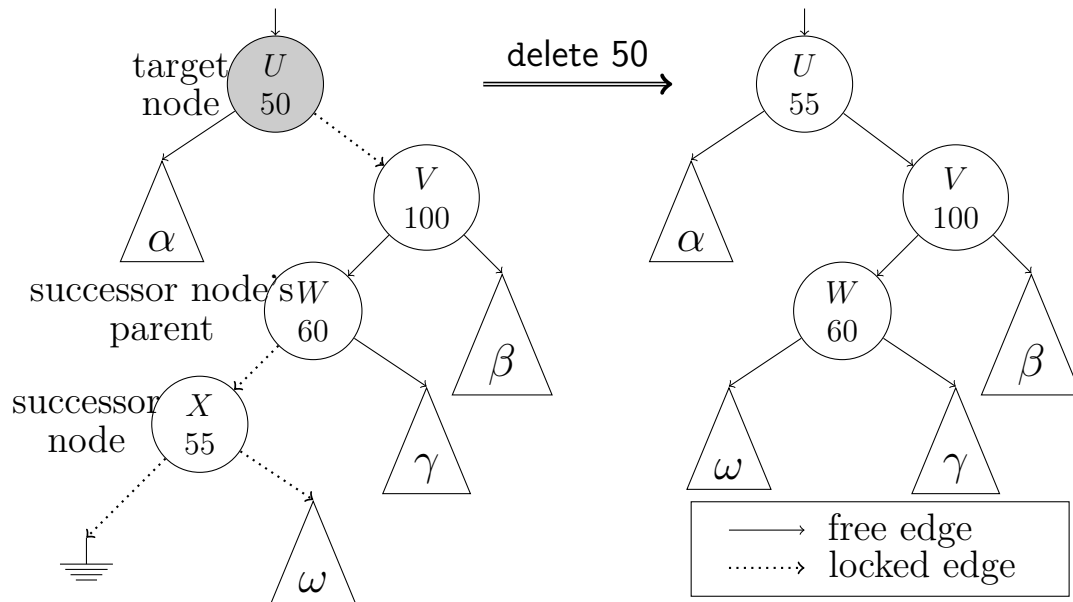


Figure 3.6: An illustration of a complex delete operation

If all the edges are locked successfully, then the operation validates that the key stored in the target node still matches the target key. If the validation succeeds, then the operation copies the key stored in  $S$  to  $T$ , and marks both the children edges of  $S$  to indicate that  $S$  is going to be removed from the tree. Next, it changes the child pointer at  $S.parent$  that is pointing to  $S$  to point to  $S.right$  using a simple write instruction. Finally, the operation releases all the locks and returns **true**.

In both cases (simple as well as complex delete), if the operation fails to obtain any of the locks, then it releases all the locks it was able to acquire up to that point, and restarts from the seek phase. Also, after obtaining all the locks, if the key validation fails, then it implies that some other delete operation has removed the key from the tree while the current execution phase was in progress. In that case, the given delete operation releases all the locks, and simply returns **false**. Note that using a CAS instruction for setting the lock-flag also enables us to *validate that the child pointer has not changed* since it was last observed in a single step.

### 3.1.3 Formal Description

We refer to our algorithm as CASTLE (Concurrent Algorithm for Binary Search Tree by Locking EdgEs).

---

**Algorithm 1:** Data Structures Used

---

```

// a tree node
1 struct Node {
2     Key key;
3     { boolean, boolean, boolean, NodePtr } child[2];
    // each child field contains four subfields: lFlag, mFlag, nFlag and address
4 };

// used to store the results of a tree traversal
5 struct SeekRecord {
6     NodePtr node;
7     NodePtr parent;
8     NodePtr nullAddress;
9 };

// used to store information about an anchor node
10 struct AnchorRecord {
11     NodePtr node;
12     Key key;
13 };

// used to store information about an edge to lock
14 struct LockRecord {
15     NodePtr node;
16     enum { LEFT, RIGHT } which;
17     { boolean, NodePtr } addressSeen;
    // addressSeen contains two subfields: nFlag and address
18 };

// local seek record used when looking for a node
19 SeekRecordPtr seekTargetKey, seekSuccessorKey;
// local array used to store the set of edges to lock
20 LockRecord lockArray[4];

```

---

A pseudo-code of our algorithm is given in Algorithms 1-7. Different data structures used in our algorithm are shown in Algorithm 1. Besides tree node, we use three additional records: (a) *seek record*: to store the outcome of a tree traversal both when looking for the target key and the successor key, (b) *anchor record*: to store information about the anchor node during the seek phase, and (c) *lock record*: to store information about a tree edge that needs to be locked.

---

**Algorithm 2:** Seek Function
 

---

```

21 SEEK( key, seekRecord )
22 begin
23   while true do
24     // initialize the variables used in the traversal
25     pNode :=  $\mathbb{R}$ ;    cNode :=  $\mathbb{S}$ ;
26     address :=  $\mathbb{S} \rightarrow \text{child}[\text{RIGHT}].\text{address}$ ;
27     anchorRecord :=  $\{\mathbb{R}, \infty_1\}$ ;
28     while true do
29       // reached terminal node; read the key stored in the current node
30       cKey := cNode  $\rightarrow$  key;
31       if key = cKey then
32         seekRecord :=  $\{cNode, pNode, address\}$ ;
33         return;
34       which := key < cKey ? LEFT : RIGHT;
35       // read the next address to dereference along with mark and null flags
36        $\langle *, *, nFlag, address \rangle := cNode \rightarrow \text{child}[which]$ ;
37       if nFlag then // the null flag is set; reached terminal node
38         aNode := anchorRecord  $\rightarrow$  node;
39         if aNode  $\rightarrow \text{child}[\text{RIGHT}].mFlag$  then
40           // the anchor node is marked; it may no longer be part of the tree
41           if anchorRecord = pAnchorRecord then
42             // the anchor record of the current traversal matches that of
43             // the previous traversal
44             seekRecord := pSeekRecord;
45             return;
46           else break;
47         else // the anchor node is definitely part of the tree
48           if aNode  $\rightarrow$  key < key then // seek can terminate now
49             seekRecord :=  $\{cNode, pNode, address\}$ ;
50             return;
51           else break;
52       // update the anchor record if needed
53       if which = RIGHT then
54         // the next edge to be traversed is a right edge; keep track of
55         // current node and its key
56         anchorRecord :=  $\{cNode, cKey\}$ ;
57       // traverse the next edge
58       pNode := cNode;    cNode := address;

```

---

---

**Algorithm 3: Search Operation**


---

```

49 boolean SEARCH( key )
50 begin
51   SEEK( key, seekTargetKey );
52   node := seekTargetKey → node;
53   if node → key = key then
54     return true;                                     // key found
55   else
56     return false;                                   // key not found

```

---



---

**Algorithm 4: Insert Operation**


---

```

57 boolean INSERT( key )
58 begin
59   while true do
60     SEEK( key, seekTargetKey );
61     node := seekTargetKey → node;
62     if node → key = key then
63       return false;                                     // key found
64     else
65       // key not found; add the key to the tree
66       newNode := create a new node;
67       // initialize its fields
68       newNode → key := key;
69       newNode → child[LEFT] := ⟨0l, 0m, 1n, null⟩;
70       newNode → child[RIGHT] := ⟨0l, 0m, 1n, null⟩;
71       // determine which child field (left or right) needs to be modified
72       which := key < node → key ? LEFT : RIGHT;
73       // fetch the address observed by the seek function in that field
74       address := seekTargetKey → nullAddress;
75       result := CAS( node → child[which],
76                     ⟨0l, 0m, 1n, address⟩,
77                     ⟨0l, 0m, 0n, newNode⟩ );
78       if result then
79         // new key successfully added to the tree
80         return true;

```

---

---

**Algorithm 5: Delete Operation**


---

```

74 boolean DELETE( key )
75 begin
76   while true do
77     SEEK( key, seekTargetKey );
78     node := seekTargetKey → node;
79     if node → key ≠ key then                                     // key not found
80       return false;
81     else                                                         // key found; read contents of target node's children fields
82       lField := CLEARFLAGS( node → child[LEFT] );
83       rField := CLEARFLAGS( node → child[RIGHT] );
84       if lField.nFlag or rField.nFlag then                       // simple delete operation
85         parent := seekTargetKey → parent;
86         if key < parent → key then which := LEFT;
87         else which := RIGHT;
88         lockArray[0] := {parent, which, ⟨0, node⟩};
89         lockArray[1] := {node, LEFT, lField};
90         lockArray[2] := {node, RIGHT, rField};
91         result := LOCKALL( lockArray, 3 );
92         if result then                                           // all locks acquired; perform the operation
93           if node → key = key then                               // key still matches; remove the node
94             REMOVECHILD( parent, which ); match := true;
95           else match := false;
96           UNLOCKALL( lockArray, 3 );
97           return match;
98       else                                                         // complex delete operation; locate the successor node
99         FINDSMALLEST(node, rField.address, seekSuccessorKey);
100        sNode := seekSuccessorKey → node; sParent := seekSuccessorKey → parent;
101        // determine the edges to be locked
102        lockArray[0] := {node, RIGHT, rField};
103        if node ≠ sParent then
104          // successor node is not the right child of target node
105          lockArray[1] := {sParent, LEFT, ⟨0, sNode⟩}; size := 4;
106        else size := 3 ;
107        lField := CLEARFLAGS( sNode → child[LEFT] );
108        rField := CLEARFLAGS( sNode → child[RIGHT] );
109        lockArray[size - 2] := {sNode, LEFT, lField};
110        lockArray[size - 1] := {sNode, RIGHT, rField};
111        result := LOCKALL( lockArray, size );
112        if result then                                           // all locks acquired; perform the operation
113          if node → key = key then
114            // key still matches; copy key in successor node to target node
115            node → key := sNode → key;
116            REMOVECHILD( sParent, LEFT ); match := true;
117          else match := false;
118          UNLOCKALL( lockArray, size );
119          return match;

```

---

---

**Algorithm 6:** Lock and Unlock Functions

---

```

117 boolean LOCKALL( lockArray, size )
118 begin
119   for  $i \leftarrow 0$  to  $size - 1$  do
120     // acquire lock for the  $i$ -th entry
121      $node := lockArray[i].node$ ;
122      $which := lockArray[i].which$ ;
123      $lockedAddress := lockArray[i].addressSeen$ ;
124      $lockedAddress.lFlag := \mathbf{true}$ ;
125     // set the lock flag in the child edge
126      $result := \mathbf{CAS}(node \rightarrow child[which], lockArray[i].addressSeen, lockedAddress)$ ;
127     if not ( $result$ ) then
128       // release all the locks acquired so far
129       UNLOCKALL( lockArray,  $i - 1$  );
130       return false;
131   return true;
132
133 UNLOCKALL( lockArray, size )
134 begin
135   for  $i \leftarrow size - 1$  to  $0$  do
136      $node := lockArray[i].node$ ;
137      $which := lockArray[i].which$ ;
138     // clear the lock flag in the child edge
139      $node \rightarrow child[which].lFlag := \mathbf{false}$ ;

```

---

The pseudo-code for the seek function is shown in Algorithm 2. The pseudo-codes for search, insert and delete operations are shown in Algorithm 3, Algorithm 4 and Algorithm 5, respectively. Algorithm 6 contains the pseudo-code for locking and unlocking a set of tree edges, as specified in an array. Finally, Algorithm 7 contains the pseudo-codes for three helper functions used by a delete operation, namely: (a) CLEARFLAGS: to clear lock and mark flags from a child field, (b) FINDSMALLEST: to locate the smallest key in a subtree, and (c) REMOVECHILD: to remove a given child of a node.

In the pseudo-code, to improve clarity, we sometimes use subscripts  $l$ ,  $m$  and  $n$  to denote lock, mark and null flags, respectively.



---

**Algorithm 7:** Helper Functions used by Delete Operation

---

```

135 word CLEARFLAGS( word field )
136 begin
137   newField := field with lock and mark flags cleared;
138   return newField;

139 FINDSMALLEST( parent, node, seekRecord )
140 begin
141   // initialize the variables used in the traversal
142   pNode := parent;    cNode := node;
143   while true do
144      $\langle *, *, nFlag, address \rangle := cNode \rightarrow child[LEFT];$ 
145     if not (nFlag) then
146       // visit the next node
147       pNode := cNode;    cNode := address;
148     else
149       // reached the successor node
150       seekRecord := {cNode, pNode, address};
151       break;

149 REMOVECHILD( parent, which )
150 begin
151   // determine the address of the child to be removed
152   node := parent  $\rightarrow$  child[which];
153   // mark both the children edges of the node to be removed
154   node  $\rightarrow$  child[LEFT].mFlag := true;
155   node  $\rightarrow$  child[RIGHT].mFlag := true;
156   // determine whether both the child pointers of the node to be removed are null
157   if node  $\rightarrow$  child[LEFT].nFlag and node  $\rightarrow$  child[RIGHT].nFlag then
158     // set the null flag only
159     parent  $\rightarrow$  child[which].nFlag := true;
160   else
161     // switch the pointer at the parent to point to its appropriate grandchild
162     if node  $\rightarrow$  child[LEFT].nFlag then
163       address := node  $\rightarrow$  child[RIGHT].address;
164     else address := node  $\rightarrow$  child[LEFT].address;
165     parent  $\rightarrow$  child[which].address := address;

```

---

### 3.1.4 Correctness Proof

It is convenient to treat insert and delete operations that do not change the tree as search operations. We call a tree node *active* if it is reachable from the root of the tree. We call a tree node *passive* if it was active earlier but is not active any more. Note that, before an active node is made passive by a delete operation, both its children edges are *marked*. Also, a CAS instruction performed on an edge (by either an insert operation or a delete operation as part of locking) is successful only if the edge is unmarked. As a result, clearly, if an insert operation completes successfully, then its target node was active when its edge was modified to make the new node (containing the target key) a part of the tree. Likewise, if a delete operation completes successfully, then all the nodes involved in the operation (up to three nodes) were active when their edges were locked.

### All Executions are Linearizable

We show that an arbitrary execution of our algorithm is linearizable by specifying the *linearization point* of each operation. Note that the linearization point of an operation is the point during its execution at which the operation appeared to have taken effect. Our algorithm supports three types of operations: search, insert and delete. We now specify the linearization point of each operation.

1. *Insert operation:* The operation is linearized at the point at which it performed the successful CAS instruction that resulted in its target key becoming part of the tree.
2. *Delete operation:* There are two cases depending on whether the delete operation is simple or complex. If the operation is simple delete, then the operation is linearized at the point at which a successful write step was performed at the parent of the target node that resulted in the target node becoming passive. Otherwise, it is linearized at the point at which the original key of the target node was replaced with its successor key.

3. *Search operation:* There are two cases depending on whether the target node was active when the operation read the key stored in the node. If the target node was not active, then the operation is linearized at the point at which the target node became passive. Otherwise, it is linearized at the point at which the read step was performed.

It can be easily verified that, for any execution of the algorithm, the sequence of operations obtained by ordering operations based on their linearization points is legal, *i.e.*, all operations in the sequence satisfy their specification.

Thus we have:

**Theorem 1.** *Every execution of our algorithm is linearizable.*

### All Executions are Deadlock-Free

We say that the system is in a *quiescent state* if no modify operation completes hereafter. We say that the system is in a *potent state* if it has one or more pending modify operations. Note that quiescence is a *stable property*; once the system is in a quiescent state, it stays in a quiescent state. We show that our algorithm is deadlock-free by proving that a potent state is necessarily non-quiescent.

Note that, in a quiescent state, no edges in the tree can be marked. This is because a delete operation marks edges only after it has successfully obtained all the locks, after which it is guaranteed to complete. This also implies that the tree cannot undergo any changes now because that would imply eventual completion of a modify operation. Thus, once a system has reached a quiescent state, all modify operation currently pending repeatedly alternate between seek and execution phases. We say that the system is in a *strongly-quiescent state* if all pending modify operations started their most recent seek phase *after* the system became quiescent. Note that, like quiescence, strong quiescence is also a stable property. Now, once the system has reached a strongly quiescent state, the following can be easily verified. First,

for a given modify operation, every traversal of the tree in the seek phase returns the same target node. Second, for a given delete operation, the set of edges it needs to lock remains the same.

Now, assume that the system eventually reaches a state that is both potent and quiescent. Clearly, from this state, the system will eventually reach a state that is potent and strongly-quiescent. Note that a delete operation in our algorithm locks edges in a *top-down, left-right* manner. As a result, there cannot be a “cycle” involving delete operations. If a delete operation continues to fail in the execution phase, then it is necessarily because it tried to acquire lock on an already locked edge. (Recall that the set of edges does not change any more and there are no marked edges in the tree.) We can construct a chain of operations such that each operation in the chain tried to lock an edge already locked by the next operation in the chain. Clearly, the length of the chain is bounded. This implies that the last operation in the chain is guaranteed to obtain all the locks and will eventually complete. This contradicts the fact that the system is in a quiescent state.

Thus, we have:

**Theorem 2.** *Every execution of our algorithm is deadlock-free.*

## CHAPTER 4

### LOCK FREE CONCURRENT BINARY SEARCH TREE

#### 4.1 The Lock-Free Algorithm

For ease of exposition, we describe our algorithm assuming no memory reclamation.

