

TECHNIQUES FOR CONSTRUCTING EFFICIENT LOCK-FREE DATA STRUCTURES

by

Trevor Brown

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

© Copyright 2017 by Trevor Brown

Abstract

Techniques for Constructing Efficient Lock-free Data Structures

Trevor Brown

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2017

Building a library of concurrent data structures is an essential way to simplify the difficult task of developing concurrent software. Lock-free data structures, in which processes can *help* one another to complete operations, offer the following progress guarantee: If processes take infinitely many steps, then infinitely many operations are performed. Handcrafted lock-free data structures can be very efficient, but are notoriously difficult to implement. We introduce numerous tools that support the development of efficient lock-free data structures, and especially trees.

We address the difficulty of using single-word compare-and-swap (CAS) to develop lock-free graph-based data structures by introducing multiword synchronization primitives called LLX and SCX. These primitives fall between multi-compare-single-swap and full multiword-CAS in expressiveness, and can be implemented much more efficiently than multiword-CAS. We use them to implement a tree update template that can be followed to produce lock-free implementations of arbitrary updates in down-trees, and use this template to produce the first lock-free implementations of several advanced trees: chromatic search trees, relaxed AVL trees, and relaxed (a,b) -trees. We also introduce a new variant of a B-tree, called a relaxed B-slack tree, which has significantly better worst-case space complexity, and produce a lock-free implementation using our template.

In these implementations, operations dynamically allocate memory for *nodes*. Additionally, in our implementation of LLX and SCX, each SCX operation allocates a *descriptor*, which contains information that another process can use to *help* the SCX operation complete. These implementations must perform dynamic lock-free memory reclamation, but traditional lock-free memory reclamation algorithms are either inefficient or cannot be used with our implementations. So, we introduce a fast epoch-based reclamation (EBR) algorithm called DEBRA+, which is the first lock-free EBR algorithm that can be used with trees.

We also devise two techniques for accelerating lock-free data structure implementations. Using the first technique, we can efficiently eliminate dynamic allocation and reclamation of *descriptors* in SCX, and a large class of other algorithms. Using the second, we exploit hardware transactional memory support in modern processors to produce accelerated implementations of the template.

Acknowledgements

According to an old African proverb, it takes a village to raise a child. Similarly, it takes an academic community to train a researcher. Over the last few years, I have been fortunate to be surrounded by many brilliant, friendly and helpful people. I would first like to express my gratitude to my supervisor Faith Ellen. She taught me the importance of laying a rigorous theoretical foundation, completely transformed my writing, encouraged me when I would otherwise have given up on a hard problem, and continually surprised me with her keen insights. I would also like to thank Eric Ruppert, who served as a mentor and research supervisor to me during my undergraduate studies, steered me towards research, and suggested that I study under Faith.

Faith and Eric have both been excellent collaborators, as well as mentors, and they were instrumental in the development of the LLX and SCX primitives and the tree update template. During an undergraduate research project, Ken Hoover used the tree update template to develop the lock-free relaxed AVL tree, and did some initial work on the lock-free relaxed (a, b) -tree, and I thank him for his help. I would also like to acknowledge Maya Arbel-Raviv for her part in the work on weak descriptors, and thank her for our many interesting discussions.

I am indebted to the members of my supervisory committee and thesis defense committee, namely, Azadeh Farzan, Vassos Hadzilacos, Maurice Herlihy, Ryan Johnson and Sam Toueg, for the time and energy they invested. Their feedback has also been very useful, and has led to numerous new insights. Most notably, discussions with Ryan led to the crash recovery mechanism in DEBRA+. I also thank Ryan for taking the time to explore parts of the Linux kernel with me, and for helping me debug hardware transactional memory issues.

The experiments in this work would not have been possible without several large-scale systems, which were graciously provided by Oracle, the University of Rochester, the Technion (Israel Institute of Technology), and the University of Toronto. Additionally, I was able to spend much less time implementing competing algorithms thanks to researchers who published their code, as well as those who were kind enough to send me their code, namely Maya Arbel-Raviv, Alex Matveev and Shahar Timnat. Funding for my research was provided by the Natural Sciences and Engineering Research Council of Canada.

Of course, I have to mention my parents and siblings, who have been an unfailing source of joy and warmth, regardless of the challenges in their own lives. Finally, I thank the love of my life, Brittany Trueman, for taking this ride with me. She has stood by me despite all of the late nights and deadline crunch-times, and serves as a gentle reminder that there is more to life than work.

Contents

1	Introduction	1
2	Model	8
3	Lock-free synchronization primitives	12
3.1	Related work	13
3.2	The primitives	15
3.2.1	Correctness properties	16
3.2.2	Progress properties	16
3.3	Implementation of primitives	17
3.3.1	Constraints	19
3.3.2	Detailed algorithm description and sketch of proofs	20
3.3.3	Additional properties	24
3.4	Formal proof	24
3.4.1	Basic properties	26
3.4.2	Changes to the <i>info</i> field of a Data-record and the <i>state</i> field of an SCX-record	27
3.4.3	Proving <i>state</i> and <i>info</i> fields change as described in Fig. 3.3	29
3.4.4	The period of time over which a Data-record is frozen	32
3.4.5	Properties of update CAS steps	34
3.4.6	Freezing works	37
3.4.7	Correctness of HELP	38
3.4.8	Linearizability of LLX/SCX/VLX	41
3.4.9	Progress guarantees	45
3.5	Additional properties of LLX/SCX/VLX	54
3.6	Modifications to enable memory reclamation	55
3.6.1	Garbage collection	55
3.6.2	Unmanaged languages	58
3.6.3	Reachability from nodes in the data structure	58
3.6.4	Reachability from private pointers	60

4 Multiset implemented with LLX/SCX	61
4.1 Implementation	62
4.2 Correctness and progress	62
4.2.1 Proof sketch	62
4.2.2 Complete proof	64
5 A template for implementing trees	71
5.1 Related work	73
5.2 LLX, SCX and VLX primitives	74
5.3 Tree update template	75
5.4 Correctness proof	79
5.5 Progress proof	83
6 Chromatic tree implemented with the template	86
6.1 Chromatic trees	87
6.2 Implementation	88
6.2.1 Detailed description of insertion	91
6.2.2 Detailed description of deletion	94
6.2.3 The rebalancing algorithm	95
6.2.4 Deciding which rebalancing step to perform	96
6.2.5 Implementing a rebalancing step	101
6.2.6 Successor queries	102
6.3 Correctness proof	102
6.4 Progress proof	108
6.5 Bounding the height of the tree	109
6.6 Allowing more violations	115
6.7 Experimental results	116
6.8 Summary	119
7 Relaxed AVL tree implemented with the template	121
7.1 AVL trees	122
7.2 Relaxed AVL trees	122
7.3 Implementation overview	124
7.4 Towards a height bound	125
8 Relaxed (a,b)-tree implemented with the template	128
8.1 (a,b) -trees and the relaxation	129
8.2 Implementation	132
8.2.1 Detailed description of insertion	133
8.2.2 Detailed description of deletion	133
8.2.3 The rebalancing algorithm	136
8.2.4 Implementing a rebalancing step	136

8.3	Correctness proof	139
8.4	Progress proof	143
8.5	Bounding the height of the tree	143
8.6	Experiments	148
9	B-slack trees: space efficient B-trees	151
9.1	Related work	152
9.2	B-slack trees	155
9.2.1	Relaxed B-slack trees	155
9.2.2	Updates to relaxed B-slack trees	156
9.2.3	Rebalancing steps	156
9.3	Analysis	159
9.3.1	Analysis of overslack trees	160
9.3.2	Relating B-slack trees to overslack trees	162
9.3.3	Amortized logarithmic rebalancing	164
9.4	B-slack trees with amortized constant rebalancing	167
9.5	Space complexity of competing trees	169
9.6	Sequential experiments	170
9.7	Implementation issues for rebalancing	171
10	Relaxed B-slack trees implemented with the template	173
10.1	Implementation	174
10.1.1	Insertion	174
10.1.2	Deletion	177
10.1.3	The rebalancing algorithm	177
10.1.4	Fixing weight violations	178
10.1.5	Fixing degree and slack violations	181
10.1.6	An optimization to avoid unnecessary searches while rebalancing	188
10.2	Correctness proof	188
10.3	Progress proof	193
10.4	Balance proof	195
10.5	Modifications for amortized constant rebalancing	199
10.5.1	Adding a range query operation	200
10.6	Experiments	200
10.6.1	Steady-state performance	202
10.6.2	Tree building performance	204
10.6.3	Memory usage of the final trees	205
10.6.4	Workloads with range queries	206

11 Reclaiming memory	209
11.1 Related work	212
11.2 DEBRA: Distributed Epoch Based Reclamation	222
11.3 Adding fault tolerance	226
11.4 A lock-free memory management abstraction	230
11.5 Experiments	233
11.6 Summary	236
12 Reusing descriptors	239
12.1 Wasteful Algorithms	242
12.1.1 Immutable descriptors	243
12.1.2 Mutable descriptors	245
12.2 Weak descriptors	247
12.2.1 Weak descriptor ADT	247
12.2.2 Transforming a class of algorithms to use the weak descriptor ADT	248
12.3 Extended Weak Descriptors	250
12.3.1 Example Algorithm: k-CAS	252
12.3.2 Example Algorithm: LLX and SCX	253
12.4 Implementing the extended weak descriptor ADT	255
12.5 Experiments	260
12.5.1 <i>k</i> -CAS microbenchmark	261
12.5.2 BST microbenchmark	265
12.5.3 Studying sequence number wraparound	265
12.6 Related Work	267
12.7 Summary	268
13 Accelerating the template with HTM	269
13.1 HTM-based LLX and SCX	273
13.1.1 Correctness and Progress	274
13.2 Accelerated template implementations	281
13.2.1 The 2-path <i>con</i> algorithm	281
13.2.2 The TLE algorithm	281
13.2.3 The 2-path <i>con̄</i> algorithm	282
13.2.4 The 3-path algorithm	282
13.3 Example data structures	283
13.3.1 Unbalanced BST	283
13.3.2 Relaxed (a,b) -tree	284
13.4 Experimental results	285
13.4.1 Light vs. Heavy workloads	285
13.4.2 Code path usage and abort rates	287
13.4.3 Comparing with hybrid transactional memory	288

13.5	Modifications for performing searches outside of transactions	290
13.6	Reclaiming memory more efficiently	290
13.7	Related work	291
13.8	Other uses for the <i>3-path</i> approach	292
13.8.1	Accelerating data structures that use read-copy-update (RCU)	292
13.8.2	Accelerating data structures that use <i>k</i> -CAS	294
13.9	Summary	294
	Bibliography	296

Chapter 1

Introduction

Building a library of concurrent data structures is an essential way to simplify the difficult task of developing concurrent software. One way to implement concurrent data structures is to use *locks*. Locks protect shared resources. Concurrent data structures can be implemented with coarse-grained locking, where a single lock protects the entire data structure, or with fine-grained locking, where there are many locks, each protecting a small part of the data structure (e.g., a tree where each node has an associated lock). When a process acquires a lock, it obtains exclusive access to the shared resource protected by the lock.