##### 4.1.1 Overview of the Algorithm

Every operation in our algorithm uses *seek* function as a subroutine. The seek function traverses the tree from the root node until it either finds the target key or reaches a non-binary node whose next edge to be followed points to a null node. We refer to the path traversed by the operation during the seek as the *access-path*, and the last node in the access-path as the *terminal node*. The operation then compares the target key with the stored key (the key present in the terminal node). Depending on the result of the comparison and the type of the operation, the operation either terminates or moves to the execution phase. In certain cases in which a key may have moved upward along the access-path, the seek function may have to restart and traverse the tree again; details about restarting are provided later. We now describe the next steps for each of the operations one-by-one.

**Search** A search operation starts by invoking seek operation. It returns **true** if the stored key matches the target key and **false** otherwise.

**Insert** An insert operation ((shown in Figure 4.1)) starts by invoking seek operation. It returns **false** if the target key matches the stored key; otherwise, it moves to the execution phase. In the execution phase, it attempts to insert the key into the tree as a child node of

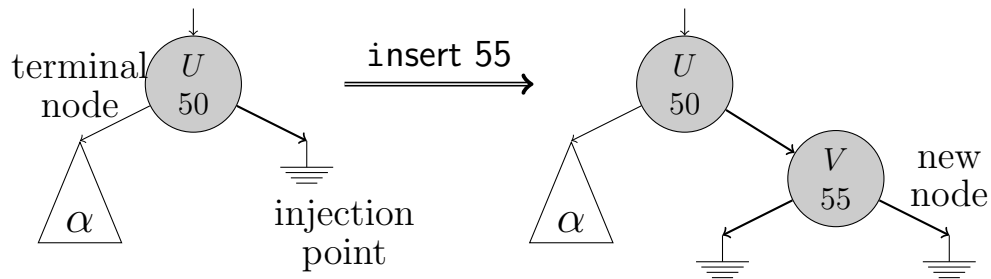


Figure 4.1: An illustration of an insert operation.

the last node in the access-path using a CAS instruction. If the instruction succeeds, then the operation returns **true**; otherwise, it restarts by invoking the seek function again.

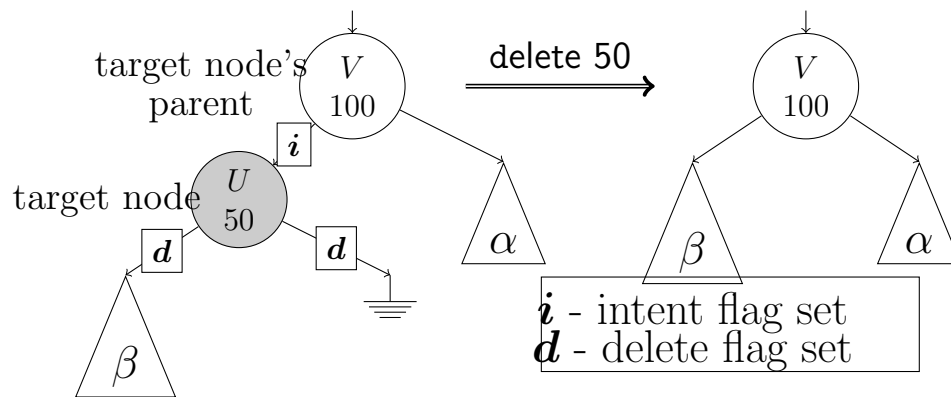


Figure 4.2: An illustration of a simple delete operation.

**Delete** A delete operation starts by invoking seek function. It returns **false** if the stored key does not match the target key; otherwise, it moves to the execution phase. In the execution phase, it attempts to remove the key stored in the terminal node of the access-path. There are two cases depending on whether the terminal node is a binary node (has two children) or not (has at most one child). In the first case, the operation is referred to as *complex delete operation*. In the second case, it is referred to as *simple delete operation*. In the case of simple delete (shown in Figure 4.2), the terminal node is removed by changing the pointer at the parent node of the terminal node. In the case of complex delete (shown in Figure 4.3),

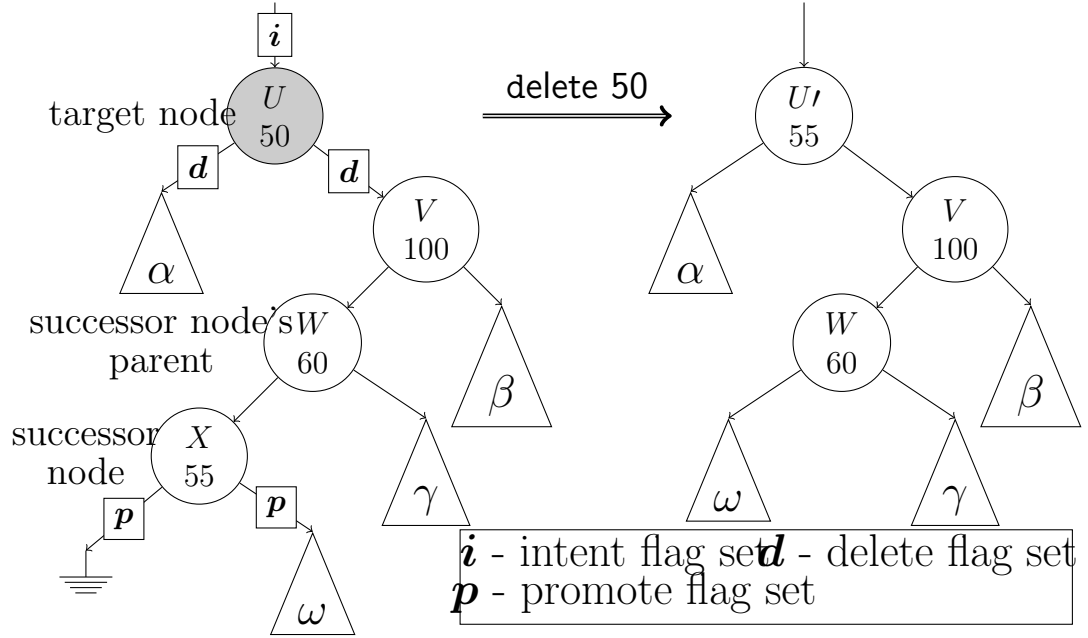


Figure 4.3: An illustration of a complex delete operation.

the key to be deleted is replaced with the *next largest* key in the tree, which will be stored in the *leftmost node* of the *right subtree* of the terminal node.

#### 4.1.2 Details of the Algorithm

As in most algorithms, we use sentinel keys and three sentinel nodes to handle the boundary cases easily. The structure of an empty tree with only sentinel keys (denoted by  $\infty_0$ ,  $\infty_1$  and  $\infty_2$  with  $\infty_0 < \infty_1 < \infty_2$ ) and sentinel nodes (denoted by  $\mathbb{R}$ ,  $\mathbb{S}$  and  $\mathbb{T}$ ) is shown in Figure 4.4.

Our algorithm, like the one in (Natarajan and Mittal, 2014), operates at edge level. A delete operation obtains ownership of the edges it needs to work on by marking them. To enable marking, we steal bits from the child addresses of a node. Specifically, we steal *three* bits from each child address to distinguish between three types of marking: (i) marking for *intent*, (ii) marking for *deletion* and (iii) marking for *promotion*. The three bits are referred to as *intent-flag*, *delete-flag* and *promote-flag*. To avoid the ABA problem, as in Howley

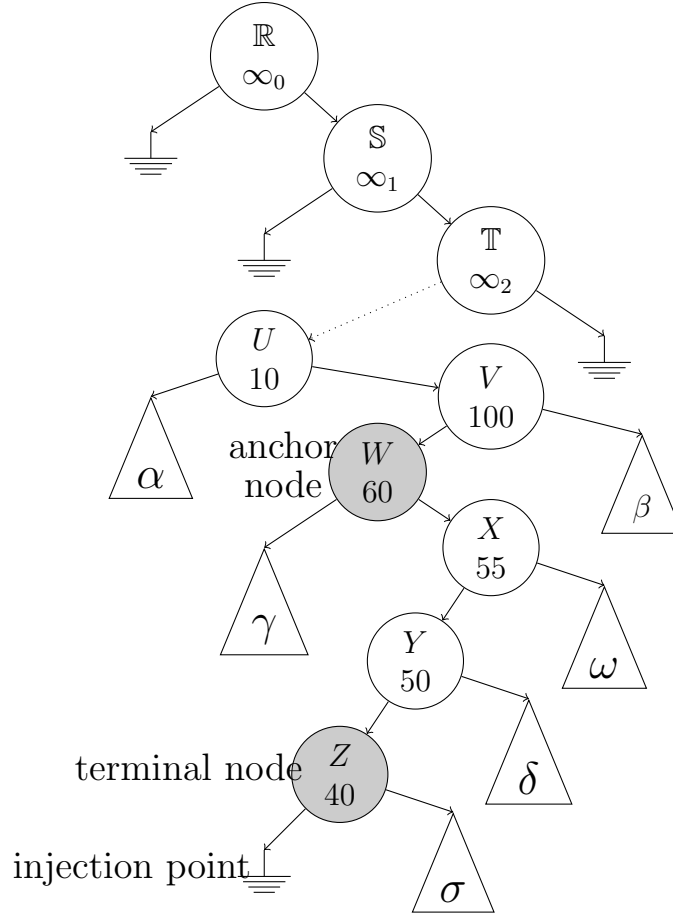


Figure 4.4: Nodes in the access path of seek along with sentinel keys and nodes ( $\infty_0 < \infty_1 < \infty_2$ )

and Jones (Howley and Jones, 2012), we use *unique* null pointers. To that end, we steal another bit from the child address, referred to as *null-flag*, and use it to indicate whether the address field contains a null or a non-null value. So, when an address changes from a non-null value to a null value, we only set the null-flag and the contents of the address field are not otherwise modified. This ensures that all null pointers are unique.

Finally, we also steal a bit from the key field to indicate whether the key stored in a node is the original key or the replacement key. This information is used in a complex delete operation to coordinate helping among processes.



We next describe the details of the seek function, which is used by all operations (search as well as modify) to traverse the tree after which we describe the details of the execution phase of insert and delete operations.

### The Seek Phase

A seek function keeps track of the node in the access-path at which it took the last “right turn” (*i.e.*, it last followed a right edge). Let this “right turn” node be referred to as *anchor node* when the traversal reaches the terminal node. Note that the terminal node is the node whose key matched the target key or whose next child edge is set to a null address. For an illustration, please see Figure 4.4. In the latter case (stored key does not match the target key), it is possible that the key may have moved up in the tree. To ascertain that the seek function did not miss the key because it may have moved up during the traversal, we use the following set of conditions that are *sufficient* (but not necessary) to guarantee that the seek function did not miss the key. First, the anchor node is still part of the tree. Second, the key stored in the anchor node has not changed since it first encountered the anchor node during the (current) traversal. To check for the above two conditions, we determine whether the anchor node is undergoing removal (either delete or promote flag set) by examining its right child edge. We discuss the two cases one-by-one.

- (a) *Right child edge not marked:* In this case, the anchor node is still part of the tree. We next check whether the key stored in the anchor node has changed. If the key has not changed, then the seek function returns the results of the (current) traversal, which consists of three addresses: (i) the address of the terminal node, (ii) the address of its parent, and (iii) the null address stored in the child field of the terminal node that caused the traversal to terminate. The last address is required to ensure that an insert operation works correctly (specifically to ascertain that the child field of the terminal node has not undergone any change since the completion of the traversal). We refer

to it as the *injection point* of the insert operation. On the other hand, if the key has changed, then the seek function restarts from the root of the tree.

- (b) *Right child edge marked:* In this case, we compare the information gathered in the current traversal about the anchor node with that in the previous traversal, if one exists. Specifically, if the anchor node of the previous traversal is same as that of the current traversal and the keys found in the anchor node in the two traversals also match, then the seek function terminates, but returns the results of the previous traversal (instead of that of the current traversal). This is because the anchor node was definitely part of the tree during the previous traversal since it was reachable from the root of the tree at the beginning of the current traversal. Otherwise, the seek function restarts from the root of the tree.

The seek function also keeps track of the *second-to-last* edge in the access-path (whose endpoints are the parent and grandparent nodes of the terminal node), which is used for helping, if there is a conflict. For insert and delete operations, we refer to the terminal node as the *target node*.

### The Execution Phase of an Insert Operation

In the execution phase, an insert operation creates a new node containing the target key. It then adds the new node to the tree at the injection point using a CAS instruction. If the CAS instruction succeeds, then (the new node becomes a part of the tree and) the operation terminates; otherwise, the operation determines if it failed because of a *conflicting* delete operation in progress. If there is no conflicting delete operation in progress, then the operation restarts from the seek phase; otherwise it performs helping and then restarts from the seek phase. Figure 4.5 shows a flow chart describing the sequence of steps of an insert operation.

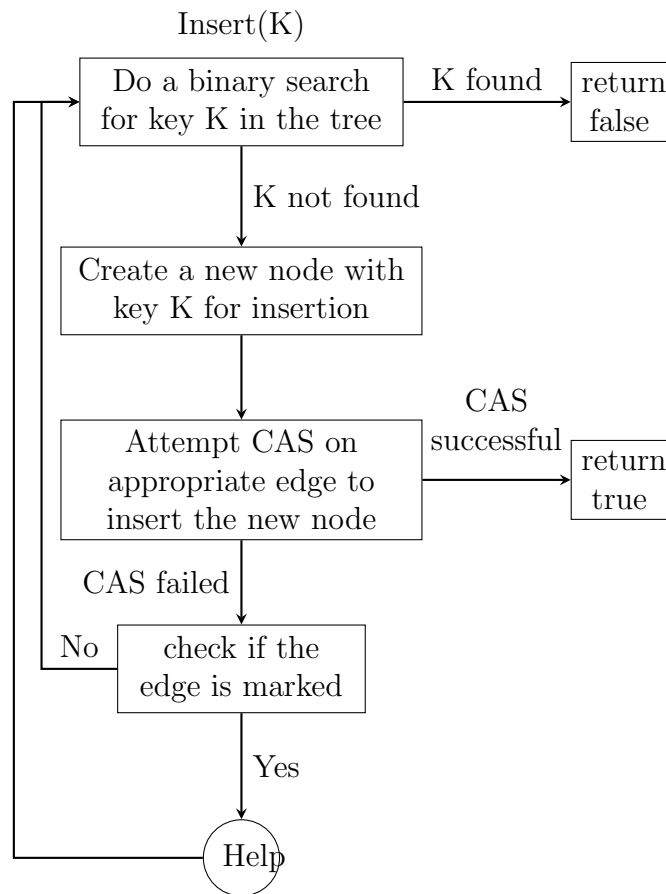


Figure 4.5: ELFTREE - Sequence of steps in an insert operation

### The Execution Phase of a Delete Operation

The execution of a delete operation starts in *injection mode*. Once the operation has been injected into the tree, it advances to either *discovery mode* or *cleanup mode* depending on the type of the delete operation.

**Injection Mode** In the injection mode, the delete operation marks the three edges involving the target node as follows: (i) It first sets the intent-flag on the edge from the parent of the target node to the target node using a CAS instruction. (ii) It then sets the delete-flag on the left edge of the target node using a CAS instruction. (iii) Finally, it sets the delete-flag on the right edge of the target node using a CAS instruction. If the CAS instruction fails at

any step, the delete operation performs helping, and either repeats the same step or restarts from the seek phase. Specifically, the delete operation repeats the same step when setting the delete-flag as long as the target node has not been claimed as the successor node by another delete operation. In all other cases, it restarts from the seek phase.

We maintain the invariant that an edge, once marked, cannot be unmarked. After marking both the edges of the target node, the operation checks whether the target node is a binary node or not. If it is a binary node, then the delete operation is classified as complex; otherwise it is classified as simple. Note that the type of the delete operation cannot change once all the three edges have been marked as described above. If the delete operation is complex, then it advances to the discovery mode after which it will advance to the cleanup mode. On the other hand, if it is simple, then it directly advances to the cleanup mode (and skips the discovery mode). Eventually, the target node is either removed from the tree (if simple delete) or replaced with a “new” node containing the next largest key (if complex delete).

For a tree node  $X$ , let  $X.parent$  denote its parent node, and  $X.left$  and  $X.right$  denote its left and right child node, respectively. Also, hereafter in this section, let  $T$  denote the target node of the delete operation under consideration.

**Discovery Mode** In the discovery mode, a complex delete operation performs the following steps:

1. **Find Successor Key:** The operation locates the next largest key in the tree, which is the smallest key in the subtree rooted at the right child of  $T$ . We refer to this key as the *successor key* and the node storing this key as the *successor node*. Hereafter in this section, let  $S$  denote the successor node.
2. **Mark Child Edges of Successor Node:** The operation sets the promote-flag on both the child edges of  $S$  using a CAS instruction. Note that the left child edge of  $S$  will be

null. As part of marking the left child edge, we also store the address of  $T$  (the target node) in the edge. This is done to enable helping in case the successor node is obstructing the progress of another operation. In case the CAS instruction fails while marking the left child edge, the operation repeats from step 1 after performing helping if needed. On the other hand, if the CAS instruction fails while marking the right child edge, then the marking step is repeated after performing helping if needed.

3. **Promote Successor Key:** The operation replaces the target node's original key with the successor key. At the same time, it also sets the mark bit in the key to indicate that the current key stored in the target node is the replacement key and not the original key.
4. **Remove Successor Node:** The operation removes  $S$  (the successor node) by changing the child pointer at  $S.parent$  that is pointing to  $S$  to point to the right child of  $S$  using a CAS instruction. If the CAS instruction succeeds, then the operation advances to the cleanup mode. Otherwise, it performs helping if needed. It then finds  $S$  again by performing another traversal of the tree starting from the right child of  $T$ . If the traversal fails to find  $S$  (recall that the left edge of  $S$  is marked for promotion and contains the address of  $T$ ), then  $S$  has already been removed from the tree by another operation as part of helping, and the delete operation advances to the cleanup mode. On advancing to the cleanup mode, the operation sets a flag in  $T$  indicating that  $S$  has been removed from the tree (and  $T$  can now be replaced with a new node) so that other operations trying to help it know not to look for  $S$ .

Figure 4.6 shows a flow chart describing the sequence of steps of a delete operation.

**Cleanup Mode** There are two cases depending on whether the delete operation is simple or complex.

- (a) **Simple Delete:** In this case, either  $T.left$  or  $T.right$  is pointing to a null node. Note that both  $T.left$  and  $T.right$  may be pointing to null nodes (which in turn will imply

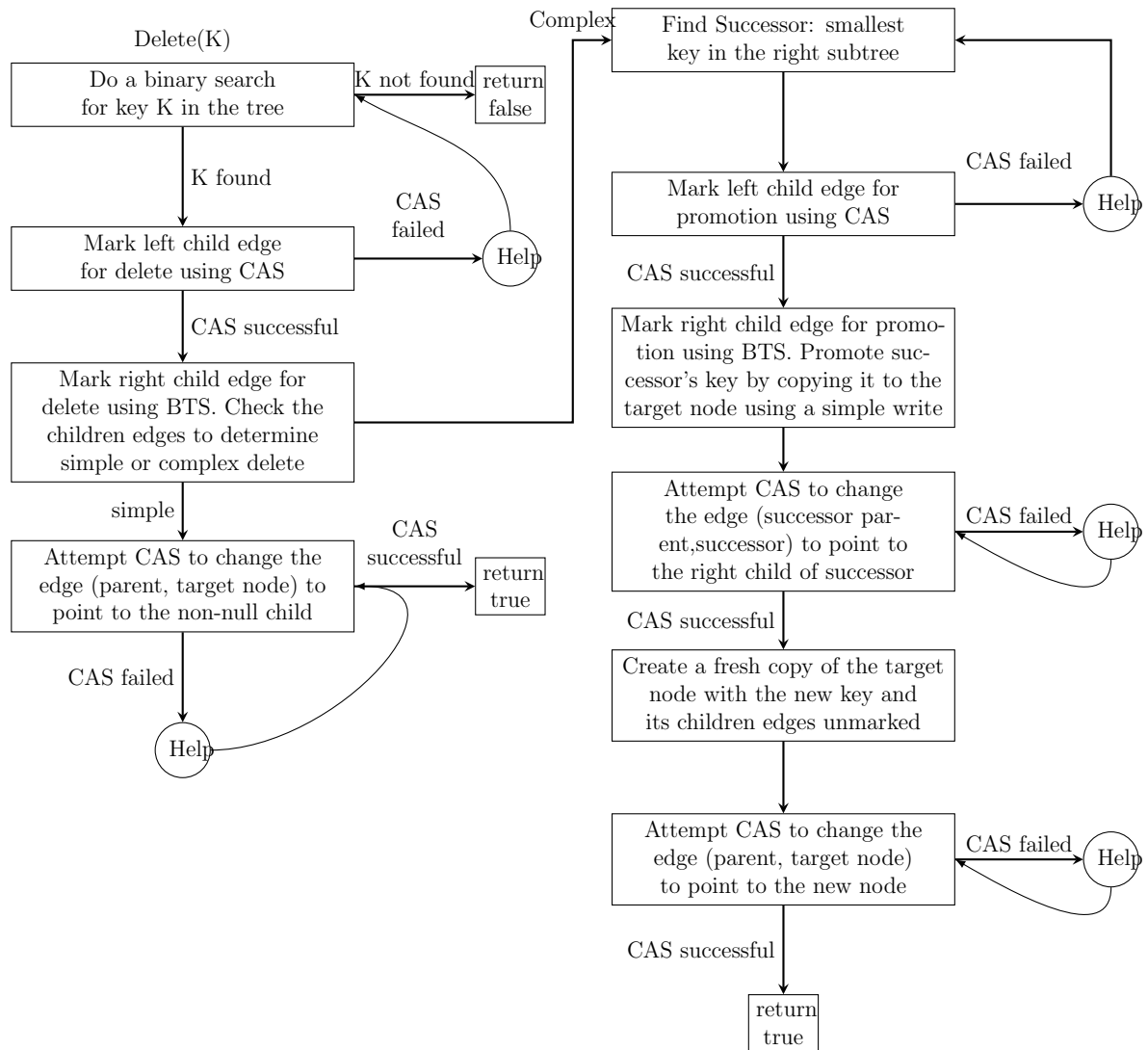


Figure 4.6: ELFTREE - Sequence of steps in a delete operation

---

**Algorithm 8:** Data Structures Used

---

```

161 struct Node {
162     {Boolean, Key} mKey;
163     {Boolean, Boolean, Boolean, Boolean, NodePtr} child[2];
164     Boolean readyToReplace;
165 };
166 struct Edge {
167     NodePtr parent, child;
168     enum which { LEFT, RIGHT };
169 };
170 struct SeekRecord {
171     Edge lastEdge, pLastEdge, injectionEdge;
172 };
173 struct AnchorRecord {
174     NodePtr node;
175     Key key;
176 };
177 struct StateRecord {
178     Edge targetEdge, pTargetEdge;
179     Key targetKey, currentKey;
180     enum mode { INJECTION, DISCOVERY, CLEANUP };
181     enum type { SIMPLE, COMPLEX };
182     // the next field stores pointer to a seek record; it is used for finding the
183     // successor if the delete operation is complex
184     SeekRecordPtr successorRecord;
185 };
186 // object to store information about the tree traversal when looking for a given key
187 // (used by the seek function)
188 SeekRecordPtr targetRecord := new seek record;
189 // object to store information about process' own delete operation
190 StateRecordPtr myState := new state;

```

---

that  $T$  is a leaf node). Without loss of generality, assume that  $T.right$  is a null node. The removal of  $T$  involves changing the child pointer at  $T.parent$  that is pointing to  $T$  to point to  $T.left$  using a CAS instruction. If the CAS instruction succeeds, then the delete operation terminates; otherwise, it performs another seek on the tree. If the seek function either fails to find the target key or returns a terminal node different from  $T$ , then  $T$  has been already removed from the tree (by another operation as part of helping) and the delete operation terminates; otherwise, it attempts to remove  $T$  from

---

**Algorithm 9: Seek Function**


---

```

186 SEEK( key, seekRecord )
187 begin
188   pAnchorRecord := { $\mathbb{S}$ ,  $\infty_1$ };
189   while true do
190     // initialize all variables used in traversal
191     pLastEdge := { $\mathbb{R}$ ,  $\mathbb{S}$ , RIGHT};    lastEdge := { $\mathbb{S}$ ,  $\mathbb{T}$ , RIGHT};
192     curr :=  $\mathbb{T}$ ;    anchorRecord := { $\mathbb{S}$ ,  $\infty_1$ };
193     while true do
194       // read the key stored in the current node
195        $\langle *, cKey \rangle := curr \rightarrow mKey$ ;
196       // find the next edge to follow
197       which := key < cKey ? LEFT: RIGHT;
198        $\langle n, *, d, p, next \rangle := curr \rightarrow child[which]$ ;
199       // check for the completion of the traversal
200       if key = cKey or n then
201         // either key found or no next edge to follow; stop the traversal
202         seekRecord  $\rightarrow$  pLastEdge := pLastEdge;
203         seekRecord  $\rightarrow$  lastEdge := lastEdge;
204         seekRecord  $\rightarrow$  injectionEdge := {curr, next, which};
205         if key = cKey then // keys match
206           return;
207         else break;
208       if which = RIGHT then
209         // next edge to be traversed is a right edge; keep track of the
210         // current node and its key
211         anchorRecord :=  $\langle curr, cKey \rangle$ ;
212         // traverse the next edge
213         pLastEdge := lastEdge;    lastEdge := {curr, next, which};    curr := next;
214       // key was not found; check if can stop
215        $\langle *, *, d, p, * \rangle := anchorRecord.node \rightarrow child[RIGHT]$ ;
216       if not (d) and not (p) then
217         // anchor node is still part of the tree; check if anchor node's key has
218         // changed
219          $\langle *, aKey \rangle := anchorRecord.node \rightarrow mKey$ ;
220         if anchorRecord.key = aKey then return;
221       else
222         // check if the anchor record (the node and its key) matches that of the
223         // previous traversal
224         if pAnchorRecord = anchorRecord then
225           // return the results of the previous traversal
226           seekRecord := pSeekRecord;
227           return;
228       // store the results of the traversal and restart
229       pSeekRecord := seekRecord;    pAnchorRecord := anchorRecord;

```

---



---

**Algorithm 10: Search Operation**


---

```

215 Boolean SEARCH( key )
216 begin
217   SEEK( key, mySeekRecord );
218   node := mySeekRecord → lastEdge.child;
219    $\langle *, nKey \rangle := node \rightarrow mKey$ ;
220   if nKey = key then return true;
221   else return false;

```

---



---

**Algorithm 11: Insert Operation**


---

```

222 Boolean INSERT( key )
223 begin
224   while true do
225     SEEK( key, targetRecord );
226     targetEdge := targetRecord → lastEdge;
227     node := targetEdge.child;
228      $\langle *, nKey \rangle := node \rightarrow mKey$ ;
229     if key = nKey then return false;

    // create a new node and initialize its fields
230     newNode := create a new node;
231     newNode → mKey :=  $\langle 0_m, key \rangle$ ;
232     newNode → child[LEFT] :=  $\langle 1_n, 0_i, 0_d, 0_p, \text{null} \rangle$ ;
233     newNode → child[RIGHT] :=  $\langle 1_n, 0_i, 0_d, 0_p, \text{null} \rangle$ ;
234     newNode → readyToReplace := false;

235     which := targetRecord → injectionEdge.which;
236     address := targetRecord → injectionEdge.child;
237     result := CAS(node → child[which],  $\langle 1_n, 0_i, 0_d, 0_p, address \rangle$ ,  $\langle 0_n, 0_i, 0_d, 0_p, newNode \rangle$ );
238     if result then return true;

    // help if needed
239      $\langle *, *, d, p, * \rangle := node \rightarrow child[which]$ ;
240     if d then HELPTARGETNODE( targetEdge );
241     else if p then HELPSUCCESSORNODE( targetEdge );

```

---

---

**Algorithm 12: Delete Operation**


---

```

242 Boolean DELETE( key )
243 begin
    // initialize the state record
244   myState → targetKey := key;      myState → currentKey := key;
245   myState → mode := INJECTION;
246   while true do
247     SEEK( myState → currentKey, targetRecord );
248     targetEdge := targetRecord → lastEdge;      pTargetEdge := targetRecord → pLastEdge;
249      $\langle *, nKey \rangle := targetEdge.child \rightarrow mKey$ ;
250     if myState → currentKey ≠ nKey then
251       // the key does not exist in the tree
252       if myState → mode = INJECTION then return false;
253       else return true;
254       // perform appropriate action depending on the mode
255       if myState → mode = INJECTION then
256         // store a reference to the target edge
257         myState → targetEdge := targetEdge;
258         myState → pTargetEdge := pTargetEdge;
259         // attempt to inject the operation at the node
260         INJECT( myState );
261         // mode would have changed if injection was successful
262         if myState → mode ≠ INJECTION then
263           // check if the target node found by the seek function matches the one
264           // stored in the state record
265           if myState → targetEdge.child ≠ targetEdge.child then return true;
266           // update the target edge information using the most recent seek
267           myState → targetEdge := targetEdge;
268           if myState → mode = DISCOVERY then
269             // complex delete operation; locate the successor node and mark its child
270             // edges with promote flag
271             FINDANDMARKSUCCESSOR( myState );
272             if myState → mode = DISCOVERY then
273               // complex delete operation; promote the successor node's key and remove
274               // the successor node
275               REMOVESUCCESSOR( myState );
276             if myState → mode = CLEANUP then
277               // either remove the target node (simple delete) or replace it with a new
278               // node with all fields unmarked (complex delete)
279               result := CLEANUP( myState );
280               if result then return true;
281               else
282                  $\langle *, nKey \rangle := targetEdge.child \rightarrow mKey$ ;
283                 myState → currentKey := nKey;

```

---

---

**Algorithm 13:** Injecting a Deletion Operation

---

```

270 INJECT( state )
271 begin
272   targetEdge := state → targetEdge;
      // try to set the intent flag on the target edge
      // retrieve attributes of the target edge
273   parent := targetEdge.parent;
274   node := targetEdge.child;
275   which := targetEdge.which;
276   result := CAS( parent → child[which],
                  ⟨0n, 0i, 0d, 0p, node⟩,
                  ⟨0n, 1i, 0d, 0p, node⟩ );
277   if not (result) then
      // unable to set the intent flag; help if needed
278     ⟨*, i, d, p, address⟩ := parent → child[which];
279     if i then HELPTARGETNODE( targetEdge );
280     else if d then
281       | HELPTARGETNODE( state → pTargetEdge );
282     else if p then
283       | HELPSUCCESSORNODE( state → pTargetEdge );
284   return;

      // mark the left edge for deletion
285   result := MARKCHILDEDGE( state, LEFT );
286   if not (result) then return;
      // mark the right edge for deletion; cannot fail
287   MARKCHILDEDGE( state, RIGHT );

      // initialize the type and mode of the operation
288   INITIALIZETYPEANDUPDATEMODE( state );

```

---

the tree again using possibly the new parent information returned by seek. This process may be repeated multiple times.

- (b) **Complex Delete:** Note that, at this point, the key stored in the target node is the replacement key (the successor key of the target key). Further, the key as well as both the child edges of the target node are marked. The delete operation attempts to replace target node with a *new* node, which is basically a copy of target node except that all its fields are unmarked. This replacement of  $T$  involves changing the child pointer at  $T.parent$  that is pointing to  $T$  to point to the new node. If the CAS instruction succeeds, then the delete operation terminates; otherwise, as in the case of simple delete,

---

**Algorithm 14: Locating the Successor Node**


---

```

289 FINDANDMARKSUCCESSOR( state )
290 begin
    // retrieve the addresses from the state record
291 node := state → targetEdge.child;
292 seekRecord := state → successorRecord;
293 while true do
    // read the mark flag of the key in the target node
294  $\langle m, * \rangle := \text{node} \rightarrow mKey$ ;
    // find the node with the smallest key in the right subtree
295 result := FINDSMALLEST( state );
296 if m or not (result) then
    // successor node had already been selected before the traversal or the
    // right subtree is empty
297     break;
    // retrieve the information from the seek record
298 successorEdge := seekRecord → lastEdge;
299 left := seekRecord → injectionEdge.child;
    // read the mark flag of the key under deletion
300  $\langle m, * \rangle := \text{node} \rightarrow mKey$ ;
301 if m then // successor node has already been selected
302     continue;
    // try to set the promote flag on the left edge
303 result := CAS( successorEdge.child →
                  child[LEFT],
                   $\langle 1_n, 0_i, 0_d, 0_p, \text{left} \rangle$ ,
                   $\langle 1_n, 0_i, 0_d, 1_p, \text{node} \rangle$  );
304 if result then break;
    // attempt to mark the edge failed; recover from the failure and retry if
    // needed
305  $\langle n, *, d, *, * \rangle := \text{successorEdge.child} \rightarrow \text{child}[\text{LEFT}]$ ;
306 if n and d then
    // the node found is undergoing deletion; need to help
307     HELPTARGETNODE( successorEdge );

    // update the operation mode
308 UPDATEMODE( state );

```

---

---

**Algorithm 15:** Removing the Successor Node

---

```

309 REMOVESUCCESSOR( state )
310 begin
    // retrieve addresses from the state record
311   node := state → targetEdge.child;
312   seekRecord := state → successorRecord;
    // extract information about the successor node
313   successorEdge := seekRecord → lastEdge;

    // ascertain that seek record for successor node contains valid information
314    $\langle *, *, *, p, address \rangle := successorEdge.child \rightarrow child[LEFT]$ ;
315   if not (p) or (address  $\neq$  node) then
316       node → readyToReplace := true;
317       UPDATEMODE( state );
318       return;

    // mark the right edge for promotion if unmarked
319   MARKCHILDEDGE( state, RIGHT );

    // promote the key
320   node → mKey :=  $\langle 1_m, successorEdge.child \rightarrow mKey \rangle$ ;
321   while true do
        // check if the successor is the right child of the target node itself
322       if successorEdge.parent = node then
            // need to modify the right edge of target node whose delete flag is set
323             dFlag := 1;    which := RIGHT;
324       else
            dFlag := 0;    which := LEFT;
325        $\langle *, i, *, *, * \rangle := successorEdge.parent \rightarrow child[which]$ ;
326        $\langle n, *, *, *, right \rangle := successorEdge.child \rightarrow child[RIGHT]$ ;
327       oldValue :=  $\langle 0_n, i, dFlag, 0_p, successorEdge.child \rangle$ ;
328       if n then // only set the null flag; do not change the address
329           newValue :=  $\langle 1_n, 0_i, dFlag, 0_p, successorEdge.child \rangle$ ;
330       else // switch the pointer to point to the grand child
331           newValue :=  $\langle 0_n, 0_i, dFlag, 0_p, right \rangle$ ;
332       result := CAS(successorEdge.parent → child[which], oldValue, newValue);
333       if result or dFlag then break;
334        $\langle *, *, d, *, * \rangle := successorEdge.parent \rightarrow child[which]$ ;
335       pLastEdge := seekRecord → pLastEdge;
336       if d and (pLastEdge.parent  $\neq$  null) then
337           HELPTARGETNODE( pLastEdge );

338       result := FINDSMALLEST( state );
339       lastEdge := seekRecord → lastEdge;
340       if not (result) or lastEdge.child  $\neq$  successorEdge.child then
341           break; // the successor node has already been removed
342       else successorEdge := seekRecord → lastEdge ;

343   node → readyToReplace := true;
344   UPDATEMODE( state );