Lock-based programming is fairly simple, and there are many concurrent lock-based data structures. As a simple example, consider an operation for inserting a key into a binary search tree. We describe one way this operation could be implemented using fine-grained locking. The operation first locks the root node, and then searches for the location where a key should be inserted using hand-over-hand locking: each time it follows a child pointer, it locks the node it reaches before accessing its contents, and then subsequently unlocks the parent node. Once the operation reaches the node where the key should be inserted, it holds locks on that node, and its parent. The operation performs the appropriate modification to the tree, and unlocks those two nodes. It is straightforward to argue that these insertion operations are atomic. However, they are also *inefficient*.

The locking performed by searches is extremely costly, for several reasons. First, the cost of acquiring locks can be very high compared to the cost of reading a node's key and child pointers. Second, locking the root, and other nodes near the top of the tree, can severely limit concurrency. Reader-writer locks (which allow a lock to be acquired by a single writer or multiple readers) can be used to improve concurrency, but they are even more costly to acquire and release. Third, locking nodes that are not modified by an operation essentially turns reads into writes, which negatively impacts cache performance, especially on systems with non-uniform memory architectures.

Locks also have several other downsides. Most notably, if a process crashes while holding a lock, then it can prevent *all* processes from making progress. Additionally, locks can be susceptible to deadlock, convoying and priority inversion [56]. Deadlock occurs when two processes attempt to lock the same resources, but in different orders, and they mutually prevent one another from making progress. Convoying occurs when many threads attempt to acquire the same lock, and processes sleep while waiting for the lock to become free. Each time the lock becomes free, all of the processes wake up, attempt to acquire the lock, and, if unsuccessful, go back to sleep. This corresponds to a long convoy of processes moving from the waiting queue (where processes reside when they are not yet ready to run) to the run queue (where they wait to be executed on a processor) and back again. This movement of processes between scheduler queues can cause significant overhead. Priority inversion occurs when a process that is given high priority by the scheduler attempts to acquire a lock that is held by a low priority process, but must wait for the low priority process to release the lock before it can acquire the lock and make progress.

Consequently, it is often preferable to use hardware synchronization primitives like compare-and-swap (CAS) instead of locks. However, the difficulty of this task has inhibited the development of *lock-free* (also called *non-blocking*) data structures. These are data structures which guarantee that the system as a whole will continue to make progress even if some processes crash. In other words, they guarantee that, at all times, some operation will eventually complete. (Naturally, these data structures cannot use locks.) This progress guarantee is typically achieved with *helping*. When a process is blocked by an operation performed by another process, it helps the other process to make progress before continuing its own operation.

Direct implementations of lock-free data structures from CAS are notoriously complicated. Many operations on interesting data structures access, or even modify, multiple words in memory. With locks, one could achieve atomicity simply by acquiring locks on all of the relevant words in memory, and then performing the appropriate modification.

However, CAS operates atomically on only a *single* word in memory. As a result, things that were straightforward with locks, such as ensuring that a node in a tree is not changed at the same time as it is deleted (or thereafter), become very difficult with CAS. Thus, complex ad-hoc synchronization mechanisms are needed to guarantee atomicity.

We briefly discuss why these synchronization mechanisms are so complex. In a lock-free algorithm, *processes* cannot acquire exclusive access to any shared resource, or else the resource could become permanently inaccessible in the event of a process crash. Consequently, instead of having a *process* acquire exclusive access to shared resources, an *operation* acquires exclusive access to shared resources, and any processes helping the operation can access these resources. It is important to note that multiple processes can simultaneously help the same operation. (To ensure that an operation could be helped by only one process at a time, a process would need to acquire exclusive access to a shared resource, which we have argued cannot happen.) Therefore, the helping algorithm must be designed so that multiple processes helping the same operation do not perform conflicting changes or erroneously repeat algorithmic steps (and it must guarantee this *without* having any process acquire exclusive access to any shared resource).

Since direct implementations of lock-free data structures from CAS are so complicated, it is extremely difficult to formally prove correctness and progress for them. Moreover, when these proofs are actually produced, they are nearly impossible to check. Consequently, most implementations are presented with only brief sketches of correctness and progress. Unfortunately, these sketches often turn out to be incorrect.

Early on, researchers recognized that synchronization primitives which atomically access multiple locations make the design of lock-free data structures much easier [17, 72, 114]. In Chapter 3, we introduce three new primitives, *load-link-extended* (LLX), *validate-extended* (VLX) and *store-conditional-extended* (SCX), which are natural generalizations of the well known *load-link* (LL), *validate* (VL) and *store-conditional* (SC) primitives. We carefully designed our primitives to strike a fine balance between ease of use and efficient implementability. We provide a practical implementation of our primitives from CAS, and give complete proofs of correctness and progress. In Chapter 4, we demonstrate their use by giving a simple implementation of a multiset from a singly-linked list, along with a full proof.

As we were developing our new primitives, there was a flurry of interest in lock-free trees. In particular, numerous papers appeared containing incredibly complicated ad-hoc implementations of lock-free balanced search trees. These implementations either had only brief correctness arguments or incredibly long proofs (with one exceeding 100 pages). The authors of these papers never released source code for their implementations. With the goal of providing a rigorous, provably correct alternative, we introduced a *tree update template* that can be followed to produce lock-free implementations of down-trees (trees in which all nodes except the root have in-degree one) with any kinds of update operations. This template, which appears in Chapter 5, uses our LLX and SCX primitives.

We demonstrate the use of our template by producing lock-free implementations of several data structures that are significantly more advanced than any lock-free data structures in the literature. Java and/or C++ code for all of our implementations is publicly available.¹

In Chapter 6, we present a lock-free implementation of a chromatic search tree, which is a generalization of a red-black tree (a kind of balanced binary search tree (BST)). The chromatic tree is a highly advanced *relaxed balance* tree. In contrast to traditional balanced trees, which require an insertion or deletion and any necessary rebalancing to be performed as one large atomic update (potentially involving an entire root-to-leaf path), relaxed balance trees *decouple* rebalancing from insertions and deletions. Specifically, rebalancing is performed as a sequence of small atomic *rebalancing steps* that can be interleaved freely with insertions, deletions, and other rebalancing steps. This

¹<http://implementations.tbrown.pro>

significantly improves concurrency. However, as a consequence, the chromatic tree can transiently become slightly less balanced than a red-black tree, and there are more cases for rebalancing to handle. In fact, there are eleven different types of rebalancing steps (each with a symmetric mirror-image version). Despite the complexity of this data structure, we provide a rigorous proof of correctness and progress that is both concise (only six pages long) and easily checked. A wide range of experiments on a large scale system showed that our implementation was significantly faster than the leading competitors.

In Chapter 7, we give a high level description of a lock-free relaxed AVL (RAVL) tree, which is a relaxed balance generalization of an AVL tree. As with the chromatic tree, decoupling rebalancing so that it can be performed in small atomic rebalancing steps makes RAVL trees more concurrency friendly, but creates more cases for rebalancing. RAVL trees provide eleven different rebalancing steps (each with a symmetric mirror-image version). Experiments show that an optimized version of RAVL trees performs as well as chromatic trees.

A B-tree of minimum degree $a \geq 2$ is a balanced tree in which all leaves have the same depth, and all nodes contain between a and $2a - 1$ keys. Whenever an operation would cause a node to contain more than $2a - 1$ or fewer than a keys, the tree is rebalanced by *splitting* the node or *joining* it with a sibling, respectively (which may necessitate splitting or joining at a sequence of ancestors). The *capacity* of a node is the maximum number of keys it can contain. The *utilization* of a node is k/c , where k is the number of keys it contains and c is its capacity.

In Chapter 8, we present a lock-free implementation of a relaxed (a,b) -tree, which is a relaxed balance generalization of a B-tree. In a standard (a,b) -tree, all leaves have the same depth and contain between a and b keys, where $a \geq 2$ and $b \geq 2a - 1$. Observe that the worst case utilization of a node in an (a,b) -tree is the same as in a B-tree when $b = 2a - 1$, and it decreases as b grows relative to a . Allowing nodes to contain fewer keys, proportional to their capacity, can reduce the number of rebalancing steps needed to maintain the structural properties of the tree. Relaxed (a,b) -trees allow the structural properties of (a,b) -trees to be violated in specific ways, and provide six different rebalancing steps to fix these violations, with the ultimate goal of transforming a relaxed (a,b) -tree into a standard (a,b) -tree. We provide a rigorous proof of correctness and progress for our implementation in just five pages. This is in stark contrast to the lock-free B-tree of Braginsky and Petrank [24], which has a 33 page proof, despite being a simpler data structure than an (a,b) -tree. Experiments show that the relaxed (a,b) -tree has a significant performance advantage over the chromatic tree for certain workloads.