```

---

---

**Algorithm 16:** Cleaning Up the Tree

---

```

346 Boolean CLEANUP( state )
347 begin
348    $\langle \text{parent}, \text{node}, pWhich \rangle := \text{state} \rightarrow \text{targetEdge};$ 
349   if  $\text{state} \rightarrow \text{type} = \text{COMPLEX}$  then
350     // replace the node with a new copy in which all fields are unmarked
351      $\langle *, nKey \rangle := \text{node} \rightarrow mKey;$ 
352      $\text{newNode} \rightarrow mKey := \langle 0_m, nKey \rangle;$ 
353     // initialize left and right child pointers
354      $\langle *, *, *, *, left \rangle := \text{node} \rightarrow \text{child}[\text{LEFT}];$ 
355      $\text{newNode} \rightarrow \text{child}[\text{LEFT}] := \langle 0_n, 0_i, 0_d, 0_p, left \rangle;$ 
356      $\langle n, *, *, *, right \rangle := \text{node} \rightarrow \text{child}[\text{RIGHT}];$ 
357     if  $n$  then
358        $\text{newNode} \rightarrow \text{child}[\text{RIGHT}] := \langle 1_n, 0_i, 0_d, 0_p, \text{null} \rangle;$ 
359     else  $\text{newNode} \rightarrow \text{child}[\text{RIGHT}] := \langle 0_n, 0_i, 0_d, 0_p, right \rangle ;$ 
360     // initialize the arguments of CAS instruction
361      $\text{oldValue} := \langle 0_n, 1_i, 0_d, 0_p, \text{node} \rangle;$ 
362      $\text{newValue} := \langle 0_n, 0_i, 0_d, 0_p, \text{newNode} \rangle;$ 
363   else // remove the node
364     // determine to which grand child will the edge at the parent be switched
365     if  $\text{node} \rightarrow \text{child}[\text{LEFT}] = \langle 1_n, *, *, *, * \rangle$  then
366        $nWhich := \text{RIGHT};$ 
367     else  $nWhich := \text{LEFT};$ 
368     // initialize the arguments of the CAS instruction
369      $\text{oldValue} := \langle 0_n, 1_i, 0_d, 0_p, \text{node} \rangle;$ 
370      $\langle n, *, *, *, address \rangle := \text{node} \rightarrow \text{child}[nWhich];$ 
371     if  $n$  then // set the null flag only
372        $\text{newValue} := \langle 1_n, 0_i, 0_d, 0_p, \text{node} \rangle;$ 
373     else // change the pointer to the grand child
374        $\text{newValue} := \langle 0_n, 0_i, 0_d, 0_p, address \rangle ;$ 
375    $\text{result} := \text{CAS}( \text{parent} \rightarrow \text{child}[pWhich],$ 
376                      $\text{oldValue}, \text{newValue} );$ 
377   return  $\text{result};$ 

```

---

---

**Algorithm 17:** Mark Child Edge
 

---

```

372 Boolean MARKCHILDEDGE( state, which )
373 begin
374   if state → mode = INJECTION then
375     edge := state → targetEdge;
376     flag := DELETE_FLAG;
377   else
378     edge := (state → successorRecord) → lastEdge;
379     flag := PROMOTE_FLAG;
380   node := edge.child;
381   while true do
382      $\langle n, i, d, p, address \rangle := node \rightarrow child[which]$ ;
383     if i then
384       helpedEdge := {node, address, which};
385       HELPTARGETNODE( helpedEdge );
386       continue;
387     else if d then
388       if flag = PROMOTE_FLAG then
389         HELPTARGETNODE( edge );
390         return false;
391       else return true;
392     else if p then
393       if flag = DELETE_FLAG then
394         HELPSUCCESSORNODE( edge );
395         return false;
396       else return true;
397     oldValue :=  $\langle n, 0_i, 0_d, 0_p, address \rangle$ ;
398     newValue := oldValue | flag;
399     result := CAS( node → child[which],
400                   oldValue,
401                   newValue );
400     if result then break;
401   return true;

```

---

---

**Algorithm 18: Find Smallest**


---

```

402 Boolean FINDSMALLEST( state )
403 begin
    // find the node with the smallest key in the subtree rooted at the right child
    // of the target node
404 node := state → targetEdge.child;
405 seekRecord := state → seekRecord;
406  $\langle n, *, *, *, right \rangle$  := node → child[RIGHT];
407 if n then // the right subtree is empty
408     return false;

    // initialize the variables used in the traversal
409 lastEdge :=  $\langle node, right, RIGHT \rangle$ ;
410 pLastEdge :=  $\langle node, right, RIGHT \rangle$ ;
411 while true do
412     curr := lastEdge.child;
413      $\langle n, *, *, *, left \rangle$  := curr → child[LEFT];
414     if n then // reached the node with the smallest key
415         injectionEdge :=  $\langle curr, left, LEFT \rangle$ ;
416         break;

        // traverse the next edge
417     pLastEdge := lastEdge;
418     lastEdge :=  $\langle curr, left, LEFT \rangle$ ;

    // initialize seek record and return
419 seekRecord → lastEdge := lastEdge;
420 seekRecord → pLastEdge := pLastEdge;
421 seekRecord → injectionEdge := injectionEdge;
422 return true;

```

---

it performs another seek on the tree, this time looking for the successor key. If the seek function either fails to find the successor key or returns a terminal node different from  $T$ , then  $T$  has been already replaced (by another operation as part of helping) and the delete operation terminates. Otherwise, it attempts to replace  $T$  again using possibly the new parent information returned by seek. This process may be repeated multiple times.

**Discussion** It can be verified that, in the absence of conflict, a delete operation performs three atomic instructions in the injection mode, three in the discovery mode (if delete is complex), and one in the cleanup mode.



---

**Algorithm 19: Helper Routines**


---

```

423 INITIALIZETYPEANDUPDATEMODE( state )
424 begin
    // retrieve the target node's address from the state record
425   node := state → targetEdge.child;
426    $\langle lN, *, *, *, * \rangle := node \rightarrow child[LEFT]$ ;
427    $\langle rN, *, *, *, * \rangle := node \rightarrow child[RIGHT]$ ;
428   if lN or rN then
    // one of the child pointers is null
429      $\langle m, * \rangle := node \rightarrow mKey$ ;
430     if m then state → type := COMPLEX;
431     else state → type := SIMPLE;
432   else // both child pointers are non-null
433     state → type := COMPLEX;
434   UPDATEMODE( state );

435 UPDATEMODE( state )
436 begin
    // update the operation mode
437   if state → type = SIMPLE then // simple delete
438     state → mode := CLEANUP;
439   else // complex delete
440     node := state → targetEdge.child;
441     if node → readyToReplace then
442       state → mode := CLEANUP;
443     else state → mode := DISCOVERY;

```

---

## Helping

To enable helping, as mentioned earlier, whenever traversing the tree to locate either a target key or a successor key, we keep track of the *last two* edges encountered in the traversal. When a CAS instruction fails, depending on the reason for failure, helping is either performed along the last edge or the second-to-last edge.

### 4.1.3 Formal Description

A pseudo-code of our algorithm is given in Algorithms 8-20.

Algorithm 8 describes the data structures used in our algorithm. Besides **Node**, three important data types in our algorithm are: **Edge**, **SeekRecord** and **StateRecord**. The data

---

**Algorithm 20:** Helping Conflicting Delete Operations
 

---

```

444 HELPTARGETNODE( helpedEdge )
445 begin
    // intent flag must be set on the edge
    // obtain new state record and initialize it
446   state → targetEdge := helpedEdge;
447   state → mode := INJECTION;

    // mark the left and right edges if unmarked
448   result := MARKCHILDEDGE( state, LEFT );
449   if not (result) then return;
450   MARKCHILDEDGE( state, RIGHT );
451   INITIALIZETYPEANDUPDATEMODE( state );

    // perform the remaining steps of a delete operation
452   if state → mode = DISCOVERY then
453     FINDANDMARKSUCCESSOR( state );

454   if state → mode = DISCOVERY then
455     REMOVESUCCESSOR( state );

456   if state → mode = CLEANUP then CLEANUP( state );

457 HELPSUCCESSORNODE( helpedEdge )
458 begin
    // retrieve the address of the successor node
459   parent := helpedEdge.parent;
460   node := helpedEdge.child;
    // promote flag must be set on the successor node's left edge
    // retrieve the address of the target node
461   ⟨*,*,*,*,left⟩ := node → child[LEFT];

    // obtain new state record and initialize it
462   state → targetEdge := {null, left, _};
463   state → mode := DISCOVERY;
464   seekRecord := state → successorRecord;
    // initialize the seek record in the state record
465   seekRecord → lastEdge := helpedEdge;
466   seekRecord → pLastEdge := {null, parent, _};
    // promote the successor node's key and remove the successor node
467   REMOVESUCCESSOR( state );
    // no need to perform the cleanup
  
```

---

type **Edge** is a structure consisting of three fields: the two endpoints and the direction (left or right). The data type **SeekRecord** is a structure used to store the results of a tree traversal. The data type **StateRecord** is a structure used to store information about a delete operation (*e.g.*, target edge, type, current mode, etc.). Note that only objects of type **Node** are shared between processes; objects of all other types (*e.g.*, **SeekRecord**, **StateRecord**) are *local* to a process and not shared with other processes.

The pseudo-code of the seek function is described in Algorithm 9, which is used by all the operations. The pseudo-codes of the search, insert and delete operations are given in Algorithm 10, Algorithm 11 and Algorithm 12, respectively. A delete operation executes function **INJECT** in injection mode, functions **FINDANDMARKSUCCESSOR** and **REMOVESUCCESSOR** in discovery mode and function **CLEANUP** in cleanup mode. Their pseudo-codes are given in Algorithm 13, Algorithm 14, Algorithm 15 and Algorithm 16, respectively. The pseudo-codes for helper routines (used by multiple functions) are given in Algorithm 18, Algorithm 17 and Algorithm 19. Finally, the pseudo-codes of functions used to help other (conflicting) delete operations are given in Algorithm 20.

#### 4.1.4 Correctness Proof

It can be shown that our algorithm satisfies linearizability and lock-freedom properties (Herlihy and Shavit, 2012). Broadly speaking, linearizability requires that an operation should appear to take effect instantaneously at some point during its execution. Lock-freedom requires that some process should be able to complete its operation in a finite number of its own steps. It is convenient to treat insert and delete operations that do not change the tree as search operations. We call a tree node *active* if it is reachable from the root of the tree. We call a tree node *passive* if it was active earlier but is not active any more. It can be verified that, if an insert operation completes successfully, then its target node was active when it performed the successful **CAS** instruction on the node's child edge. Likewise, if a

delete operation completes successfully, then its target node was active when it marked the node's left edge for deletion. Further, for a complex delete, the successor node was active when it marked the node's left edge for promotion.

### All Executions are Linearizable

We show that an arbitrary execution of our algorithm is linearizable by specifying the *linearization point* of each operation. Note that the linearization point of an operation is the point during its execution at which the operation appeared to have taken effect. Our algorithm supports three types of operations: search, insert and delete. We now specify the linearization point of each operation.

1. *Insert operation:* The operation is linearized at the point at which it performed the successful CAS instruction that resulted in its target key becoming part of the tree.
2. *Delete operation:* There are two cases depending on whether the delete operation is simple or complex. If the operation is simple delete, then the operation is linearized at the point at which a successful CAS instruction was performed at the parent of the target node that resulted in the target node becoming passive. Otherwise, it is linearized at the point at which the original key of the target node was replaced with its successor key.
3. *Search operation:* There are two cases depending on whether the target node was active when the operation read the key stored in the node. If the target node was not active, then the operation is linearized at the point at which the target node became passive. Otherwise, it is linearized at the point at which the read action was performed.

It can be easily verified that, for any execution of the algorithm, the sequence of operations obtained by ordering operations based on their linearization points is legal, *i.e.*, all operations in the sequence satisfy their specification. This establishes that *our algorithm generates only linearizable executions*.

## All Executions are Lock-Free

We say that the system is in a *quiescent state* if no modify operation completes hereafter. We say that the system is in a *potent state* if it has one or more pending modify operations. Note that a quiescence is a *stable* property; once the system is in a quiescent state, it stays in a quiescent state. We show that our algorithm is lock-free by proving that a potent state is necessarily non-quiescent provided assuming that some process with a pending modify operation continues to take steps.

Assume, by the way of contradiction, that there is an execution of the system in which the system eventually reaches a state that is potent as well as quiescent. Note that, once the system has reached a quiescent state, it will eventually reach a state after which the tree will not undergo any structural changes. This is because a modify operation makes at most two structural changes to the tree. So, if the tree is undergoing continuous structural changes, then it clearly implies that modify operations are continuously completing their responses, which contradicts the assumption that the system is in a quiescent state. Further, on reaching such a state, the system will reach a state after which no new edges in the tree are marked. Again, this is because a modify operation marks at most four edges and the set of edges in the tree does not change any more. We call such a system state after which neither the set of edges nor the set of *marked* edges in the tree change any more as a *strongly quiescent state*. Note that, like quiescence, strong quiescence is also a stable property.

From the above discussion, it follows that the system in a quiescent state will eventually reach a state that is strongly quiescent. Consider the search tree in such a strongly quiescent state. It can be verified that no more modify operations can now be injected into the tree, and, moreover, all modify operations already injected into the tree are delete operations currently “stuck” in either discovery or cleanup mode. Now, consider a process, say  $p$ , that continues to take steps to execute either its own operation or another operation blocking its progress (directly or indirectly) as part of helping. Consider the recursive chain of the *helpee*

operations that  $p$  proceeds to help in order to complete its own operation. Let  $\alpha_i$  denote the  $i^{\text{th}}$  helpee operation in the chain. It can be shown that:

**Lemma 1.** *Let  $\mathcal{C}_D$  denote the set of all complex delete operations already injected into the tree that are "stuck" in the discovery mode. Then,*

1.  $\alpha_1 \in \mathcal{C}_D$ , and
2. *Suppose  $p$  is currently helping  $\alpha_i$  for some  $i \geq 1$  and assume that  $\alpha_i \in \mathcal{C}_D$ . Let  $\alpha_{i+1}$  denote the next operation that  $p$  selects to help. Then, (a)  $\alpha_{i+1}$  exists, (b)  $\alpha_{i+1} \in \mathcal{C}_D$ , and (c) the target node of  $\alpha_{i+1}$  is at strictly larger depth than the target node of  $\alpha_i$ .*

Using the above lemma, we can easily construct a chain of distinct helpee operations whose length exceeds the number of processes—a contradiction. This establishes that *our algorithm only generates lock-free executions.*

## CHAPTER 5

### LOCAL RECOVERY FOR CONCURRENT BINARY SEARCH TREES

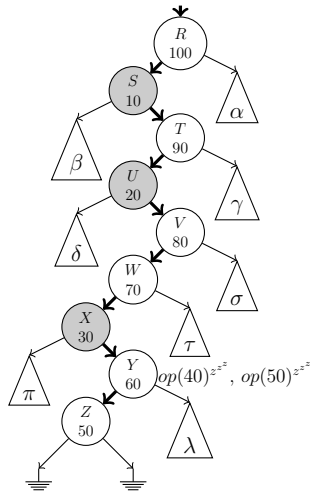
#### 5.1 The Local Recovery Algorithm

We first present the main idea behind the algorithm and then provide its pseudo-code with more details.

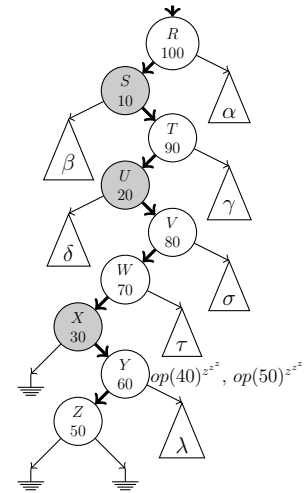
##### 5.1.1 Overview of the Algorithm

As mentioned earlier, every operation on a BST involves first traversing the tree from top to down starting from the root node and following either the left or the right child pointer until either the target key is found or a null pointer is encountered (termination condition). Depending on the outcome of the traversal and the type of the operation, the tree may then need to be modified to actually realize the operation. We refer to the period during which the tree is being traversed as *seek phase*. Further, we refer to the period during which the tree is being modified as *execution phase*.

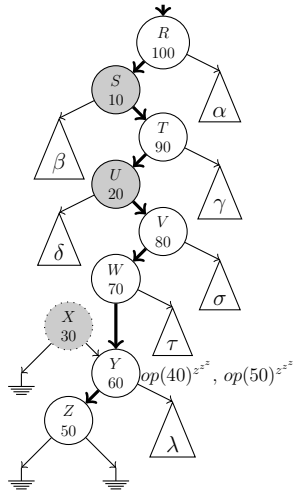
During the seek phase, the target key may move from its current location to a new location up the tree. As a result, the traversal may miss the key both at its old location as well as its new location. For an illustration, see Figure 5.1. In the illustration, key 50 has moved up by five nodes. In most concurrent BST algorithms, if it is suspected that the key may have moved up the tree, then the traversal is simply restarted from the root node. Different algorithms use different approaches to detect possible key movement. For example, in (Howley and Jones, 2012), the traversal is restarted if the *last right-turn* node is detected to have undergone some change. For example, in Figure 5.1a, the last right-turn node for the



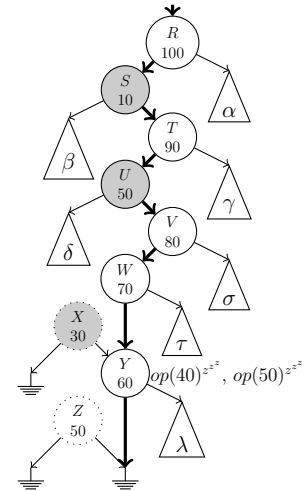
(a) Operation  $op(50)$  is suspended at node  $Y$  during its traversal.