When chromatic trees, RAVL trees and relaxed (a,b) -trees were proposed, it was assumed that rebalancing steps would be completely decoupled from insertion and deletion, so there would be no worst-case upper bound on the height of these trees. For each tree, we introduce a simple algorithm for performing rebalancing steps that yields strong upper bounds on their heights. We prove that the height of the chromatic tree is at most $O(c + \log n)$, where c is the number of insertions and deletions currently in progress. We sketch a proof that the RAVL tree has the same upper bound on its height. We also prove that the height of the relaxed (a,b) -tree is at most $O(c + \log_a n)$. The proofs of these upper bounds are subtle, and unlike anything in the literature on lock-free trees or relaxed balance trees.

One downside of B-trees is that half of the capacity of each node is wasted in the worst case (and even more capacity is wasted in (a,b) -trees with $b > 2a - 1$). The most obvious consequence of wasting space in nodes is increased tree height. This increases search times, which are typically the dominant factor in the performance of search trees. However, there is another, more subtle consequence in environments where memory is allocated in blocks, with a limited set of block sizes. This is often the case in hardware, such as internet routers, where allocators are very simplistic, and typically use only a single block size to avoid fragmentation. In this case, if we allocate a block for each node, then half of all memory is wasted in the worst case. Fast memory is expensive, and hardware

developers are very interested in building devices with less memory. Since hardware must include sufficient resources to handle the worst case, it is *not* sufficient to design a data structure with good *average-case* behaviour—it must have good worst-case behaviour.

To address this problem, in Chapter 9, we introduce a novel data structure called a *B-slack tree*, which is a variant of a B-tree with substantially better worst-case space complexity. In a B-slack tree, all nodes contain between 0 and b keys, and internal nodes contain between 2 and b child pointers. The *degree* of an internal node (resp., leaf) is the number of pointers (resp., keys) it contains, and a node’s *slack* is $b - d$, where d is its degree. Slack represents the part of a node that is wasted. Rather than imposing strong constraints on the degree or slack of individual nodes, the key idea is to constrain each internal node so that the sum of the slack at its children is less than b . Surprisingly, this invariant is fairly straightforward to maintain, and it yields a tree with very good worst-case behaviour. In the worst case, the average degree of nodes is very high, exceeding $b - 2$ for trees of height at least three. A rigorous and thorough mathematical analysis of the characteristics of B-slack trees is presented. The space complexity of B-slack trees is significantly better than all of their competitors. We also introduced relaxed B-slack trees, which are a relaxed balance version of B-slack trees. Relaxed B-slack trees allow the properties of B-slack trees to be violated in specific ways, and provide six different rebalancing steps to fix these violations, with the ultimate goal of transforming a relaxed B-slack tree into a standard B-slack tree. The operations on relaxed B-slack trees are relatively simple and efficient. A Java implementation of a single-threaded B-slack tree is publicly available.

We present a provably correct lock-free implementation of a relaxed B-slack tree using our template in Chapter 10. This involved developing a new algorithm for determining when and where rebalancing steps should be applied, and proving that, when there are no ongoing updates, the tree is a (strict) B-slack tree. Experimental results show that an optimized variant of the relaxed B-slack tree yields significantly better performance than the chromatic tree for workloads consisting mostly of searches, and is reasonably efficient even for workloads with many updates.

In concurrent data structures that use locks, it is typically straightforward to free memory to the operating system after an object is removed from the data structure. For example, consider a set implemented with a singly-linked list using hand-over-hand locking. Recall that hand-over-hand locking allows a process to lock a node (other than the head node) only if it holds a lock on the node’s predecessor. To traverse from a locked node u to its successor v , the process first locks v , then it unlocks u . To delete a node u , a process first locks the head node, then performs hand-over-hand locking until it reaches u ’s predecessor and locks it. The process then locks u (to ensure no other process holds a lock on u), removes it from the list, frees it to the operating system, and finally unlocks u ’s predecessor. It is easy to argue that no other process has a pointer to u when u is freed.

In contrast, memory reclamation is one of the most challenging aspects of lock-free data structure design. The main difficulty in performing memory reclamation for a lock-free data structure is that a process can be sleeping while holding a pointer to an object that is about to be freed. Thus, carelessly freeing an object can cause a sleeping process to access freed memory when it wakes up, crashing the program or producing subtle errors. Since nodes are not locked, processes must coordinate to let each other know which nodes are safe to reclaim, and which might still be accessed. (Note that reclaiming memory is similarly challenging for lock-based algorithms that have lock-free searches.)

Managed languages such as Java and C# have automatic garbage collection, which greatly simplifies the implementation of lock-free data structures. However, in unmanaged languages such as C and C++, programmers must manually implement lock-free memory reclamation. Prior to this work, existing lock-free memory reclamation schemes were either inefficient, or could not easily be used with algorithms implemented from LLX and SCX (or our template). To remedy this, we introduced DEBRA, a distributed epoch-based reclamation algorithm (Chapter 11). Experiments

show that DEBRA is highly efficient, even on systems with non-uniform memory architectures (NUMA). However, a process that crashes in the middle of an operation can prevent all processes from reclaiming memory. DEBRA is a good choice when process failures cannot occur, or in a lock-based data structure with lock-free searches. We also present a fault-tolerant version of DEBRA called DEBRA+, and describe a large class of lock-free data structures that can be used with DEBRA+.

Up to this point, we have discussed tools for designing and implementing lock-free data structures. However, there is also a need for techniques to *accelerate* lock-free data structure implementations, e.g., to obtain the fastest code for data structure libraries. We describe two techniques for doing so.

In many lock-free data structures (e.g., [52, 70, 113]), each time a process performs an operation, it first creates a *descriptor* that specifies the steps the process will take to perform the operation, and stores a pointer to the descriptor in the data structure so other processes can see it. Then, whenever a process is blocked by an operation, it uses the information stored in the operation’s descriptor to help it complete. Our implementation of LLX and SCX takes this approach. We call implementations that create a new descriptor for each operation *wasteful algorithms*.

In Chapter 12, we introduce two simple *descriptor* abstract data types (ADT) that attempt to capture how descriptors are used by wasteful algorithms (including our implementation of LLX and SCX): the immutable descriptor ADT (which represents descriptors whose fields are all immutable after they are first initialized), and the mutable descriptor ADT (which represents descriptors in which fields can be mutable).

Naturally, the descriptors used for helping must eventually be freed to the operating system, or reused. In many applications, it is crucial that the reclamation of descriptors imposes very little runtime overhead. Additionally, for some applications, such as embedded systems, it may be important to have a small, predictable number of descriptors in the system. Thus, we want to reclaim or reuse descriptors in a way that minimizes time overhead and *descriptor footprint*, i.e., the largest number of descriptors in the system at one time. (Having a smaller descriptor footprint can also significantly improve the performance of processor caches, since less cache space is occupied by descriptors.)

One approach is to use a memory reclamation scheme to reclaim descriptors once they are no longer needed. However, reclaiming descriptors for each operation can introduce significant overhead (since the memory reclamation scheme incurs overhead for each descriptor). Additionally, it is not always easy to implement memory reclamation for descriptors, so this approach can add considerable complexity. As a better alternative, we introduce a *weak descriptor* ADT that has slightly *weaker semantics* than the mutable descriptor ADT, but can be implemented *without memory reclamation*. We also identify a class of lock-free algorithms that use the descriptor ADT, and which can be *transformed* to use the weak descriptor ADT. We then present an extension to our weak descriptor ADT, and show how an even larger class of lock-free algorithms can be transformed to use this extension. We prove correctness and progress for the extended transformation, and demonstrate its use by transforming wasteful implementations of a k -compare-and-swap (k -CAS) primitive [62] and the LLX and SCX primitives in Chapter 3.

We use known techniques to produce an efficient, provably correct implementation of our extended weak descriptor ADT. With this implementation, the transformed algorithms for k -CAS, and LLX and SCX, have some desirable properties. In the original k -CAS algorithm, *each operation attempt* allocates at least $k + 1$ new descriptors. In contrast, the transformed algorithm allocates only two descriptors *per process, once, at the beginning of the execution*, and processes simply reuse these descriptors. Similarly, whereas, in the original algorithm for LLX and SCX, each SCX operation creates a new descriptor, the transformed algorithm allocates only one descriptor per process, at the beginning of the execution. Observe that this entirely eliminates dynamic allocation *and* memory reclamation for descriptors (significantly reducing overhead), and results in an extremely small descriptor footprint. Extensive ex-

periments show that our transformed implementations perform at least as well as the original implementations, and *significantly* outperform them (by up to 5x) in some workloads.

Another way to accelerate lock-free data structures is to take advantage of new hardware features. Recently, Intel introduced hardware transactional memory (HTM) in its processors. HTM allows a programmer to run blocks of code in transactions, which either commit and take effect atomically, or abort and have no effect on shared memory. HTM has the potential to improve the performance of handcrafted algorithms significantly. This is because hardware transactions are extremely fast, and they can be used to optimistically avoid using other, more expensive synchronization mechanisms. As a trivial example, a sequence of CAS instructions can be accelerated by replacing it with a transaction that performs reads, if-statements and writes. Note that this represents a non-standard use of HTM: we are *not* interested in its ease of use, but, rather, in its ability to reduce synchronization costs.

Although hardware transactions are extremely fast, it is surprisingly difficult to obtain the full performance benefit of HTM. Intel’s HTM implementation is best-effort, which means it does not guarantee that transactions will *ever* be able to commit. Even in a single threaded system, a transaction can repeatedly abort because of internal buffer overflows, page faults, interrupts, and many other events. So, to guarantee progress, any code that uses HTM must also provide a non-transactional *fallback path* to be executed if a transaction fails. The decision of whether to allow operations on the fallback path to run concurrently with hardware transactions profoundly impacts the overall performance of algorithms implemented using HTM. In order to support this concurrency, hardware transactions must be *instrumented* with code that synchronizes with operations on the fallback path. This can add significant overhead to hardware transactions, negating much of their benefit. However, if operations on the fallback path are not permitted to run concurrently with hardware transactions, then numerous other performance pathologies arise.