(b) All keys in subtree  $\pi$  are deleted one-by-one.



(c) Key 30 is deleted (simple delete); node  $X$  is removed.



(d) Key 20 is deleted (complex delete); key 20 is replaced with key 50 in node  $U$  and node  $Z$  is removed.

Figure 5.1: An illustration of a key moving up the tree



operation  $op(50)$  is node  $X$ . On reaching the terminal node in Figure 5.1d, after resuming running,  $op(50)$  needs to restart since  $X$  has since been removed from the tree.

A *re-traversal* of the tree may also be required if the operation encounters any failure during the execution phase. For example, in (Ramachandran and Mittal, 2015b), which is a lock-based algorithm, execution phase is aborted if, after locking the relevant edges, the validation step fails. This happens if the portion of the tree that lies within the “operation’s window”, which typically consists of a small constant number of nodes, has undergone some change since it was last observed. In that case, the operation moves to the seek phase again.

In most concurrent BST algorithms, (a single instance of) the execution phase of an operation typically tends to have constant time complexity. The seek phase is where an operation may end up spending most of its time especially if the tree is large. Hence, it is desirable to make the seek phase of an operation more efficient by: (i) reducing the number of restarts due to “suspected” key movement, and (ii) restarting the traversal from a point “close” to the operation’s window. This leads to two separate but related questions that any local recovery algorithm needs to address. First, “If a key is not found, then does the traversal need to restart?”. Second, “If the traversal needs to be restarted, then from which node should the traversal restart?”

Consider an operation  $\alpha$  currently traversing the tree; let  $\Pi(\alpha)$  denote the path taken by  $\alpha$  so far. For example, in Figure 5.1a, considering only the subtree shown in the figure,  $\Pi(op(50)) = \langle R, S, T, U, V, W, X, Y \rangle$ . At each node in the path (except the last node),  $\alpha$  either followed the left or the right child pointer. We say that a node in the path is an *anchor* node if the operation followed its *right* child pointer; otherwise we say that is a *non-anchor* node. For example, in the path  $\Pi(op(50))$ , nodes  $S$ ,  $U$  and  $X$  are anchor nodes, whereas nodes  $R$ ,  $T$ ,  $V$ ,  $W$  and  $Y$  are non-anchor nodes. We assume that the first anchor node in a traversal path is always a *sentinel* node that is never marked (not shown in the figure). Of course, as an operation is traversing the tree, the tree may undergo changes as a

result of which the path taken by the operation may no longer be correct. For example, in Figure 5.1d, the new access-path of  $op(50)$  in the subtree, which is obtained from the subtree in Figure 5.1a after applying several delete operations, is now given by  $\langle R, S, T, U \rangle$ .

Note that, since in a complex delete operation we assume that the key being deleted is replaced with its successor key, the value of a key stored in a node can only increase. Therefore, the child pointer followed by an operation at a node, if it is still part of the access-path, may change (from right to left) for an anchor node but cannot change for a non-anchor node. For example, as shown in Figure 5.1a, the node  $U$  is an anchor node for  $op(40)$ . But due to the changes made to the tree, the key at  $U$  has now become 50, as shown in Figure 5.1d. Hence, the pointer that  $op(40)$  now needs to follow at  $U$  is left and not right. We say that an anchor node is *consistent* with respect to an operation if its key is still less than the operation's key; otherwise, we say that it is *inconsistent*. For example, in Figure 5.1d, anchor nodes  $S$  and  $X$  are still consistent with respect to  $op(40)$  but node  $U$  has become inconsistent with respect to  $op(40)$ .

Clearly, in case an operation needs to restart, *no node* in the path *after an inconsistent anchor node* in general can serve as a restart point since the path taken and the path that needs to be taken may now diverge. This implies that, to find a restart point, a local recovery algorithm should locate the *shallowest* inconsistent anchor node in the path (with the least depth) and discard the suffix of the path after such a node. Moreover, a restart point has to be a node that is still a part of the tree.

This leads to the following approach to find a restart point for an operation  $\alpha$  when needed. Find a node  $C$  in the path taken so far by  $\alpha$  such that the following two conditions hold. First,  $C$  is not marked. Second, every anchor node in the path preceding  $C$  is consistent with respect to  $\alpha$ . To check for the second condition, it is not necessary to examine every anchor node in the path preceding  $C$  as stated in the following lemma.

**Lemma 2.** *Consider an operation  $\alpha$  and let  $\Pi(\alpha)$  denote the path taken by  $\alpha$  when traversing the tree. Let  $A$  be an anchor node in  $\Pi(\alpha)$  and let  $\sigma = A_0, A_1, \dots, A_k$ , where  $A_k = A$ , denote the sequence of anchor nodes in  $\Pi(\alpha)$  up to and including  $A$ . Then, if  $A$  is unmarked and consistent with respect to  $\alpha$ , then, for every  $i$  with  $0 \leq i \leq k$ ,  $A_i$  is also consistent with respect to  $\alpha$ . Moreover, the access-path of  $\alpha$  in the current tree includes  $A$ .*

We say that an anchor node is *critical* with respect to a node  $C$  in the path if it is the *closest* preceding anchor node to  $C$  that is also unmarked. For example, in Figure 5.1c, the critical anchor node with respect to node  $Y$  is node  $U$  since node  $X$  is marked. Using the above lemma, the second condition can now be replaced with the following: the critical anchor node with respect to  $C$  in the path, say  $A$ , as well as every anchor node in the path that lies between  $A$  and  $C$ , which will be marked, is consistent with respect to  $\alpha$ .

The next lemma states a useful property about an inconsistent anchor node.

**Lemma 3.** *Consider an operation  $\alpha$  with target key  $k$  and let  $\Pi(\alpha)$  denote the path taken by  $\alpha$  when traversing the tree. Let  $A$  be an anchor node in  $\Pi(\alpha)$ . Assume that  $A$  is now inconsistent with respect to  $\alpha$ . Then, at the time  $A$  became inconsistent, the tree did not contain  $k$ .*

To see why the above lemma holds, let  $k_{old}$  ( $k_{new}$ ) be the key stored in  $A$  just before (after)  $A$  became inconsistent. Clearly,  $k_{old} < k < k_{new}$ . Let  $t$  denote the time just after which  $k_{old}$  was replaced with  $k_{new}$  in  $A$ . Note that  $k_{new}$  must be the next smallest key in the right subtree of  $A$  at time  $t$ . This implies that the right subtree of  $A$  did not contain  $k$  at time  $t$ . Further, from Lemma 2,  $A$  is on the access-path of  $\alpha$  at time  $t$ . Hence, we can conclude that the tree does not contain  $k$  at time  $t$ . Note that, if a key is not present in the tree at some point while a search/delete operation is in progress, it is acceptable for the operation to say that key was not found. In this case, the operation will be linearized after the delete operation that removed the key from the tree.

When an operation fails to find the target key after traversing the tree from top to bottom, it examines the path it took to check whether or not the key has moved up the tree and/or a re-traversal is required. To that end, it examines the anchor nodes in the reverse order in which they were visited, starting from the one closest to the terminal node. We now discuss the behavior of each operation one-by-one.

**Search Operation:** A search operation *does not* need to restart. When it examines an anchor node as mentioned above, there are three possibilities. First, if the anchor node's key matches the target key, then the key has been found and the operation terminates. Second, if the anchor node's key is greater than the target key (the anchor node has become inconsistent), then the operation concludes that the key is not present in the tree and terminates. Finally, if the anchor node's key is smaller than the target key (the anchor node is still consistent), then the operation terminates if the node is not marked; otherwise it moves to the preceding anchor node and repeats the comparison.

**Insert Operation:** An insert operation needs to restart only if one of the anchor nodes in the path has become inconsistent. When it examines an anchor node as mentioned above, there are three possibilities. First, if the anchor node's key matches the target key, then the key has been found and the operation terminates. Second, if the anchor node's key is greater than the target key (the anchor node has become inconsistent), then it discards the suffix of the path after the anchor node and restarts the traversal from a restart point. Finally, if the anchor node's key is smaller than the target key (the anchor node is still consistent), then the traversal terminates if the node is not marked (the terminal node is returned as the injection point); otherwise it moves to the preceding anchor node and repeats the comparison.

**Delete Operation:** A delete operation also *does not* need to restart except when there is a failure in the execution phase. When it examines an anchor node as mentioned above,

---

**Algorithm 21:** Data Structures Used

---

```

// Used to store information about a node visited during tree traversal
468 struct StackEntry {
469     NodePtr node;
470     enum Direction which;
471     integer anchor;
472 };

// Used to store the path from the root node to the current node in the tree
473 struct State {
474     StackEntry[ ] stack;
475     integer top;
476 };

// Used to store information about the operation currently in progress
477 struct OpRecord {
478     enum Type type;
479     Key key;
480     State targetStack, successorStack;
481     NodePtr injectionPoint;

    // algorithm-specific fields
482 };

// Used to store the outcome of a tree traversal
483 struct SeekRecord{
    // algorithm-specific fields (e.g., target node and its parent)
484 };

```

---

there are three possibilities. First, if the anchor node's key matches the target key, then the key has been found and the operation moves to the execution phase. Second, if the anchor node's key is greater than the target key (the anchor node has become inconsistent), then the operation concludes that the key is not present in the tree and terminates. Finally, if the anchor node's key is smaller than the target key (the anchor node is still consistent), then the traversal terminates if the node is not marked; otherwise it moves to the preceding anchor node and repeats the comparison.

### 5.1.2 Details of the Algorithm

A pseudo-code of the local recovery algorithm is given in Pseudo-codes 21-28. The pseudo-code only shows the seek phase of an algorithm and not its execution phase since the exe-

---

**Algorithm 22:** Functions for Manipulating Traversal Stack

---

```

// Returns the number of elements in the stack
485 integer SIZE( state )
486 begin
487   return state → top + 1;

// Returns the topmost node in the stack
488 NodePtr GETTOP( state )
489 begin
490   {stack, top} := state;
491   return stack[top] → node;

// Returns the second topmost node in the stack
492 NodePtr GETSECONDTOTOP( state )
493 begin
494   {stack, top} := state;
495   return stack[top - 1] → node;

// Adds the given node to the stack along with its anchor node
496 ADDTOTOP( state, node, which )
497 begin
498   {stack, top} := state;
499   // find the anchor node
500   if which = RIGHT then anchor := top ;
501   else anchor := stack[top] → anchor ;
502   // push the node into the stack
503   stack[top + 1] := {node, which, anchor};
504   state → top := top + 1;

// Removes the topmost node from the stack
505 REMOVEFROMTOP( state )
506 begin
507   {stack, top} := state;
508   // update the anchor node of the penultimate entry if needed
509   anchor := stack[top - 1] → anchor;
510   if stack[top] → anchor < stack[anchor] → anchor then
511     stack[anchor] → anchor := stack[top] → anchor;
512   // pop the node from the stack
513   state → top := top - 1;

// Pops the stack until a given entry
514 REMOVEUNTIL( state, index )
515 begin
516   state → top := index;

```

---

cution phase is algorithm-specific. We have also moved the pseudo-code for local recovery when looking for a successor key to the appendix due to lack of space.

---

**Algorithm 23:** Functions for Manipulating Traversal Stack (Continued)

---

```

// Remember the critical node (to avoid locating it again)
513 REMEMBERCRITICAL( state, critical )
514 begin
515   {stack, top} := state;
516   anchor := stack[top] → anchor;
517   if critical < stack[anchor] → anchor then
518     stack[anchor] → anchor := critical;

// Returns a given entry in the stack
519 { NodePtr, enum Direction, integer }
520   GETFULLENTRY( state, index )
521 begin
522   {stack, top} := state;
523   if index = ⊤ then return stack[top] ;
524   else return stack[index] ;

// initializes the traversal stack
525 INITIALIZE TRAVERSAL STATE( state, type )
526 begin
527   if type = TARGET_STACK then
528     // initialize the stack using sentinel nodes
529     // sentinel nodes are never removed from the stack
530     // a sentinel node is always a safe starting point for the traversal
531   else state → top := -1 ;

```

---

The local recovery algorithm assumes that the original algorithm supports the following functions: (a) `GETKEY( )`, `ISMARKED( )` and `GETCHILD( )` returns the various attributes of a tree node, (b) `ISNULL( )` returns true if a reference is null and false otherwise, (c) `GETADDRESS( )` returns the node address stored in a reference, if non-null, (d) `MOVE( )` enables the original algorithm to move along an edge, which may invoke helping and restarting of the traversal as in (Howley and Jones, 2012), (e) `NEEDCLEANPARENTNODE( )` returns true if the operation needs the parent node to be clean and have no operation in progress (needed for a delete operation since it needs to modify a child pointer at the parent node), and (f) `POPULATESEEKRECORD( )` copies the relevant information from the traversal state required by the algorithm into a seek record.

Pseudo-code 21 shows the data structures used by the local recovery algorithm. Note that all the data structures shown in Pseudo-code 21 are *local* to a process not shared

among processes. A process uses three main data structures, namely **State**, **OpRecord** and **SeekRecord**. A **State** (lines 473-476) is essentially a stack used to store the nodes visited during tree traversal when looking for a key (target or successor). Note that the traversal stack satisfies the last-in-first-out (LIFO) semantics but our algorithm sometimes uses it in a non-traditional way by accessing entries in the middle of the stack. One way to implement such an “augmented” stack is to use an auto-resizing vector provided as part of C++ STL library or Java package. Each entry in a traversal stack (lines 468-472) stores the address of the node, the location of its closest anchor node (within the stack’s vector) and whether the node is a left or right child of its parent. An **OpRecord** (lines 477-482) stores information about the operation such as type and key as well two stacks: one used when looking for the target key (all operations) and one used when looking for the successor key (only complex delete operations). Finally, a **SeekRecord** (lines 483-484) is used to return the outcome of a tree traversal to the original algorithm. Its fields are algorithm-specific. For example, for CASTLE, **SeekRecord** contains three fields: (a) two addresses, namely those of the target node and its parent, and (b) the contents of the injection point where an insert operation needs to attach the new node.

Pseudo-code 22 shows the functions used to manipulate a traversal stack. The function **SIZE** (lines 485-487) returns the number of entries in the stack. The functions **GETTOP** (lines 488-491) and **GETSECONDTOTOP** (lines 492-495) return the address of the node stored in the topmost entry and the entry below it, respectively. The function **ADDTOTOP** (lines 496-502) adds an entry to the top of the stack while **REMOVEFROMTOP** (lines 503-509) removes an entry from the top of the stack. The function **REMOVEUNTIL** (lines 510-512) removes the entries from the top of the stack until a given point. The function **REMEMBER-CRITICAL** (lines 513-518) updates the anchor field of the anchor node of the topmost entry in the stack. The function **GETFULLENTY** (lines 520-524) returns all the three fields of a given entry in the stack (may not be the topmost entry). The function **INITIALIZETRAVERSALSTATE** (lines 525-528) initializes a traversal stack. The stack for target key



Pseudo-codes 24 & 25 shows the functions used to find the target key by a search operation. The function `SEEKFORSEARCH` (lines 560-566) first traverses the tree starting from the root node (line 562). If the traversal fails to locate the key, then the key may have moved up the tree. To address this possibility, the function examines the traversal stack to determine whether or not that is the case (line 564). The function `TRAVERSE` (lines 529-542) first initializes the traversal stack (line 532) and then, starting from the topmost node in the stack (line 533), follows either the left or the right child pointer (line 536) until it either finds the key (line 538) or encounters a null pointer (line 539). It also populates the traversal stack as it moves (line 542). The function `EXAMINESTACK` (lines 543-559) examines the anchor nodes stored in the stack in the reverse order in which they were visited, starting from the anchor node closest to the topmost node in the traversal stack (line 547). If the anchor node's key matches the target key, then the function returns true (lines 552-554). If the anchor node is no longer consistent or is unmarked, then the function returns false (lines 555-556). Otherwise, the function backtracks and examines the preceding anchor node in the stack (lines 557-558).

Pseudo-codes 26-27 & 28 show the functions used to find the target key by a modify (insert or delete) operation. The function `SEEKFORMODIFY` (lines 603-631) first backtracks to a safe node in the stack (line 608). Initially, the starting point is typically a sentinel node which is a safe node. The function then traverses the tree from top to down by following either the left or the right child pointer (line 613) until it either finds the key or encounters a null pointer (lines 615-624). In case the terminal node's key is greater than the target key, the function checks whether the path stored in the traversal stack is still valid (line 617). If not, the traversal is restarted. As the traversal moves down the tree, the function also populates the traversal stack (lines 625-629). The function `VALIDATEPATH` (lines 567-588) checks whether or not the path stored in the stack is still valid. To that end, it examines the anchor nodes in the stack in the reverse order in which they were visited, starting from

the anchor node closest to the topmost node in the traversal stack. There are three possible cases. First, the anchor node is still consistent (lines 574-579). In this case, the path is deemed to be valid if the anchor node is unmarked; otherwise, the function moves to the preceding anchor node. Second, the anchor node is no longer consistent (lines 580-585). In this case, the path is deemed to be invalid. However, if the operation is a delete operation, then it can be deduced that the key did not exist in the tree continuously and the function returns indicating that the key was not found (thereby causing the operation to terminate). Finally, the anchor node's key matches the target key (lines 586-588). In this case, if the anchor node is marked and the operation is a delete operation, then the path is deemed to be invalid (and further backtracking is required). This is because the key may be in the process of moving up the tree. Otherwise, the function returns indicating that the key was found. The function `FINDASAFENODE` (lines 589-602) finds a safe node on the path stored in the stack from which the operation can restart its traversal. To that end, it backtracks to an unmarked node with a clean parent if required (lines 592-600). It then checks whether or not the remaining path in the stack is still valid (line 601). If not, it repeats the above-mentioned steps.

---

**Algorithm 24:** Seek Function for Target Key (Search Operation)

---

```

    // Traverses the tree starting from the root until either the key is found or a null
    // pointer is encountered
529 boolean TRAVERSE( opRecord, seekRecord )
530 begin
531     state := opRecord→targetStack;
    // initialize the stack and the variables used in the traversal
532     INITIALIZE TRAVERSAL STATE( state, TARGET_STACK );
533     current := GET TOP( state );
    // traverse the tree (starting from current)
534     while true do
535         key := GET KEY( current );
536         which := opRecord→key < key ? LEFT : RIGHT;
    // read the next address to de-reference
537         reference := GET CHILD( current, which );
538         if opRecord→key = key then return true ;
539         if IS NULL( reference ) then return false ;
    // traverse the next edge
540         address := GET ADDRESS( reference );
541         current := address;
    // push the next node to be visited into the stack
542         ADD TO TOP( state, address, which );

    // Checks if the key being searched for has moved up in the path stored in the stack
543 boolean EXAMINE STACK( opRecord, seekRecord )
544 begin
545     result := false;
546     state := opRecord→targetStack;
    // start with the anchor closest to the topmost node in the stack
547     {*,*,critical} := GET FULL ENTRY( state,  $\top$  );
548     while true do
549         // retrieve the node and its closest anchor node from the stack
550         {node*,anchor} := GET FULL ENTRY( state, critical );
    // read the attributes of the node
551         marked := IS MARKED( node );
552         key := GET KEY( node );
553         if opRecord→key = key then
554             // the key stored in the node matches the one being searched for
555             result := true;
556             break;
557         else if (opRecord→key < key) or not (marked) then
558             // the target key did not exist continuously in the tree
559             break;
560         else // examine the preceding anchor node
561             critical := anchor;
562     return result;

```

---

---

**Algorithm 25:** Seek Function for Target Key (Search Operation) (continued)

---

```

// Looks for a given key in the tree (invoked by a search operation)
560 boolean SEEKFORSEARCH( opRecord, seekRecord )
561 begin
    // traverse the tree from top to down
562     result := TRAVERSE( opRecord, seekRecord );
563     if not (result) then
        // check if the key has moved up in the path stored in the stack
564         result := EXAMINESTACK( opRecord, seekRecord );
    // return the outcome
565     POPULATESEEKRECORD( seekRecord, state );
566 return result;

```

---

---

**Algorithm 26:** validate path
 

---

```

    // Determines if the path stored in the stack is still valid
    // Returns one of the following four values:
    // { SAFE, NOT_SAFE, FOUND, NOT_FOUND}
    // May backtrack along the path under certain situations
567 enum Outcome VALIDATEPATH( opRecord, state )
568 begin
    // check if any of the anchor nodes in the stack has become inconsistent
    // starting with the one immediately preceding the topmost node in the stack
569 {*,*,critical} := GETFULLENTY( state,  $\top$  );
570 while true do
    // retrieve the node and its anchor from the stack
571 {node,*,anchor} := GETFULLENTY( state, critical );
    // read the attributes of the node
572 marked := ISMARKED( node );
573 key := GETKEY( node );
574 if opRecord→key > key then
    // the anchor node is still consistent
575 if not (marked) then
    // the access-path is still valid
576 REMEMBERCRITICAL( state, critical );
577 return SAFE;
578 else // need to check the previous anchor node
579 | critical := anchor;
580 else if opRecord→key < key then
    // the anchor node is no longer consistent
581 if opRecord→type = DELETE then
    // the target key did not exist continuously in the tree
582 return NOT_FOUND;
583 else // the path is not valid
584 | REMOVEUNTIL( state, critical );
585 | return NOT_SAFE;
586 else // the two keys match
587 | REMOVEUNTIL( state, critical );
    // stop the traversal
588 | return FOUND;
  
```

---

---

**Algorithm 27:** find a safe node
 

---

```

    // Backtracks along the path stored in the stack until a suitable restart point is
    // found
589 enum Outcome FINDASAFENODE( opRecord, state )
590 begin
591   while true do
592     // backtrack until an unmarked node
593     current := GETTOP( state );
594     while ISMARKED( current ) do
595       REMOVEFROMTOP( state );
596       current := GETTOP( state );
597
598     // check if the algorithm needs a clean parent node
599     if NEEDCLEANPARENTNODE( opRecord, current ) then
600       parent := GETSECONDTOTOP( state );
601       if not (ISCLEAN( parent )) then
602         // need to backtrack even further
603         REMOVEFROMTOP( state );
604         continue;
605
606     // check if the topmost node in the stack is a suitable restart point
607     result := VALIDATEPATH( opRecord, state );
608     if result ≠ NOT_SAFE then return result ;

```

---

---

**Algorithm 28:** Seek Function for Target Key (Modify Operation)

---

```

// Looks for a given key in the tree (invoked by insert/delete operation)
603 boolean SEEKFORMODIFY( opRecord, seekRecord )
604 begin
605   state := opRecord → targetStack;
606   result := NOT_SAFE;
607   while result = NOT_SAFE do
608     // backtrack to a suitable restart point in the path stored in the stack
609     result := FINDASAFENODE( opRecord, state );
610     if result ∈ {FOUND, NOT_FOUND} then break;
611     // traverse the tree starting from the topmost node in the stack
612     current := GETTOP( state );
613     while true do
614       key := GETKEY( current );
615       which := opRecord → key < key ? LEFT : RIGHT;
616       // read the next address to de-reference
617       reference := GETCHILD( current, which );
618       if (opRecord → key = key) or ISNULL( reference ) then
619         // either stop or backtrack & restart
620         if opRecord → key < key then
621           // check if the path traversed is still valid
622           result := VALIDATEPATH( opRecord, state );
623         else // determine what value to return
624           result := FOUND;
625           if opRecord → key ≠ key then
626             // remember the address read
627             opRecord → injectionPoint :=
628               GETADDRESS( reference );
629             result := NOT_FOUND;
630         // terminate the current traversal
631         break;
632     // traverse the next edge
633     address := GETADDRESS( reference );
634     restart := MOVE( current, address, which );
635     if restart then
636       // the algorithm wants to restart the traversal
637       break;
638     // push the node visited into the stack
639     ADDTOTOP( state, address, which );
640
641   // return the outcome
642   POPULATESEEKRECORD( seekRecord, opRecord );
643   return (result = FOUND ? true: false);

```

---

## CHAPTER 6

### EXPERIMENTAL EVALUATION

We now describe the results of the comparative evaluation of different implementations of a concurrent BST using simulated workloads. This chapter is organized as follows. Performance evaluation of our lock-based algorithm is described in Section 6.2 followed by our lock-free algorithm described in Section 6.3. Performance evaluation of our local recovery technique is described in Section 6.4.

#### 6.1 Experimental Setup

We conducted our experiments on a single large-memory node in stampede<sup>1</sup> cluster at TACC (Texas Advanced Computing Center). This node is a Dell PowerEdge R820 server with 4 Intel E5-4650 8-core processors (32 cores in total) and 1TB of DDR3 memory. Hyper-threading has been disabled on the node. It runs CentOS 6.3 operating system.

To better understand the scalability of our algorithms we also conducted experiments on a single Intel Xeon Phi SE10P Coprocessor<sup>2</sup> having 61 1.1 GHz cores with 4 hardware threads per core and 8GB of GDDR5 memory.

We used Intel C/C++ compiler (version 2013.2.146) with optimization flag set to O3. We used GNU Scientific Library to generate random numbers. We used Intel’s *TBBMalloc* (Reindeers, 2007) as the dynamic memory allocator since it provided superior performance to C/C++ default allocator in a multi-threaded environment.

---

<sup>1</sup><https://www.tacc.utexas.edu/systems/stampede>

<sup>2</sup><http://www.intel.com/content/www/us/en/processors/xeon/xeon-phi-detail.html>



To compare the performance of different implementations, we considered the following parameters:

1. **Relative Distribution of Operations:** We considered three different workload distributions: (a) *read-dominated*: 90% search, 5% insert and 5% delete, (b) *mixed*: 70% search, 15% insert and 15% delete, and (c) *write-dominated*: 0% search, 50% insert and 50% delete.
2. **Maximum Degree of Contention:** This depends on number of threads that can concurrently operate on the tree. On 32 core machine, we varied the number of threads from 1 to 32 in powers of two. On 61 core machine we varied the number of threads from 1 to 244 in multiples of 61.
3. **Maximum Tree Size:** This depends on the size of the key space. To get the peak throughput, we set the number of threads to be the value where the peak performance is achieved and we varied key space size from  $2^{13}$  (8Ki) to  $2^{24}$  (16Mi). To understand the scalability of the algorithms, we varied the number of threads and considered four different key ranges: 2,000 (2k), 20,000 (20K), 200,000 (200K) and 2 million (2M) keys.

We compared the performance of different algorithms with respect to *system throughput*, given by the number of operations executed per unit time. In each run of the experiment, we ran each algorithm for 10 seconds, and calculated the total number of operations completed by the end of the run to determine the system throughput. The results were averaged over 10 runs. To capture only the steady state behaviour, we *pre-populated* the tree to 50% of its maximum size, prior to starting a simulation run. The beginning of each run consisted of a 1 second “warm-up” phase whose numbers were excluded in the computed statistics to avoid initial caching effects.

## 6.2 Lock based tree

In this section we evaluate CASTLE against three other implementations of a concurrent BST, namely those based on:

- (i) the lock-free internal BST by Howley and Jones (Howley and Jones, 2012), denoted by LF-IBST,
- (ii) the lock-free external BST by Natarajan and Mittal (Natarajan and Mittal, 2014), denoted by LF-EBST, and
- (iii) the RCU-based internal BST by Arbel and Attiya (Arbel and Attiya, 2014), denoted by CITRUS.

Note that CITRUS is a blocking implementation. The above three implementations were obtained from their respective authors. All implementations were written in C/C++. In our experiments, none of the implementations used garbage collection to reclaim memory. The experimental evaluation in (Howley and Jones, 2012; Natarajan and Mittal, 2014) showed that, in all cases, either LF-IBST or LF-EBST outperformed the concurrent BST implementation based on Ellen *et al.*'s lock-free algorithm in (Ellen et al., 2010). So we did not consider it in our experiments.

The results of our experiments are shown in Figure 6.1 and Figure 6.2. In Figure 6.1, the absolute value of the system throughput is plotted against the number of threads (varying from 1 to 32 in powers of 2). Here each column represents a specific workload (read-dominated, mixed or write-dominated) and each row represents a specific key space size (50K, 500K or 5M). In Figure 6.2, the relative value of the system throughput with respect to that of LF-IBST is plotted against the key space size. Here each column represents a range of key space sizes (small, medium and large) and each row represents a specific

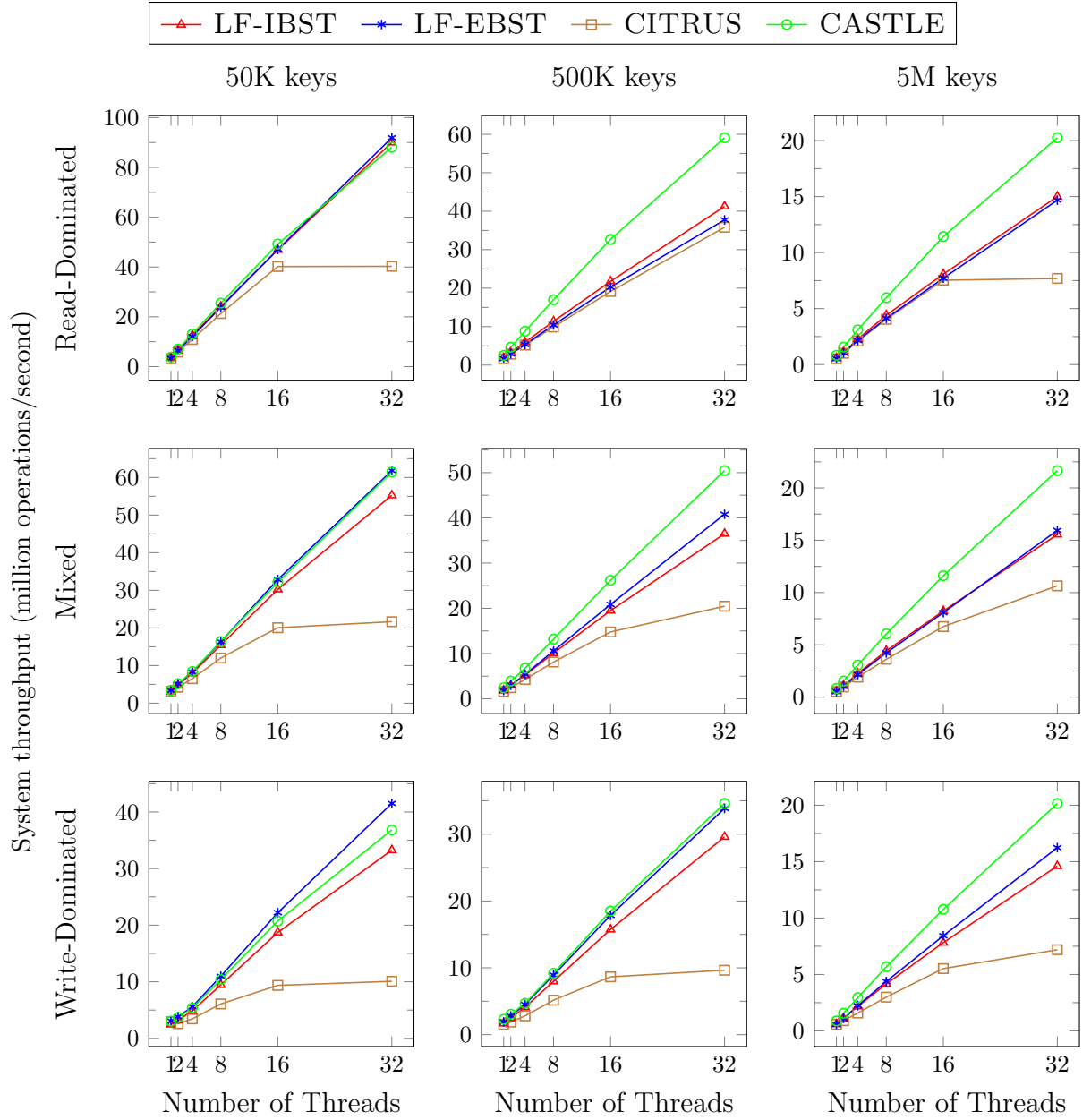


Figure 6.1: Comparison of system throughput of different algorithms. Each row represents a key space size and each column represents a workload type. Higher the throughput, better the performance of the algorithm.

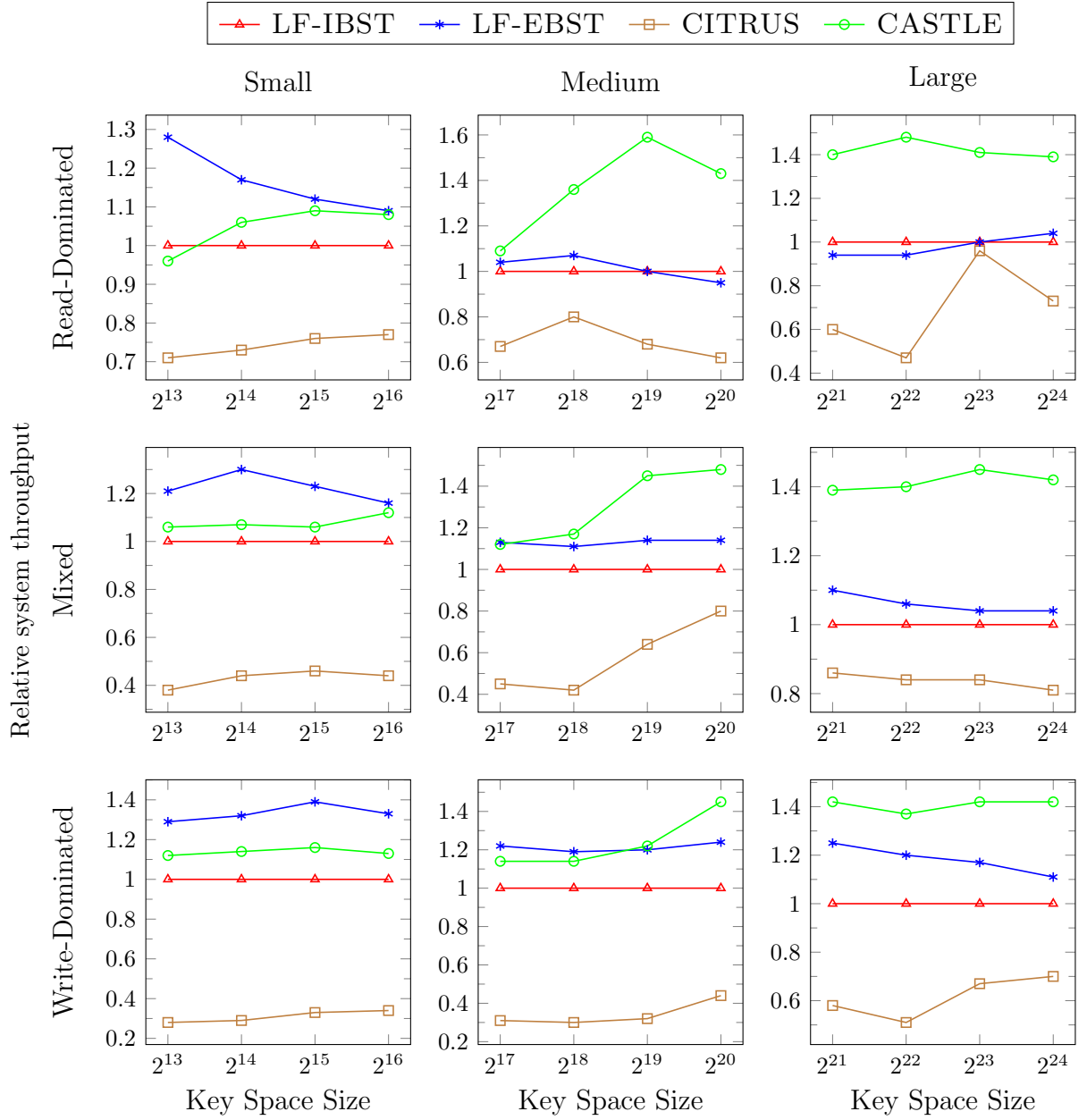


Figure 6.2: Comparison of system throughput of different algorithms *relative to that of* LF-IBST at 32 threads. Each row represents a workload type. Each column represents a range of key space size. Higher the ratio, better the performance of the algorithm.

workload. As the peak performance for all the four algorithms (for all cases) occurred at 32 threads, we set the number of threads to 32 while varying the key space size from  $2^{13}$  to  $2^{24}$ .