In Chapter 13, we explore this design space, developing three accelerated implementations of the tree update template which use two execution paths, and one which uses *three* execution paths to combine the benefits of the two-path algorithms while avoiding their downsides. We performed experiments to evaluate our new template algorithms by comparing them with the original template algorithm. In order to compare the different template algorithms, we used each algorithm to implement two data structures: an unbalanced BST and a relaxed (a,b) -tree. We then ran microbenchmarks to compare the performance (operations performed per second) of the different implementations in a variety of workloads. The results show that our new template algorithms offer significant performance improvements. For example, on an Intel system with 72 concurrent processes, our best implementation of the relaxed (a,b) -tree outperformed the implementation using the original template algorithm by an average of 410% over all workloads.

Chapter 2

Model

We consider an asynchronous shared memory system with n processes numbered one through n . Processes can run at arbitrarily different speeds, and the speeds of processes are not fixed. They can also experience crash failures (sometimes known as halting failures, or stop failures). In this model, a crashed process is indistinguishable from an extremely slow process. Each process has local memory that is not accessible by any other process, and there is a shared memory accessible by all processes.

Primitive objects Shared memory is divided into primitive objects, which have atomic operations that are provided directly by the hardware. Examples include read/write registers, compare-and-swap (CAS) objects, and double-wide compare-and-swap (DWCAS) objects.

A read/write register contains a value and provides read and write operations to retrieve the current value, and replace it with a new value, respectively. The read and write operations are atomic, and the read operation always returns the value written by the last write operation.

A CAS object is a register that also offers a CAS operation, which has two arguments: an expected value *exp* and a new value *new*. A CAS operation *atomically* does the following. It first reads the value in the CAS object. If the value is equal to *exp*, then the CAS operation stores *new* in the CAS object and returns TRUE. Otherwise, the CAS operation simply returns FALSE.

As a theoretical object, CAS is allowed to contain arbitrarily large values. However, in real systems, memory is organized as a finite sequence of words, which each have a finite set of possible values, and CAS operates on a *single word* in memory. For many applications, it turns out to be useful to have the ability to atomically perform CAS on *two adjacent words* in memory, and DWCAS was introduced to do exactly that. Although DWCAS offers no additional power over unbounded CAS in a theoretical sense, it is very useful in real systems, and is implemented on all modern Intel and AMD processors.

Configurations and executions A *configuration* of the system consists of the states of all processes, and the state of shared memory. In any given configuration, each process has a set of *steps* (operations on primitive objects) that it can perform, and this set is determined by the state of the process in that configuration. Note that a process may be able to perform a given step in one configuration, but not in another configuration. Performing a step can change the state of the process and/or the state of shared memory. An *execution* is a (possibly infinite) sequence of alternating configurations and steps $C_0 \cdot s_0 \cdot C_1 \cdot s_1 \cdot C_2 \cdot s_2 \dots$, where each C_i is a configuration, and each s_i is a step that can be performed by a process in configuration C_i . We say that a configuration is *reachable* if it appears in some execution.

Records and data structures A *record* is a collection of primitive objects, which we refer to as *fields*. Fields can contain pointers to other records. A data structure has a fixed set of *entry points*, which are pointers to records. A record is *in the data structure* if it is reachable by following pointers starting from an entry point. A record is *removed from the data structure* when it changes from being in the data structure to not being in the data structure. A record is *inserted into the data structure* when it changes from not being in the data structure to being in the data structure. Any fields of a record that cannot change while the record is in the data structure are said to be *immutable*. Immutable fields can be arbitrarily large. Other fields are said to be *mutable*. Each mutable field fits in a single machine word.

Allocating and freeing records The system has an memory *allocator* that provides operations to *allocate* and *free* records. Initially, all records in shared memory are *unallocated*. Accessing an unallocated record will cause program

failure. Allocating a record provides the process that requested it with a pointer to it, and makes it accessible by any process that has a pointer to it. A record can also be *freed*, which returns it to the *unallocated* state.

Memory hierarchy The memory is organized into a hierarchy where the lowest level is *main memory*. Without loss of generality, we consider the *cache line* granularity in main memory as the smallest unit of data. The next levels of the hierarchy are *cache* levels, which contain copies of cache lines that appear in main memory. The cache levels are numbered, L1, L2, L3 and so on, starting with the highest level (L1) and moving down the hierarchy. In practice, systems typically use between two and four cache levels. A cache coherence protocol ensures that processors see a consistent view of main memory despite the existence of multiple cached copies of some memory locations. At the highest level of the memory hierarchy are *registers*, special memory locations reserved in each processor for temporary computations. Generally, operations lower in the memory hierarchy are orders of magnitude slower than operations higher in the hierarchy.

We consider a modified-exclusive-shared-invalid (MESI) cache coherence protocol. Whenever a process p attempts to access a cache line, it checks each of its caches one-by-one, starting from the highest level, to see whether the cache line already resides in any of its caches. At each level of the cache hierarchy where p fails to find the desired cache line, we say that it experiences a *cache miss*. The last cache level before main memory is called the *last-level cache*, and a cache miss at that level is called a *last-level cache miss*. If p experiences a last-level cache miss, it must retrieve a copy of the cache line from main memory.

When reading from memory, a process loads a cache line into its cache in *shared* mode. Many processes can simultaneously have the same cache line in their private caches, provided that they all loaded it in shared mode. When writing to memory, a process loads a cache line into its cache in *exclusive* mode, and makes the change to its own cached copy (gradually flushing the change downward to all lower levels of the memory hierarchy). Loading a cache line in exclusive mode also causes any copies of that cache line in other process' caches to be *invalidated*. Whenever a process p tries to access a cache line that was invalidated since p last accessed it, p will experience a last-level cache miss.

Non-uniform memory architectures (NUMAs) Some of the systems studied in this work have non-uniform memory architectures, in which different parts of memory can have drastically different access times for different processes. We take a simplified view of NUMAs that captures the most important costs on many modern systems. Processes run on one or more *sockets*. All processes on a socket share the same last-level cache (but processes on one socket cannot access the last-level cache shared by processes on another socket). Writes by a process do **not** cause cache invalidations for any processes that share the same cache.

As an example, consider two processes p and q that are on the same socket (and, hence, share the same last-level cache). Suppose q loads a cache line, then p writes to it, and then q loads it again. Since p and q share the same last-level cache, when p performs its write, it simply modifies q 's copy that is already in the last-level cache, and does **not** invalidate it. However, p 's write will still invalidate any copies of the cache line in the higher level caches, and in the last-level caches of other sockets. Thus, the next time q accesses this cache line, it will use the copy in its last-level cache. In contrast, if p and q were on different sockets, then q 's second load would have to retrieve the cache line from main memory.

Hardware transactional memory Transactional memory allows a programmer to execute arbitrary blocks of code atomically as transactions. Each transaction either *commits* and appears to take effect instantaneously, or *aborts* and has no effect on shared memory. The set of memory locations read (resp., written) by a transaction is called its *read-set* (resp., *write-set*). The *data-set* of a transaction is the union of its read-set and write-set. If the write-set of a transaction intersects the data-set of another transaction, then the two transactions are said to *conflict*. When two transactions conflict, one of them must abort to ensure consistency.

Hardware support for transactional memory support has appeared in numerous commercially available processors, including Intel’s Haswell, Broadwell and Skylake microarchitectures, and recent IBM POWER, BlueGene/Q and zEC architectures. This support consists of instructions for starting, committing and aborting transactions, and various other platform specific offerings. In these HTM implementations, transactions can abort, not only because of conflicts, but also for other spurious reasons. These HTM implementations are *best-effort*, which means they offer no guarantee that any transaction will ever commit.¹ We consider Intel’s implementation of HTM.

Linearizability All of the implementations discussed in this work are *linearizable*. Linearizability is a correctness condition introduced by Herlihy and Wing [69]. A concurrent execution α is linearizable if linearization points can be selected for each completed operation, and for a subset of the operations that started but did not complete, such that the linearization point for an operation occurs during the operation, and the result of each completed operation in α is the same as it would be if the operations were executed atomically at their linearization points. An algorithm is linearizable if every concurrent execution is linearizable.

Progress conditions In this work, we consider lock-free (also called non-blocking) data structures. Lock-free data structures guarantee that, from every reachable system configuration, some operation will eventually complete even if some processes crash. However, individual operations may starve. Note that this rules out the use of locks, since a process that crashes while holding a lock can cause deadlock (potentially blocking all other processes). In contrast, wait-free data structures offer a stronger guarantee that *every* operation terminates after a finite number of steps (unless the process executing the operation crashes). This stronger guarantee typically comes at the cost of reduced performance and increased complexity over lock-free algorithms.

¹Technically, some IBM architectures also offer *constrained transactions*, which guarantee certain types of transactions will not fail spuriously. However, the constraints are highly restrictive. For example, a constrained transaction can contain at most 32 instructions, cannot use loops, and cannot write to more than four different cache lines. There are many additional restrictions.

Chapter 3

Lock-free synchronization primitives

The goal of this chapter is to facilitate the implementation of high-performance, provably correct, non-blocking data structures on any system that supports a hardware CAS instruction. We introduce three new operations, *load-link-extended* (LLX), *validate-extended* (VLX) and *store-conditional-extended* (SCX), which are natural generalizations of the well known *load-link* (LL), *validate* (VL) and *store-conditional* (SC) operations. We provide a practical implementation of our new operations from CAS, and give complete proofs of correctness and progress. We also show how these operations make the implementation of non-blocking data structures and their proofs of correctness substantially less difficult, as compared to using LL, VL, SC, and CAS directly.

LLX, SCX and VLX operate on *Data-records*. Any number of types of Data-records can be defined, each type containing a fixed number of *mutable* fields (which can be updated), and a fixed number of *immutable* fields (which cannot). Each Data-record can represent a natural unit of a data structure, such as a node of a tree or a table entry. A successful LLX operation returns a snapshot of the mutable fields of one Data-record. (The immutable fields can be read directly, since they never change.) An SCX operation by a process p is used to atomically store a value in one mutable field of one Data-record *and finalize* a set of Data-records, meaning that those Data-records cannot undergo any further changes. The SCX succeeds only if each Data-record in a specified set has not changed since p last performed an LLX on it. A successful VLX on a set of Data-records simply assures the caller that each of these Data-records has not changed since the caller last performed an LLX on it. A more formal specification of the behaviour of these operations is given in Section 3.2.