As both Figure 6.1 and Figure 6.2 show, for smaller key space sizes, LF-EBST achieves the best system throughput. This is not surprising since LF-EBST is optimized for high contention scenarios. For medium and large key space sizes, CASTLE achieves the best system throughput for all three workload types in almost all the cases (except when the workload is write-dominated and the key space size is in the lower half of the medium range). The maximum gap between CASTLE and the next best performer is around 59% which occurs at 500K key space size, read-dominated workload and 32 threads.

We believe that some of the reasons for the better performance of CASTLE over the other three concurrent algorithms, especially when the contention is relatively low, are as follows. First, as explained in (Natarajan and Mittal, 2014), operating at edge-level rather than at node-level reduces the contention among operations. Second, using a CAS instruction for locking an edge also validates that the edge has not undergone any change since it was last observed. Third, locking edges as late as possible minimizes the blocking effect of the locks.

Table 6.1: Comparison of different concurrent algorithms in the absence of contention.

Algorithm	Number of Objects Allocated		Number of Synchronization Primitives Executed	
	Insert	Delete	Insert	Delete
LF-IBST	2	1	3	simple: 4 complex: 9
LF-EBST	2	0	1	3
CASTLE	1	0	1	simple: 3 complex: 4

Table 6.1 shows a comparison of LF-IBST, LF-EBST and CASTLE with respect to the number of objects allocated dynamically and the number of synchronization primitives executed per modify operation in the absence of contention. We omitted CITRUS in this comparison since it based on a different framework. As Table 6.1 shows, our algorithm allocates fewer objects dynamically than the two lock-free algorithms (one for insert operations

and none for a delete operations). Further, again as Table 6.1 shows, our algorithm executes much fewer synchronization primitives than LF-IBST. It executes the same number of synchronization primitives as LF-EBST for insert and simple delete operations and only one more for complex delete operations. This is important since a synchronization primitive is usually much more expensive to execute than a simple read or write instruction. Finally, we observed in our experiments that CASTLE had a smaller memory footprint than all the three implementations (by a factor of two or more) since it uses internal representation of a search tree and allocates fewer objects dynamically. As a result, it was likely able to benefit from caching to a larger degree than the other algorithms.

In our experiments, we observed that, for key space sizes larger than 10K, the likelihood of an operation restarting was extremely low (less than 0.1%) even for a write-dominated workload. This implies that, in at least 99.9% of the cases, an operation was able to complete without encountering any conflicts. Thus, for key space sizes larger than 10K, we expect CASTLE to outperform the implementation based on the lock-free algorithm described in (Ellen et al., 2014), which is basically derived from the one in (Ellen et al., 2010).

### 6.3 Lock free tree

In this section we evaluate ELFTREE against three other implementations of a concurrent BST, namely those based on:

- (i) the lock-free internal BST by Howley and Jones (Howley and Jones, 2012), denoted by LF-IBST,
- (ii) the lock-free external BST by Natarajan and Mittal (Natarajan and Mittal, 2014), denoted by LF-EBST, and
- (iii) the RCU-based internal BST by Arbel and Attiya (Arbel and Attiya, 2014), denoted by CITRUS.

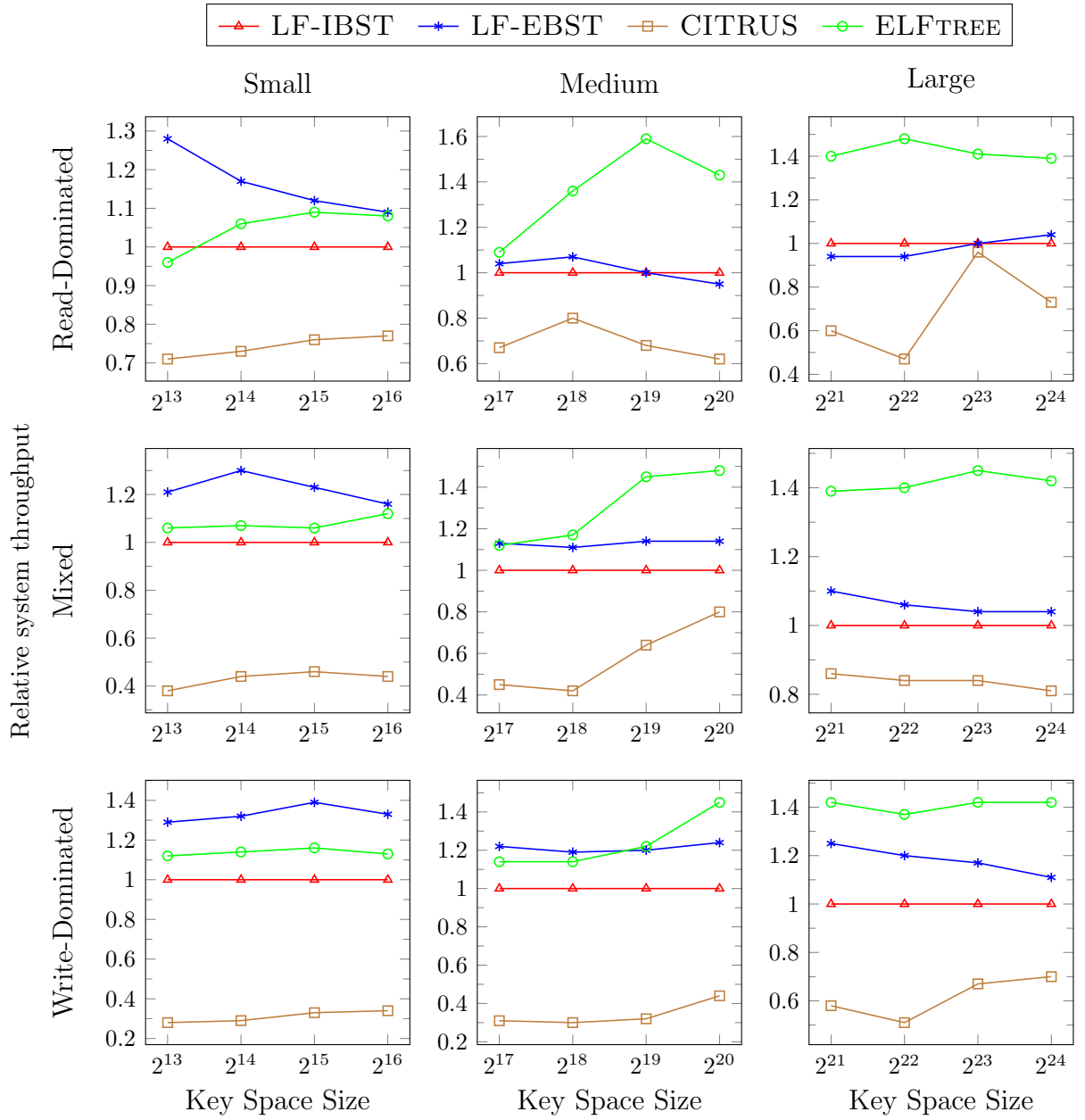


Figure 6.3: Comparison of system throughput of different algorithms *relative to that of* LF-IBST at 32 threads. Each row represents a workload type. Each column represents a range of key space size. Higher the ratio, better the performance of the algorithm.

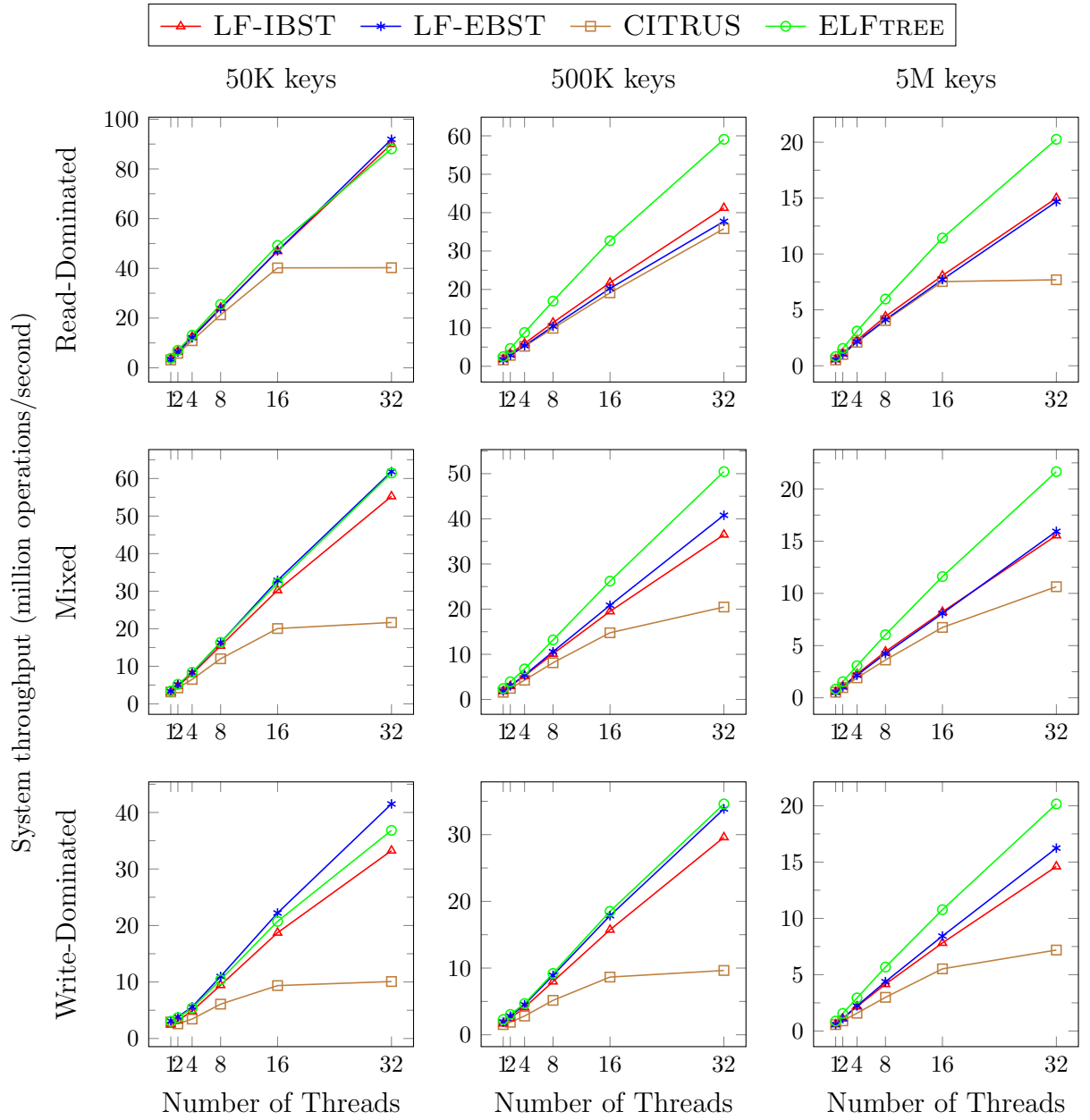


Figure 6.4: Comparison of system throughput of different algorithms. Each row represents a key space size and each column represents a workload type. Higher the throughput, better the performance of the algorithm.



The results of our experiments are shown in Figure 6.3 and Figure 6.4. In Figure 6.3, each row represents a specific workload (read-dominated, mixed or write-dominated) and each column represents a specific key space size; *small* (8Ki to 64Ki), *medium* (128Ki to 1Mi) and *large* (2Mi to 16Mi). Figure 6.4 shows the scaling with respect to the number of threads for key space size of  $2^{19}$  (512Ki). We do not show the numbers for CITRUS in the graphs as it had the worst performance among all implementations (slower by a factor of four in some cases). This is not surprising as CITRUS is optimized for read operations (*e.g.*, 98% reads & 2% updates) (Arbel and Attiya, 2014).

As the graphs show, ELFTREE achieved nearly same or higher throughput than the other two implementations for medium and large key space sizes (except for medium key space size with write-dominated workload). Specifically, at 32 threads and for a read-dominated workload, ELFTREE had 35% and 24% higher throughput than the next best performer for key space sizes of 512Ki and 1Mi, respectively. Also, at 32 threads and for a mixed workload, ELFTREE had 27% and 19% higher throughput than the next best performer for key space sizes of 1Mi and 2Mi, respectively. Overall, ELFTREE outperformed the next best implementation by as much as 35%; it outperformed LF-IBST by as much as 44% and LF-EBST by as much as 35% (both achieved for medium key space sizes). For large key space sizes, the overhead of traversing the tree appears to dominate the overhead of actually modifying the operation’s window, and the gap between various implementations becomes smaller.

Table 6.2: Comparison of different lock-free algorithms in the absence of contention.

Algorithm	Number of Objects Allocated		Number of Atomic Instructions Executed	
	Insert	Delete	Insert	Delete
LF-IBST	2	simple: 1 complex: 1	3	simple: 4 complex: 9
LF-EBST	2	0	1	3
ELFTREE	1	simple: 0 complex: 1	1	simple: 4 complex: 7

There are several reasons why `ELFTREE` outperformed the other two implementations in many cases. First, as Table 6.2 shows, our algorithm allocates fewer objects than the two other algorithms on average considering the fact that the fraction of insert operations will generally be larger than the fraction of delete operations in any realistic workload. Further, we observed in our experiments that the number of simple delete operations outnumbered the number of complex delete operations by two to one, and our algorithm does not allocate any object for a simple delete operation. Second, again as Table 6.2 shows, our algorithm executes the same number of atomic instructions as in (Natarajan and Mittal, 2014) for insert operations; and, in all the cases, executes same or fewer atomic instructions than in (Howley and Jones, 2012). This is important since an atomic instruction is more expensive to execute than a simple read or write instruction. Third, we observed in our experiments that `ELFTREE` had a smaller memory footprint than the other two implementations (by almost a factor of two) since it uses internal representation and allocates fewer objects. As a result, it was likely able to benefit from caching to a larger degree than `LF-IBST` and `LF-EBST`.

#### 6.4 Impact of local recovery

In this section we evaluate our local recovery technique.

To show that our local recovery algorithm is sufficiently general, we implemented it for three different concurrent internal BSTs, namely those based on: (i) the lock-free BST by Howley and Jones (Howley and Jones, 2012), denoted by `LF-IBST`, (ii) the lock-based BST by Ramachandran and Mittal (Ramachandran and Mittal, 2015b), denoted by `CASTLE` and (iii) the RCU (Read-Copy-Update) lock-based BST by Arbel and Attiya (Arbel and Attiya, 2014), denoted by `CITRUS`. These implementations were chosen so that we covered both lock-free and lock-based approaches. We choose two lock-based implementations as

one is based on locking edges (Ramachandran and Mittal, 2015b) and the other is based on locking nodes using RCU framework (Arbel and Attiya, 2014).

Usually uniform key distribution (where all keys have same frequency of occurrence) have been used to evaluate concurrent BSTs. But in many of the real world workloads, keys have skewed distribution (Clauset et al., 2009) where some keys are more popular than others. Zipfian distribution, a type of power-law distribution simulates this behavior (Breslau et al., 1999; Faloutsos and Jagadish, 1992; Gray et al., 1994). It is characterized by a parameter  $\alpha$  which usually lies between 0.5 and 1 (Breslau et al., 1999; Adamic and Huberman, 2002). In our experiments we used both uniform and Zipfian distributions to evaluate the local recovery algorithm.

To better understand the effect of local recovery, we also measure *seek time* which is defined as the total time an operation spends on tree traversal including all restarts as well as stack processing time. For both uniform and Zipfian distribution CASTLE and LF-IBST reached peak performance at 244 threads while CITRUS reached its peak performance at 61 threads for smaller trees and at 122 and 183 threads for larger trees. Performance comparison of these algorithms is not shown here as we are more interested in studying the impact of our local recovery algorithm on each of these algorithms.