Early on, researchers recognized that operations accessing multiple locations atomically make the design of non-blocking data structures much easier [17, 72, 114]. Our new primitives do this in three ways. First, they operate on Data-records, rather than individual words, to allow the data structure designer to think at a higher level of abstraction. Second, and more importantly, a VLX or SCX can depend upon multiple LLXs. Finally, the effect of an SCX can apply to multiple Data-records, modifying one and finalizing others.

The precise specification of our operations was chosen to balance ease of use and efficient implementability. They are more restricted than multi-word CAS [72], multi-word RMW [5], or transactional memory [114]. On the other hand, the ability to finalize Data-records makes SCX more general than k -compare-single-swap [92], which can only change one word. We found that atomically changing one pointer and finalizing a collection of Data-records provides just enough power to implement numerous pointer-based data structures in which operations replace a small portion of the data structure. To demonstrate the usefulness of our new operations, in Chapter 4, we give an implementation of a simple, linearizable, non-blocking multiset based on an ordered, singly-linked list.

Our implementation has some desirable performance properties. A VLX on k Data-records only requires reading k words of memory. If SCXs being performed concurrently depend on LLXs of disjoint sets of Data-records, they all succeed. If an SCX encounters no contention with any other SCX and finalizes f Data-records, then a total of $k + 1$ CAS steps and $f + 2$ writes are used for the SCX and the k LLXs on which it depends. We also prove progress properties that suffice for building non-blocking data structures using LLX and SCX.

3.1 Related work

Transactional memory [68, 114] is a general approach to simplifying the design of concurrent algorithms by providing atomic access to multiple objects. It allows a block of code designated as a transaction to be executed atomically, with respect to other transactions. Our LLX/VLX/SCX primitives may be viewed as implementing a restricted kind of transaction, in which each transaction can perform any number of reads followed by a single write and then finalize

any number of words. It is possible to implement general transactional memory in a non-blocking manner (e.g., [56, 114]). However, at present, implementations of transactional memory in software incur significant overhead, and hardware transactional memory (HTM) implementations have significant limitations and are not widely available. So, there is still a need for more specialized techniques for designing shared data structures that combine ease of use and efficiency.

Most shared-memory systems provide CAS operations in hardware. However, LL and SC operations have often been seen as more convenient primitives for building algorithms. Anderson and Moir gave the first wait-free implementation of small LL/SC objects from CAS using $O(1)$ steps per operation [9]. See [74] for a survey of other implementations that use less space or handle larger LL/SC objects.

Many non-blocking implementations of primitives that access multiple objects use the *cooperative technique*, first described by Turek, Shasha and Prakash [120] and Barnes [17]. Instead of using locks that give a process exclusive access to a part of the data structure, this approach gives exclusive access to *operations*. If the process performing an operation that holds a lock is slow, other processes can *help* complete the operation and release the lock.

The cooperative technique was also used recently for a wait-free universal construction [34] and to obtain non-blocking binary search trees [52] and Patricia tries [113]. The approach used here is similar.

Israeli and Rappoport [72] used a version of the cooperative technique to implement multi-word CAS from single-word CAS (and sketched how this could be used to implement multi-word SC operations). However, their approach applies single-word CAS to very large words. The most efficient implementation of k -word CAS (k -CAS) [117] first uses single-word CAS to replace each of the k words with a pointer to a *descriptor* record containing information about the operation, and then uses single-word CAS to replace each of these pointers with the desired new value and update a *status* field of the descriptor. In the absence of contention, this takes $2k + 1$ CAS steps.

Whenever a Data-record r is removed from a data structure by a multi-word CAS, care must be taken to ensure that other processes do not concurrently update any field of r . Furthermore, other processes must not update any field of r *after* it is removed. The standard way to prevent these erroneous changes is to set a *marked* bit in r when it is removed, and to have each multi-word CAS depend on the *marked* bits of the nodes it accesses. Consider an operation that atomically marks k Data-records and changes a pointer to remove them from the data structure. This can be done with a $(k + 1)$ -CAS, or with an SCX that depends on LLXs of k Data-records. As we saw above, if this $(k + 1)$ -CAS is implemented using the most efficient multi-word CAS algorithm currently available, it will perform $2k + 3$ single-word CAS steps with no contention. In contrast, in our implementation, an SCX will perform only $k + 1$ CAS steps with no contention. So, our weaker primitives can be significantly more efficient than multi-word CAS or multi-word RMW [5, 13], which is even more general.

Luchangco, Moir and Shavit [92] defined the k -compare-single-swap (k -CSS) primitive, which atomically tests whether k specified memory locations contain specified values and, if all tests succeed, writes a value to one of the locations. They provided an *obstruction-free* implementation of k -CSS, meaning that a process performing a k -CSS is guaranteed to terminate if it runs alone. They implemented k -CSS using an obstruction-free implementation of LL/SC from CAS. Specifically, to try to update location v using k -CSS, a process performs LL(v), followed by two collects of the other $k - 1$ memory locations. If v has its specified value, both collects return their specified values, and the contents of these memory locations do not change between the two collects, the process performs SC to change the value of v . Unbounded version numbers are used both in their implementation of LL/SC and to avoid the ABA problem between the two collects.

Our LLX and SCX primitives can be viewed as multi-Data-record-LL and single-Data-record-SC primitives, with

the additional power to finalize Data-records. We shall see that this extra ability is extremely useful for implementing pointer-based data structures. In addition, our implementation of LLX and SCX allows us to develop shared data structures that satisfy the non-blocking progress condition, which is stronger than obstruction-freedom.

3.2 The primitives

Our primitives operate on a collection of Data-records of various user-defined types. Each type of Data-record has a fixed number of mutable fields (each fitting into a single word), and a fixed number of immutable fields (each of which can be large). Each field is given a value when the Data-record is created. Fields can contain pointers that refer to other Data-records. Data-records are accessed using LLX, SCX and VLX, and reads of individual mutable or immutable fields of a Data-record. Reads of mutable fields are permitted because a snapshot of a Data-record's fields is sometimes excessive, and it is sometimes sufficient (and more efficient) to use reads instead of LLXs.

An implementation of LL and SC from CAS has to ensure that, between when a process performs LL and when it next performs SC on the same word, the value of the word has not changed. Because the value of the word could change and then change back to a previous value, it is not sufficient to check that the word has the same value when the LL and the SC are performed. This is known as the ABA problem. It also arises for implementations of LLX and SCX from CAS. A general technique to overcome this problem is described in Section 3.3.1. However, if the data structure designer can guarantee that the ABA problem will not arise (because each SCX never attempts to store a value into a field that previously contained that value), our implementation can be used in a more efficient manner.

Before giving the precise specifications of the behaviour of LLX and SCX, we describe how to use them, with the implementation of a multiset as a running example. The multiset abstract data type supports three operations: *GET(key)*, which returns the number of occurrences of *key* in the multiset, *INSERT(key, count)*, which inserts *count* occurrences of *key* into the multiset, and *DELETE(key, count)*, which deletes *count* occurrences of *key* from the multiset and returns TRUE, provided there are at least *count* occurrences of *key* in the multiset. Otherwise, it simply returns FALSE.

Suppose we would like to implement a multiset using a sorted, singly-linked list. We represent each node in the list by a Data-record with an immutable field *key*, which contains a key in the multiset, and mutable fields: *count*, which records the number of times *key* appears in the multiset, and *next*, which points to the next node in the list. The first and last elements of the list are sentinel nodes with count 0 and with special keys $-\infty$ and ∞ , respectively, which never occur in the multiset.

Figure 3.5 shows how updates to the list are handled. Insertion behaves differently depending on whether the key is already present. Likewise, deletion behaves differently depending on whether it removes all copies of the key. For example, consider the operation *DELETE(d, 2)* depicted in Figure 3.5(c). This operation removes node *r* by changing *p.next* to point to a new copy of *rnext*. A new copy is used to avoid the ABA problem, since *p.next* may have pointed to *rnext* in the past. To perform the *DELETE(d, 2)*, a process first invokes LLXs on *p*, *r*, and *rnext*. Second, it creates a copy *rnext'* of *rnext*. Finally, it performs an SCX that depends on these three LLXs. This SCX attempts to change *p.next* to point to *rnext'*. This SCX will succeed only if none of *p*, *r* or *rnext* have changed since the aforementioned LLXs. Once *r* and *rnext* are removed from the list, we want subsequent invocations of LLX and SCX to be able to detect this, so that we can avoid, for example, erroneously inserting a key into a deleted part of the list. Thus, we specify in our invocation of SCX that *r* and *rnext* should be *finalized* if the SCX succeeds. Once a Data-record is finalized, it can never be changed again.

LLX takes (a pointer to) a Data-record r as its argument. Ordinarily, it returns either a snapshot of r 's mutable fields or FINALIZED. If an LLX(r) is concurrent with an SCX involving r , it is also allowed to fail and return FAIL. SCX takes four arguments: a sequence V of (pointers to) Data-records upon which the SCX depends, a subsequence R of V containing (pointers to) the Data-records to be finalized, a mutable field fld of a Data-record in V to be modified, and a value new to store in this field. VLX takes a sequence V of (pointers to) Data-records as its only argument. Each SCX and VLX and returns a Boolean value.

For example, in Figure 3.5(c), the $\text{DELETE}(d, 2)$ operation invokes $\text{SCX}(V, R, fld, new)$, where $V = \langle p, r, rnext \rangle$, $R = \langle r, rnext \rangle$, fld is the next pointer of p , and new points to the node $rnext'$.

A terminating LLX is called *successful* if it returns a snapshot or FINALIZED, and *unsuccessful* if it returns FAIL. A terminating SCX or VLX is called *successful* if it returns TRUE, and *unsuccessful* if it returns FALSE. Our operations are wait-free, but an operation may not terminate if the process performing it fails, in which case the operation is neither successful nor unsuccessful. We say an invocation I of LLX(r) by a process p is *linked to* an invocation I' of $\text{SCX}(V, R, fld, new)$ or $\text{VLX}(V)$ by process p if r is in V , I returns a snapshot, and between I and I' , process p performs no invocation of LLX(r) or $\text{SCX}(V', R', fld', new')$ and no unsuccessful invocation of $\text{VLX}(V')$, for any V' that contains r . Before invoking $\text{VLX}(V)$ or $\text{SCX}(V, R, fld, new)$, a process must perform an LLX(r) linked to the invocation for each r in V .