For uniform distribution, the performance gain is marginal and, in many cases, is actually slightly worse due to the overhead of stack maintenance (graphs are provided in the appendix). This is not surprising because, for small trees, even though contention is higher, seek time is small to begin with and any benefit of local recovery is nullified by additional overhead of stack maintenance. For larger trees, even though seek time is larger, contention is low as key accesses are spread evenly.

Figure 6.5 shows the behavior for Zipfian distribution with  $\alpha=1$ . In general, Zipfian distribution causes more contention than uniform distribution. So even for small trees for which seek times are smaller, we still see performance gains for mixed and write-dominated workloads. Figure 6.6 shows how seek time improves (reduces) due to local recovery.

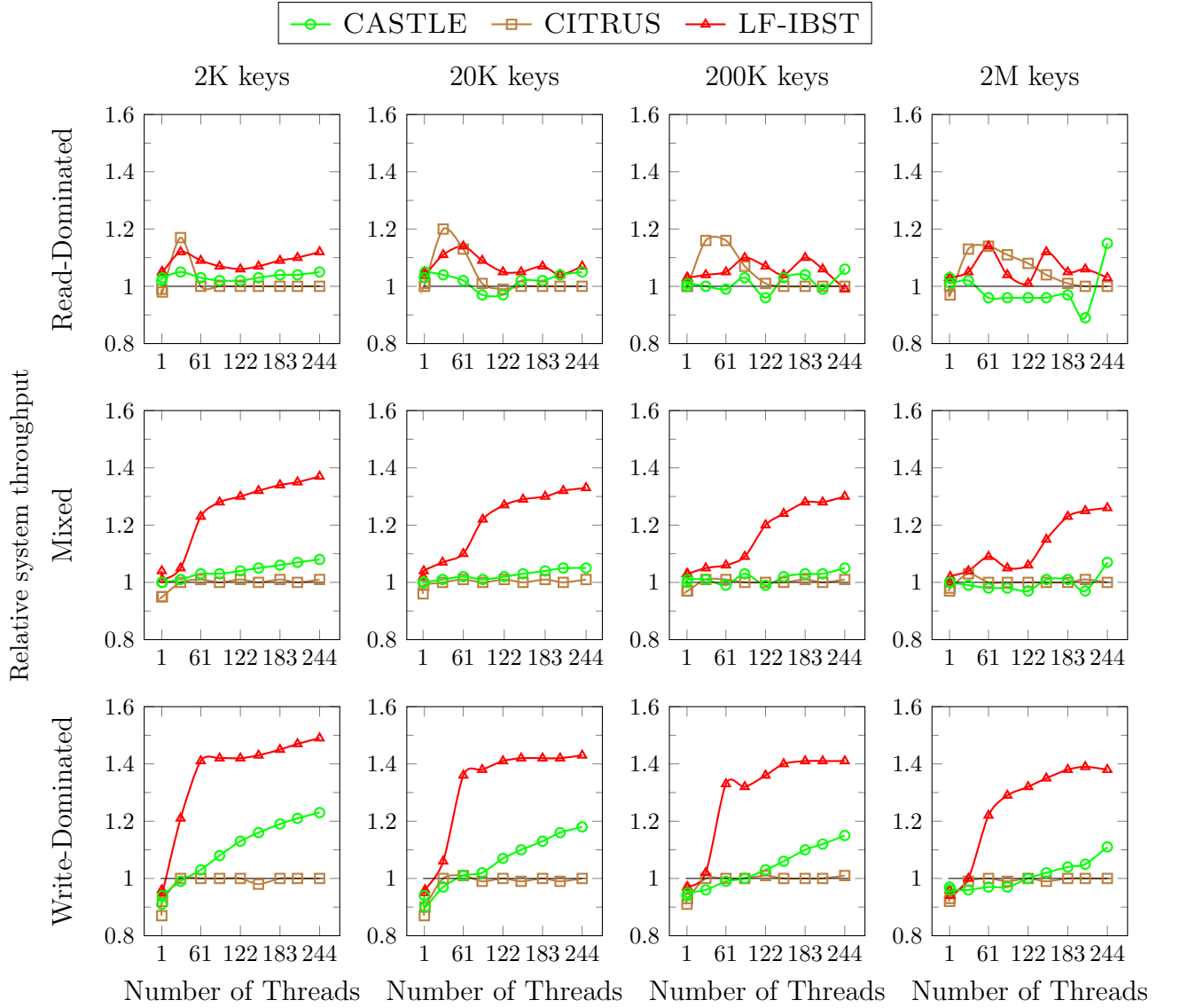


Figure 6.5: Effect of local recovery on system throughput when varying the number of threads from 1 to 244 for zipf distribution. Each row represents a workload type and each column represents a key space size. Higher the relative throughput, better the performance of the algorithm with local recovery.

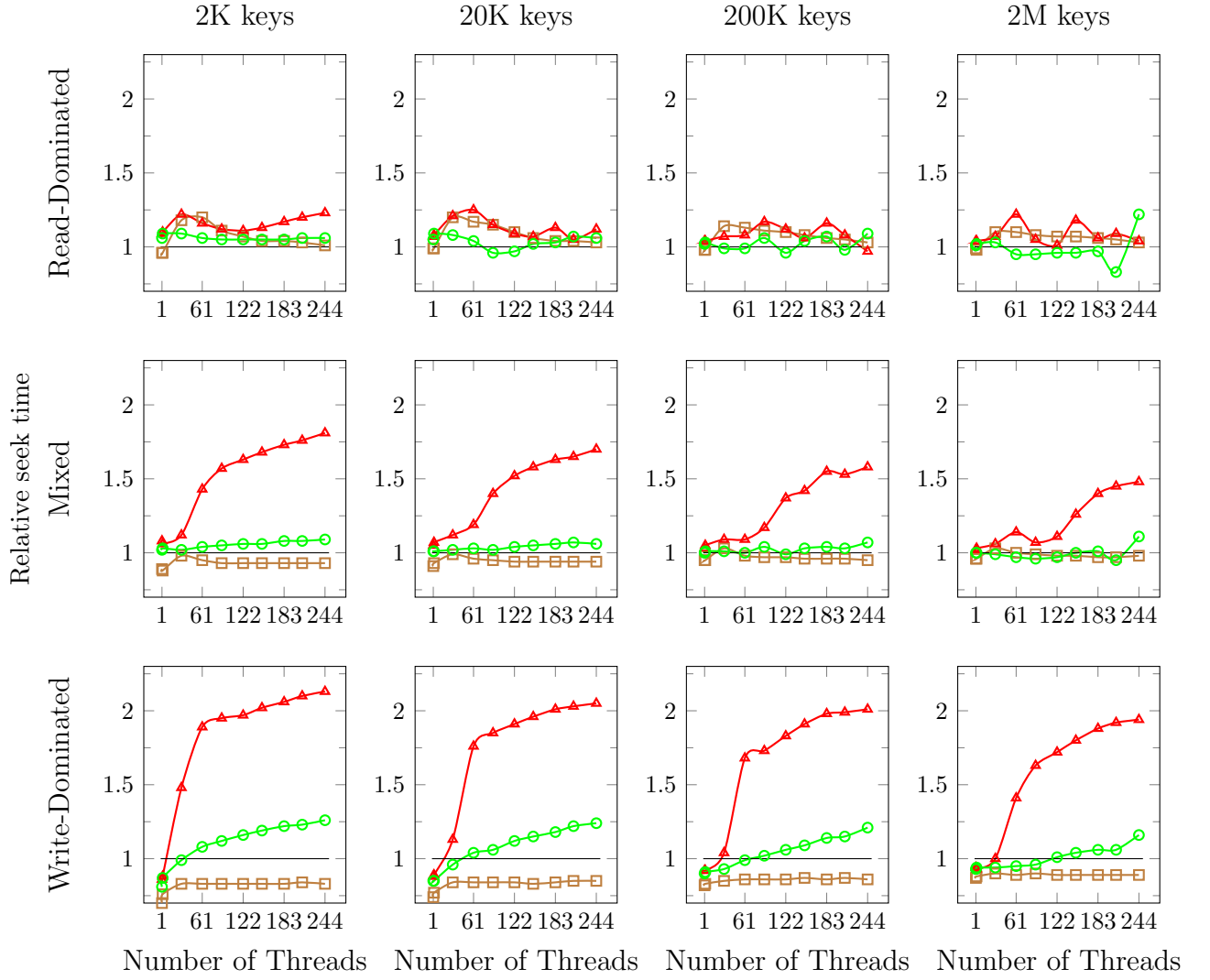


Figure 6.6: Effect of local recovery on seek time when varying the number of threads from 1 to 244 for zipf distribution. Each row represents a workload type and each column represents a key space size. Higher the relative seek time, better the performance of the algorithm with local recovery.

From Figures 6.5 & 6.6, we see a clear correlation between seek time and system throughput. As seek time reduces due to local recovery, the system throughput also improves. We see maximum improvement for LF-IBST. Since CASTLE and CITRUS are lock-based algorithms, no helping is performed during tree traversal. But in LF-IBST, during the tree traversal, if a pending operation is seen it is helped and then the current operation is restarted. This results in frequent restarts and hence local recovery improves performance by a larger margin.

Table 6.3: Effect of local recovery on system throughput. Positive number indicates a gain while a negative number indicates a drop.

<b>Algorithm</b>	<b>Uniform (max%, min%)</b>	<b>Zipfian (max%, min%)</b>
<b>CASTLE</b>	(6, -7)	(23, -11)
<b>CITRUS</b>	(23, -11)	(20, -13)
<b>LF-IBST</b>	(9, -6)	(49, -6)

Table 6.3 summarizes the performance gap (with respect to system throughput) between the base algorithm and its extension using local recovery for uniform and Zipfian distributions.

## CHAPTER 7

### CONCLUSION

In this dissertation we presented a blocking and a non-blocking algorithm for concurrent manipulation of a binary search tree in an asynchronous shared memory system that supports search, insert and delete operations.

Our *lock-based* algorithm is very simple and looks almost identical to a sequential algorithm. In contrast to other lock-based algorithms, it locks edges rather than nodes. This minimizes the contention window of an operation and improves the system throughput. Since the locks are based on edges and as we steal bits from the children edges, the tree node structure is identical to a sequential tree node. This keeps the memory foot print low and reduces the impact of *memory-allocation*. A desirable feature of this algorithm is that its search and insert operations are lock-free; they do not obtain any locks. As indicated by our experiments, our algorithm has the best performance—compared to other concurrent algorithms for a binary search—when the contention is relatively low. Specifically, it achieved the best performance for medium-sized and larger trees with mixed workloads and read-dominated workloads.

Our *lock-free* algorithm combined ideas from two existing lock-free algorithms and is especially *optimized for the conflict-free scenario*. Specifically, when compared to modify operations existing internal binary search trees, its modify operations (a) have a smaller contention window, (b) allocate fewer objects, (c) execute fewer atomic instructions, and (d) have a smaller memory footprint. Our experiments indicated that our new lock-free algorithm outperforms other lock-free algorithms in most cases.

We also presented a new approach to recover from such failures more efficiently in a concurrent binary search tree based on internal representation using *local recovery* by restarting

the traversal from the “middle” of the tree in order to locate an operation’s window. Our approach is sufficiently general in the sense that we were able to apply it to a variety of concurrent binary search trees based on both blocking and non-blocking approaches.

We also presented a framework to allow a concurrent algorithm for maintaining an internal BST to recover locally when traversing the tree to locate a key. Our framework is sufficiently general that we were able to apply it to a variety of concurrent binary search trees based on both blocking and non-blocking approaches. we showed by experiments that our local recovery framework improved the performance of concurrent BST algorithms under non-uniform key distribution (*e.g.*, Zipfian) for many different workloads.

As a future work, we would like to analyze our local recovery algorithm (and possibly refine it if needed) so that, when applied to a *non-blocking* BST, it yields a concurrent BST whose operations have provably low amortized time complexity. Also, we would like to analyze the effect of local recovery for other standard non-uniform distributions like normal and Poisson on real workloads.

We also would like to extend the ideas used in our algorithms and our local recovery technique to other data structures. A simple extention would be to apply them to  $k$ -ary search trees and then extend it further to  $B$ -trees. We also plan to explore other data structures like Bloom-filters which are commonly used in big-data applications.



## REFERENCES

- Adamic, L. A. and B. A. Huberman (2002). Zipfs Law and the Internet. *Glottometrics* 3(1), 143–150.
- Arbel, M. and H. Attiya (2014, July). Concurrent Updates with RCU: Search Tree as an Example. In *Proceedings of the 33rd ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 196–205.
- Breslau, L., P. Cao, L. Fan, G. Phillips, and S. Shenker (1999, March). Web Caching and Zipf-like Distributions: Evidence and Implications. In *Proceedings of the 18th IEEE Conference on Computer Communications (INFOCOM)*, pp. 126–134.
- Chatterjee, B., N. N. Dang, and P. Tsigas (2014). Efficient Lock-Free Binary Search Trees. In *Proceedings of the 33rd ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 321–331.
- Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 661–703.
- Cormen, T. H., C. E. Leiserson, and R. L. Rivest (1991). *Introduction to Algorithms*. The MIT Press.
- Drachsler, D., M. Vechev, and E. Yahav (2014, February). Practical Concurrent Binary Search Trees via Logical Ordering. In *Proceedings of the 19th ACM Symposium on Principles and Practice of Parallel Programming (PPoPP)*, pp. 343–356.

- Ellen, F., P. Fataourou, E. Ruppert, and F. van Breugel (2010, July). Non-Blocking Binary Search Trees. In *Proceedings of the 29th ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 131–140.
- Ellen, F., P. Fatourou, J. Helga, and E. Ruppert (2014). The Amortized Complexity of Non-Blocking Binary Search Trees. In *Proceedings of the 33rd ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 332–340.
- Faloutsos, C. and H. V. Jagadish (1992, August). On B-tree Indices for Skewed Distributions. In *Proceedings of the 18th International Conference on Very Large Data Bases (VLDB)*, pp. 364–373.
- Gray, J., P. Sundaresan, S. Englert, K. Baclawski, and P. J. Weinberger (1994, May). Quickly Generating Billion-Record Synthetic Databases. In *Proceedings of the 23rd ACM SIGMOD International Conference on Managment of Data*, pp. 243–252.
- Heller, S., M. Herlihy, V. Luchangco, M. Moir, W. N. S. III, and N. Shavit (2005). A Lazy Concurrent List-Based Set Algorithm. In *Proceedings of the 9th International Conference on Principles of Distributed Systems (OPODIS)*, Pisa, Italy, pp. 3–16.
- Herlihy, M. and N. Shavit (2012). *The Art of Multiprocessor Programming, Revised Reprint*. Morgan Kaufmann.
- Herlihy, M. and J. M. Wing (1990, July). Linearizability: A Correctness Condition for Concurrent Objects. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 12(3), 463–492.
- Howley, S. V. and J. Jones (2012, June). A Non-Blocking Internal Binary Search Tree. In *Proceedings of the 24th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*, pp. 161–171.

- Lea, D. (2003). Java Community Process, JSR 166, Concurrent Utilities. <http://gee.cs.oswego.edu/dl/concurrency-interest/index.html>.
- Lev, Y., M. Herlihy, V. Luchangco, and N. Shavit (2007, June). A Simple Optimistic Skiplist Algorithm. In *Proceedings of the 14th International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, Castiglioncello, Italy, pp. 124–138.
- Michael, M. M. (2002). High Performance Dynamic Lock-Free Hash Tables and List-based Sets. In *Proceedings of the 14th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*, pp. 73–82.
- Michael, M. M. (2004). Hazard Pointers: Safe Memory Reclamation for Lock-Free Objects. *IEEE Transactions on Parallel and Distributed Systems (TPDS)* 15(6), 491–504.
- Michael, M. M. and M. L. Scott (1996). Simple, Fast, and Practical Non-Blocking and Blocking Concurrent Queue Algorithms. In *Proceedings of the 15th ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 267–275.
- Natarajan, A. and N. Mittal (2014, February). Fast Concurrent Lock-Free Binary Search Trees. In *Proceedings of the 19th ACM Symposium on Principles and Practice of Parallel Programming (PPoPP)*, pp. 317–328.
- Ramachandran, A. and N. Mittal (2015a, January). A Fast Lock-Free Internal Binary Search Tree. In *Proceedings of the 16th International Conference on Distributed Computing and Networking (ICDCN)*.
- Ramachandran, A. and N. Mittal (2015b, February). CASTLE: Fast Concurrent Internal Binary Search Tree using Edge-Based Locking. In *Proceedings of the 20th ACM Symposium on Principles and Practice of Parallel Programming (PPoPP)*, pp. 281–282.

Reindeers, J. (2007). *Intel Threading Building Blocks: Outfitting C++ for Multi-Core Processor Parallelism*. O'Reilly Media, Inc.

## **VITA**

Arunmoezhi Ramachandran was born in Madurai, India and was brought up in Rajapalayam, India. He completed his schooling in 2003. He pursued his Bachelor's degree in Electronics and Communication at College of Engineering Guindy, Chennai. Then he worked at Infosys Technologies Ltd, Chennai as a software engineer from 2007 to 2010. Meanwhile he also pursued his Masters in Software Engineering at Birla Institute of Technology and Science, Pilani. Upon graduating with a Masters degree in 2011, he joined the PhD program in Computer Science at University of Texas, Dallas, USA.