3.2.1 Correctness properties

An implementation of LLX, SCX and VLX is *correct* if, for every execution, there is a linearization of all successful LLXs, all successful SCXs, a subset of the non-terminating SCXs, all successful VLXs, and all reads, such that the following conditions are satisfied.

- C1:** Each read of a field f of a Data-record r returns the last value stored in f by an SCX linearized before the read (or f 's initial value, if no such SCX has modified f).
- C2:** Each linearized LLX(r) that does not return FINALIZED returns the last value stored in each mutable field f of r by an SCX linearized before the LLX (or f 's initial value, if no such SCX has modified f).
- C3:** Each linearized LLX(r) returns FINALIZED if and only if it is linearized after an $\text{SCX}(V, R, fld, new)$ with r in R .
- C4:** For each linearized invocation I of $\text{SCX}(V, R, fld, new)$ or $\text{VLX}(V)$, and for each r in V , no $\text{SCX}(V', R', fld', new')$ with r in V' is linearized between the LLX(r) linked to I and I .

The first three properties assert that successful reads and LLXs return correct answers. The last property says that an invocation of SCX or VLX does not succeed when it should not. However, an SCX can fail if it is concurrent with another SCX that accesses some Data-record in common. LL/SC also exhibits analogous failures in real systems. Our progress properties limit the situations in which this can occur.

3.2.2 Progress properties

In our implementation, LLX, SCX and VLX are wait-free, but they can fail spuriously. So, to be able to build non-blocking data structures, we must state our progress properties in terms of *successful* LLX, SCX and VLX operations. The first progress property guarantees that LLXs on finalized Data-records succeed.

- P1:** Each terminating LLX(r) returns FINALIZED if it begins after the end of a successful $\text{SCX}(V, R, fld, new)$ with r in R or after another LLX(r) has returned FINALIZED.

The next progress property guarantees non-blocking progress for queries built from LLX and VLX, and for updates built from LLX and SCX. Recall that VLX and SCX can be invoked only after performing a sequence of LLXs that return snapshots. Specifying a progress guarantee for SCX and VLX is subtle, because if processes repeatedly perform LLX on Data-records that have been finalized, or repeatedly perform failed LLXs, then they may never be able to invoke SCX or VLX. In particular, it is *not* sufficient to simply prove that LLXs return snapshots infinitely often, since *all* of the LLXs in a sequence must return snapshots before a process can invoke SCX or VLX. Additionally, to ensure that changes to a data structure can continue to occur, we make a general assumption that there is always at least one non-finalized Data-record.

We give two definitions which are helpful for clearly stating the progress properties for SCX and VLX. An SCX-UPDATE algorithm performs LLXs on a sequence V of Data-records and invokes $\text{SCX}(V, R, \text{fld}, \text{new})$ if all of these LLXs return snapshots. A successful SCX-UPDATE is one in which the SCX returns TRUE. Similarly, a VLX-QUERY algorithm performs LLXs on a sequence V of Data-records and invokes $\text{VLX}(V)$ if all of these LLXs return snapshots. A successful VLX-QUERY is one in which the VLX returns TRUE.

P2: Suppose that (a) there is always some non-finalized Data-record reachable by following pointers from an entry point, (b) for each Data-record r , each process performs finitely many invocations of $\text{LLX}(r)$ that return FINALIZED, and (c) processes perform infinitely many executions of SCX-UPDATE and/or VLX-QUERY algorithms. Then, infinitely many SCX or VLX operations succeed.

One way to prevent processes from performing an infinite number of invocations of $\text{LLX}(r)$ on a finalized Data-record r is to have each process keep track of the Data-records it knows are finalized. However, in many natural applications, for example, the multiset implementation in Chapter 4, this kind of explicit bookkeeping can be avoided.

Our implementation of LLX, SCX and VLX in Section 3.3 actually satisfies stronger progress properties than the ones described above. For example, a $\text{VLX}(V)$ or $\text{SCX}(V, R, \text{fld}, \text{new})$ is guaranteed to succeed if there is no concurrent $\text{SCX}(V', R', \text{fld}', \text{new}')$ such that V and V' have one or more elements in common. However, for the purposes of the specification of the primitives, we decided to give progress guarantees that are sufficient to prove that algorithms that use the primitives are non-blocking, but weak enough that it may be possible to design other, even more efficient implementations of the primitives. For example, our specification would allow some spurious failures of the type that occur in common implementations of ordinary LL/SC operations (as long as there is some guarantee that not all operations can fail spuriously).

3.3 Implementation of primitives

The shared data structure used to implement LLX, SCX and VLX consists of a set of Data-records and a set of SCX-records. (See Figure 3.1.) Each Data-record contains user-defined mutable and immutable fields. It also contains a *marked* bit, which is used to finalize the Data-record, and an *info* field. The marked bit is initially FALSE and only ever changes from FALSE to TRUE. The *info* field points to an SCX-record that describes the last SCX that accessed the Data-record. Initially, it points to a *dummy* SCX-record. When an SCX accesses a Data-record, it changes the *info* field of the Data-record to point to its SCX-record. While this SCX is active, the *info* field acts as a kind of lock on the Data-record, granting exclusive access to this SCX, rather than to a process. (To avoid confusion, we call this *freezing*, rather than locking, a Data-record.) We ensure that an SCX S does not change a Data-record for its own purposes while it is frozen for another SCX S' . Instead, S uses the information in the SCX-record of S' to help S' complete (successfully or unsuccessfully), so that the Data-record can be unfrozen. This cooperative approach is used

```

type Data-record
  ▷ User-defined fields
   $m_1, \dots, m_y$  ▷ mutable fields
   $i_1, \dots, i_z$  ▷ immutable fields
  ▷ Fields used by LLX/SCX algorithm
  info ▷ pointer to an SCX-record
  marked ▷ Boolean

type SCX-record
  V ▷ sequence of Data-records
  R ▷ subsequence of V to be finalized
  fld ▷ pointer to a field of a Data-record in V
  new ▷ value to be written into the field fld
  old ▷ value previously read from the field fld
  state ▷ one of {InProgress, Committed, Aborted}
  allFrozen ▷ Boolean
  infoFields ▷ sequence of pointers, one read from the
  ▷ info field of each element of V

```

Figure 3.1: Type definitions for shared objects used to implement LLX, SCX, and VLX.

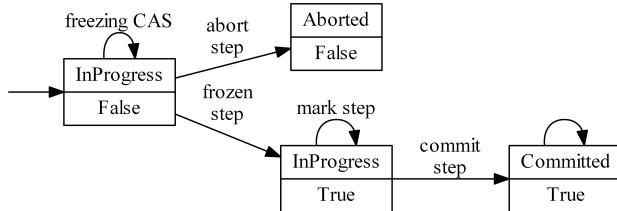


Figure 3.2: Possible [*state*, *allFrozen*] field transitions of an SCX-record.

to ensure progress.

An SCX-record contains enough information to allow any process to complete an SCX operation that is in progress. V, R, fld and *new* store the arguments of the SCX operation that created the SCX-record. Recall that R is a subsequence of V and *fld* points to a mutable field f of some Data-record r' in V . The value that was read from f by the LLX(r') linked to the SCX is stored in *old*. The SCX-record has one of three states, InProgress, Committed or Aborted, which is stored in its *state* field. This field is initially InProgress. The SCX-record of each SCX that terminates is eventually set to Committed or Aborted, depending on whether or not it successfully makes its desired update. The dummy SCX-record always has *state* = Aborted. The *allFrozen* bit, which is initially FALSE, gets set to TRUE after all Data-records in V have been frozen for the SCX. The values of *state* and *allFrozen* change in accordance with the diagram in Figure 3.2. The steps in the algorithm that cause these changes are also indicated. The *infoFields* field stores, for each r in V , the value of r 's *info* field that was read by the LLX(r) linked to the SCX.

We say that a Data-record r is *marked* when $r.marked = \text{TRUE}$. A Data-record r is *frozen* for an SCX-record U if $r.info$ points to U and either $U.state$ is InProgress, or $U.state$ is Committed and r is marked. While a Data-record r is frozen for an SCX-record U , a mutable field f of r can be changed only if f is the field pointed to by $U.fld$ (and it can only be changed by a process helping the SCX that created U). Once a Data-record r is marked and $r.info.state$

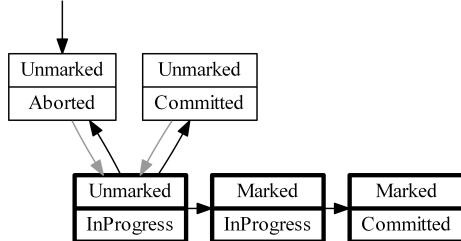


Figure 3.3: Possible transitions for the *marked* field of a Data-record and the *state* of the SCX-record pointed to by the *info* field of the Data-record.

becomes Committed, r will never be modified again in any way. Figure 3.3 shows how a Data-record can change between frozen and unfrozen. The three bold boxes represent frozen Data-records. The other two boxes represent Data-records that are not frozen. A Data-record r can only become frozen when $r.info$ is changed (to point to a new SCX-record whose state is InProgress). This is represented by the grey edges. The black edges represent changes to $r.info.state$ or $r.marked$. A frozen Data-record r can only become unfrozen when $r.info.state$ is changed.

3.3.1 Constraints

For the sake of efficiency, we have designed our implementation of LLX, VLX and SCX to work only if the primitives are used in a way that satisfies certain constraints, described in this section. We also describe general ways to ensure these constraints are satisfied. However, there are often quite natural ways to ensure the constraints are satisfied without resorting to the extra work required by the general solutions.

Since our implementation of LLX, SCX and VLX uses helping to guarantee progress, each CAS of an SCX might be repeatedly performed by several helpers, possibly after the original invocation of SCX has terminated. To avoid difficulties, we must show there is no ABA problem in the fields affected by these CAS steps.

The *info* field of a Data-record r is modified by CAS steps that attempt to freeze r for an SCX. All such steps performed by processes helping one invocation of SCX try to CAS the *info* field of r from the same old value to the same new value, and that new value is a pointer to a newly created SCX-record. Because the SCX-record was freshly allocated for the SCX, the ABA problem will not arise in the *info* field.

We must also avoid the ABA problem in mutable fields of Data-records, which are modified using CAS steps by invocations of SCX. Formally, it suffices to prove the following constraint is satisfied.

- **Constraint:** For every invocation S of $\text{SCX}(V, R, fld, new)$, new is not the initial value of fld , and no invocation of $\text{SCX}(V', R', fld, new)$ was linearized before the $\text{LLX}(r)$ linked to S was linearized, where r is the Data-record that contains fld .

This constraint can be satisfied using a similar approach to the one used for *info* fields. Specifically, the new value can be placed inside a wrapper object that is freshly allocated. (This is referred to as Solution 3 in [42], which discusses several general solutions to the ABA problem.) However, the extra level of indirection slows down accesses to fields. Observe that, if mutable fields are only ever changed to point to newly allocated Data-records, then the Data-records themselves serve the same function as wrapper objects. In this case, the constraint is satisfied without the overhead of wrapper objects. The multiset implementation in Chapter 4 works this way.

To guarantee progress for invocations of SCX, we place a constraint on the way SCX is used. Our implementation

of $\text{SCX}(V, R, fld, new)$ does something similar to locking each Data-record in V . Livelock could occur if different invocations of SCX do not process Data-records in the same order. One way to prevent this is to define a total ordering on all Data-records (for example, ordering them by their locations in memory) and having each invocation of SCX sort its V sequence using this ordering. However, this sorting can be expensive. In fact, to guarantee progress, it is not necessary for *all* SCXs to order their V sequences consistently. Instead, it suffices to show that, if all the Data-records stop changing, then the sequences passed to later invocations of SCX are all consistent with some total order. Formally, use of our implementation of SCX requires adherence to the following constraint.

- **Constraint:** Consider each execution that contains a configuration C after which the value of no field of any Data-record changes. There must be a total order on all Data-records created during this execution such that, if Data-record r_1 appears before Data-record r_2 in the sequence V passed to an invocation of SCX whose linked LLXs begin after C , then $r_1 < r_2$.

This property is often easy to satisfy in a natural way. In many *search* data structures (where operations first traverse the data structure and then perform an update), this constraint is satisfied simply by ordering the elements of each V sequence using the order they are encountered during the traversal. For example, if one were using LLX and SCX to implement an *unsorted* singly-linked list, then this constraint would be satisfied if the nodes in each sequence V occur in the order they are encountered by following next pointers from the beginning of the list, *even if* some operations could reorder the nodes in the list. While the list is changing, such a sequence may have repeated elements and might not be consistent with any total order.

Although this simple approach satisfies the constraint for many data structures, it does not always work. For example, consider a doubly-linked list in which some operations traverse left-to-right and some operations traverse right-to-left. In a static list, a left-to-right traversal can visit Data-records A then B , and a right-to-left traversal can visit B then A . In this case, it is not sufficient to order V sequences by the order in which Data-records are visited by traversals. Instead, an update after a left-to-right traversal could use the order in which it visited Data-records during the traversal, and an update after a right-to-left traversal could use the *reverse* of the order in which it visited Data-records.

3.3.2 Detailed algorithm description and sketch of proofs

Pseudocode for our implementation of LLX, VLX and SCX appears in Figure 3.4. If x contains a pointer to a record, then $x.y := v$ assigns the value v to field y of this record, $\&x.y$ denotes the address of this field and all other occurrences of $x.y$ denote the value stored in this field.

Theorem *The algorithms in Figure 3.4 satisfy properties C1 to C4 and P1 to P2 in every execution where the constraints of Section 3.3.1 are satisfied.*

An $\text{LLX}(r)$ returns a snapshot, FAIL, or FINALIZED. At a high level, it works as follows. If the LLX determines that r is not frozen and r 's *info* field does not change while the LLX reads the mutable fields of r , the LLX returns the values read as a snapshot. Otherwise, the LLX helps the SCX that it saw freeze r , and returns FAIL or FINALIZED. If the LLX returns FAIL, it is not linearized. We now discuss in more detail how LLX operates and is linearized in the other two cases.

First, suppose the $\text{LLX}(r)$ returns a snapshot at line 11. Then, the test at line 7 evaluates to TRUE. So, either $\text{state} = \text{Aborted}$, which means r is not frozen at line 5, or $\text{state} = \text{Committed}$ and $\text{marked}_2 = \text{FALSE}$. This also means r is not frozen at line 5, since $r.\text{marked}$ cannot change from TRUE to FALSE. The LLX reads r 's mutable fields (line 8)

```

1   LLX( $r$ ) by process  $p$ 
2   ▷ Precondition:  $r \neq \text{NIL}$ .
3    $\text{marked}_1 := r.\text{marked}$                                      ▷ order of lines 3–6 matters
4    $\text{rinfo} := r.\text{info}$ 
5    $\text{state} := \text{rinfo.state}$ 
6    $\text{marked}_2 := r.\text{marked}$ 
7   if  $\text{state} = \text{Aborted}$  or ( $\text{state} = \text{Committed}$  and not  $\text{marked}_2$ ) then      ▷ if  $r$  was not frozen at line 5
8     read  $r.m_1, \dots, r.m_y$  and record the values in local variables  $m_1, \dots, m_y$ 
9     if  $r.\text{info} = \text{rinfo}$  then                                         ▷ if  $r.\text{info}$  points to the same
10    store  $\langle r, \text{rinfo}, \langle m_1, \dots, m_y \rangle \rangle$  in  $p$ 's local table
11    return  $\langle m_1, \dots, m_y \rangle$                                          ▷ SCX-record as on line 4

12   if  $\text{state} = \text{InProgress}$  then HELP( $\text{rinfo}$ )
13   if  $\text{marked}_1$  then
14     return FINALIZED
15   else
16     return FAIL

17   SCX( $V, R, \text{fld}, \text{new}$ ) by process  $p$ 
18   ▷ Preconditions: (1) for each  $r$  in  $V$ ,  $p$  has performed an invocation  $I_r$  of LLX( $r$ ) linked to this SCX
19   (2)  $\text{new}$  is not the initial value of  $\text{fld}$ 
20   (3) for each  $r$  in  $V$ , no SCX( $V', R', \text{fld}, \text{new}$ ) was linearized before  $I_r$  was linearized
21   Let  $\text{infoFields}$  be a pointer to a newly created table in shared memory containing,
22   for each  $r$  in  $V$ , a copy of  $r$ 's  $\text{info}$  value in  $p$ 's local table of LLX results
23   Let  $\text{old}$  be the value for  $\text{fld}$  stored in  $p$ 's local table of LLX results
24   return HELP(pointer to new SCX-record( $V, R, \text{fld}, \text{new}, \text{old}, \text{InProgress}, \text{FALSE}, \text{infoFields}$ ))

25   HELP( $\text{scxPtr}$ )
26   ▷ Freeze all Data-records in  $\text{scxPtr.V}$  to protect their mutable fields from being changed by other SCXs
27   for each  $r$  in  $\text{scxPtr.V}$  enumerated in order do
28   Let  $\text{rinfo}$  be the pointer indexed by  $r$  in  $\text{scxPtr.infoFields}$ 
29   if not CAS( $r.\text{info}$ ,  $\text{rinfo}$ ,  $\text{scxPtr}$ ) then                                ▷ freezing CAS
30   if  $r.\text{info} \neq \text{scxPtr}$  then
31   ▷ Could not freeze  $r$  because it is frozen for another SCX
32   if  $\text{scxPtr.allFrozen} = \text{TRUE}$  then                                         ▷ frozen check step
33   ▷ the SCX has already completed successfully
34   return TRUE
35   else
36   ▷ Atomically unfreeze all Data-records frozen for this SCX
37    $\text{scxPtr.state} := \text{Aborted}$                                          ▷ abort step
38   return FALSE

39   ▷ Finished freezing Data-records (Assert:  $\text{state} \in \{\text{InProgress}, \text{Committed}\}$ )
40    $\text{scxPtr.allFrozen} := \text{TRUE}$                                          ▷ frozen step
41   for each  $r$  in  $\text{scxPtr.R}$  do  $r.\text{marked} := \text{TRUE}$                                ▷ mark step
42   CAS( $\text{scxPtr.fld}$ ,  $\text{scxPtr.old}$ ,  $\text{scxPtr.new}$ )                                ▷ update CAS

43   ▷ Finalize all  $r$  in  $R$ , and unfreeze all  $r$  in  $V$  that are not in  $R$ 
44    $\text{scxPtr.state} := \text{Committed}$                                          ▷ commit step
45   return TRUE

46   VLX( $V$ ) by process  $p$ 
47   ▷ Precondition: for each Data-record  $r$  in  $V$ ,  $p$  has performed an LLX( $r$ ) linked to this VLX
48   for each  $r$  in  $V$  do
49   Let  $\text{rinfo}$  be the  $\text{info}$  field for  $r$  stored in  $p$ 's local table of LLX results
50   if  $\text{rinfo} \neq r.\text{info}$  then return FALSE      ▷  $r$  changed since LLX( $r$ ) read  $\text{info}$ 
51   return TRUE      ▷ At some point during the loop, all  $r$  in  $V$  were unchanged

```

Figure 3.4: Pseudocode for LLX, SCX and VLX.

and rereads $r.info$ at line 9, finding it the same as on line 4. In Section 3.3.1, we explained why this implies that $r.info$ did not change between lines 4 and 9. Since r is not frozen at line 5, we know from Figure 3.3 that r is unfrozen at all times between line 5 and 9. We prove that mutable fields can change only while r is frozen, so the values read by line 8 constitute a snapshot of r 's mutable fields. Thus, we can linearize the LLX at line 9.

Now, suppose the LLX(r) returns FINALIZED. Then, the test on line 13 evaluated to TRUE, so r was already marked when line 3 was performed (and it will remain marked forever). Immediately before line 13, the LLX performs HELP($r.info$) if $state = \text{InProgress}$. Below, we argue that either $state = \text{Committed}$ or this invocation of HELP($r.info$) will perform a commit step and change $r.info.state$ to Committed before returning. By Figure 3.3(a), the $state$ of an SCX-record never changes after it is set to Committed. So, after line 13, $r.info.state = \text{Committed}$ and, thus, r has been finalized. Hence, the LLX can be linearized at line 14.

When a process performs an SCX, it first creates a new SCX-record and then invokes HELP (line 21). The HELP routine performs the real work of the SCX. It is also used by a process to help other processes complete their SCXs (successfully or unsuccessfully). The values in an SCX-record's *old* and *infoFields* come from a table in the local memory of the process that invokes the SCX, which stores the results of the last LLX it performed on each Data-record. (In practice, the memory required for this table could be greatly reduced when a process knows which of these values are needed for future SCXs. Alternatively, instead of using a process-local table of values, one could have LLX return the values it reads, and then simply pass the appropriate values as arguments to invocations of SCX and VLX.)

Consider an invocation of HELP(U) by process p to carry out the work of the invocation S of SCX(V, R, fld, new) that is described by the SCX-record U . First, p attempts to freeze each r in V by performing a *freezing CAS* to store a pointer to U in $r.info$ (line 26). Process p uses the value read from $r.info$ by the LLX(r) linked to S as the old value for this CAS and, hence, it will succeed only if r has not been frozen for any other SCX since then. If p 's freezing CAS fails, it checks whether some other helper has successfully frozen the Data-record with a pointer to U (line 27).

If every r in V is successfully frozen, p performs a *frozen step* to set $U.allFrozen$ to TRUE (line 37). After this frozen step, the SCX is guaranteed not to fail, meaning that no process will perform an abort step while helping this SCX. Then, for each r in R , p performs a *mark step* to set $r.marked$ to TRUE (line 38) and, from Figure 3.3, r remains frozen from then on. Next, p performs an *update CAS*, storing new in the field pointed to by fld (line 39), if successful. We prove that, among all the update CAS steps on fld performed by the helpers of U , only the first can succeed. Finally, p unfreezes all r in V that are not in R by performing a *commit step* that changes $U.state$ to Committed (line 41).

Now suppose that, when p performs line 27, it finds that some Data-record r in V is already frozen for another invocation S' of SCX. If $U.allFrozen$ is FALSE at line 29, then we can prove that no helper of S will ever reach line 37, so p can abort S . To do so, it unfreezes each r in V that it has frozen by performing an *abort step*, which changes $U.state$ to Aborted (line 34), and then returns FALSE (line 35) to indicate that S has been aborted. If $U.allFrozen$ is TRUE at line 29, it means that each element of V , including r , was successfully frozen by some helper of S and then, later, a process froze r for S' . Since S cannot be aborted after $U.allFrozen$ was set to TRUE, its state must have changed from InProgress to Committed before r was frozen for another SCX-record. Therefore, S was successfully completed and p can return TRUE at line 31.

We linearize an invocation of SCX at the first update CAS performed by one of its helpers. We prove that this update CAS always succeeds. Thus, all SCXs that return TRUE are linearized, as well as possibly some non-terminating SCXs. The first update CAS of SCX(V, R, fld, new) modifies the value of fld , so a read(fld) that occurs immediately

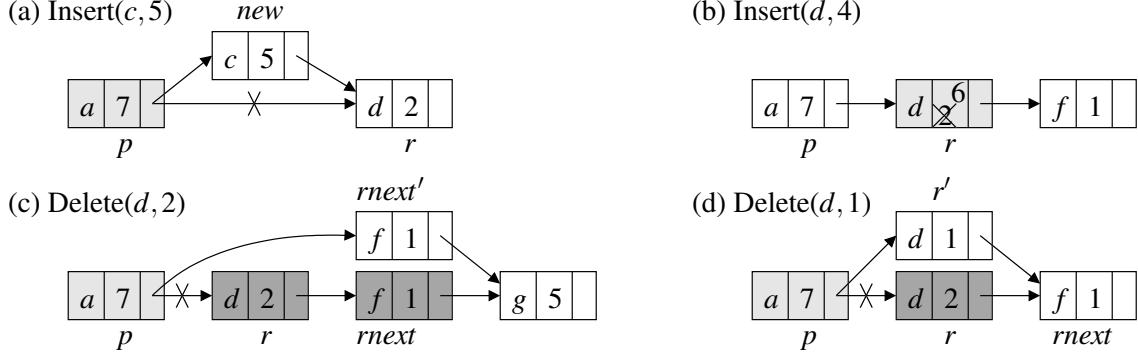


Figure 3.5: Using SCX to update a multiset. LLXs of all shaded nodes are linked to the SCX. Darkly shaded nodes are finalized by the SCX. Where a field has changed, the old value is crossed out.

after the update CAS will return the value of *new*. Hence, the linearization point of an SCX must occur at its first update CAS. There is one subtle issue about this linearization point: If an LLX(*r*) is linearized between the update CAS and commit step of an SCX that finalizes *r*, it might not return FINALIZED, violating condition C3. However, this cannot happen. Before the LLX is linearized on line 14, the LLX either sees *state* \neq InProgress or helps the SCX. Since the LLX has seen that the node *r* is marked, and *r* *cannot* be marked and point to an SCX-record with *state* Aborted (as Figure 3.3 shows), the LLX will either see *state* = Committed, or help the SCX perform a commit step.

An invocation *I* of VLX(*V*) is executed by a process *p* after *p* has performed an invocation of LLX(*r*) linked to *I*, for each *r* in *V*. VLX(*V*) simply checks, for each *r* in *V*, that the *info* field of *r* is the same as when it was read by *p*'s last LLX(*r*) and, if so, VLX(*V*) returns TRUE. In this case, we prove that each Data-record in *V* does not change between the linked LLX and the time its *info* field is reread. Thus, the VLX can be linearized at the first time it executes line 47. Otherwise, the VLX returns FALSE to indicate that the LLX results may not constitute a snapshot.

We remark that our use of the cooperative method avoids costly recursive helping. If, while *p* is helping *S*, it cannot freeze all of *S*'s Data-records because one of them is already frozen for a third SCX, then *p* will simply perform an abort step, which unfreezes all Data-records that *S* has frozen.

We briefly sketch why the progress properties described in Section 3.2.2 are satisfied. It follows easily from the code that an invocation of LLX(*r*) returns FINALIZED if it begins after the end of an SCX that finalized *r* or another LLX sees that *r* is finalized. This establishes P1.

We now sketch the proof of P2. Consider an execution in which processes execute infinitely many SCX-UPDATE and/or VLX-QUERY algorithms, but only a finite number succeed. In this execution, only finitely many SCX-records are created. Each process calls HELP(*U*) if it sees that *U.state* = InProgress, which it can do at most once for each SCX-record *U*. Since every CAS is performed inside the HELP routine, there is some time *t* after which no process performs a CAS, calls HELP, or sees a SCX-record whose *state* is InProgress. An SCX or VLX can fail only when an *info* field is modified by a concurrent SCX operation, and an LLX can only fail for the same reason, or when it sees a SCX-record whose *state* is InProgress. Therefore, all LLXs, VLXs and SCXs that begin after *t* will succeed. Since processes execute infinitely many SCX-UPDATE and/or VLX-UPDATE algorithms, they must execute infinitely many after *t*. The first SCX-UPDATE or VLX-UPDATE algorithms to occur after *t* will succeed, yielding a contradiction.

3.3.3 Additional properties

Our implementation of SCX satisfies some additional properties, which are helpful for designing certain kinds of non-blocking data structures so that query operations can run efficiently. Consider a pointer-based data structure with a fixed set of Data-records called *entry points*. An operation on the data structure starts at an entry point and follows pointers to visit other Data-records. (For example, in our multiset example, the head of the linked list is the sole entry point for the data structure.) We say that a Data-record is *in the data structure* if it can be reached by following pointers from an entry point, and a Data-record r is *removed from the data structure* by an SCX if r is in the data structure immediately prior to the linearization point of the SCX and is not in the data structure immediately afterwards.

If the data structure is designed so that a Data-record is finalized when (and only when) it is removed from the data structure, then we have the following additional properties.

Proposition *Suppose each linearized $\text{SCX}(V, R, \text{fld}, \text{new})$ removes precisely the Data-records in R from the data structure.*

- If $\text{LLX}(r)$ returns a value different from FAIL or FINALIZED, r is in the data structure just before the LLX is linearized.
- If an $\text{SCX}(V, R, \text{fld}, \text{new})$ is linearized and new is (a pointer to) a Data-record, then this Data-record is in the data structure just after the SCX is linearized.
- If an operation reaches a Data-record r by following pointers read from other Data-records, starting from an entry point, then r was in the data structure at some earlier time during the operation.

The first two properties are straightforward to prove. The last property is proved by induction on the Data-records reached. For the base case, entry points are always reachable. For the induction step, consider the time when an operation reads a pointer to r from another Data-record r' that the operation reached earlier. By the induction hypothesis, there was an earlier time t during the operation when r' was in the data structure. If r' already contained a pointer to r at t , then r was also in the data structure at that time. Otherwise, an SCX wrote a pointer to r in r' after t , and just after that update occurred, r' and r were in the data structure (by the second part of the proposition).

The last property is a particularly useful one for linearizing query operations. It means that operations that search through a data structure can use simple reads of pointers instead of the more expensive LLX operations. Even though the Data-record that such a search operation reaches may have been removed from the data structure by the time it is reached, the lemma guarantees that there *was* a time during the search when the Data-record was in the data structure. For example, we use this property to linearize searches in our multiset algorithm in Chapter 4.

3.4 Formal proof

We now give a formal proof of correctness for our implementation of LLX, SCX and VLX. A dependency graph, which appears in Figure 3.6, gives a visual overview of the following results, and how they refer to one another. An edge is drawn from A to B if B refers to A . To save space, rather than annotating nodes of the graph with “Lemma 3.60” and “Corollary 3.13,” we simply write “60” and “13.”

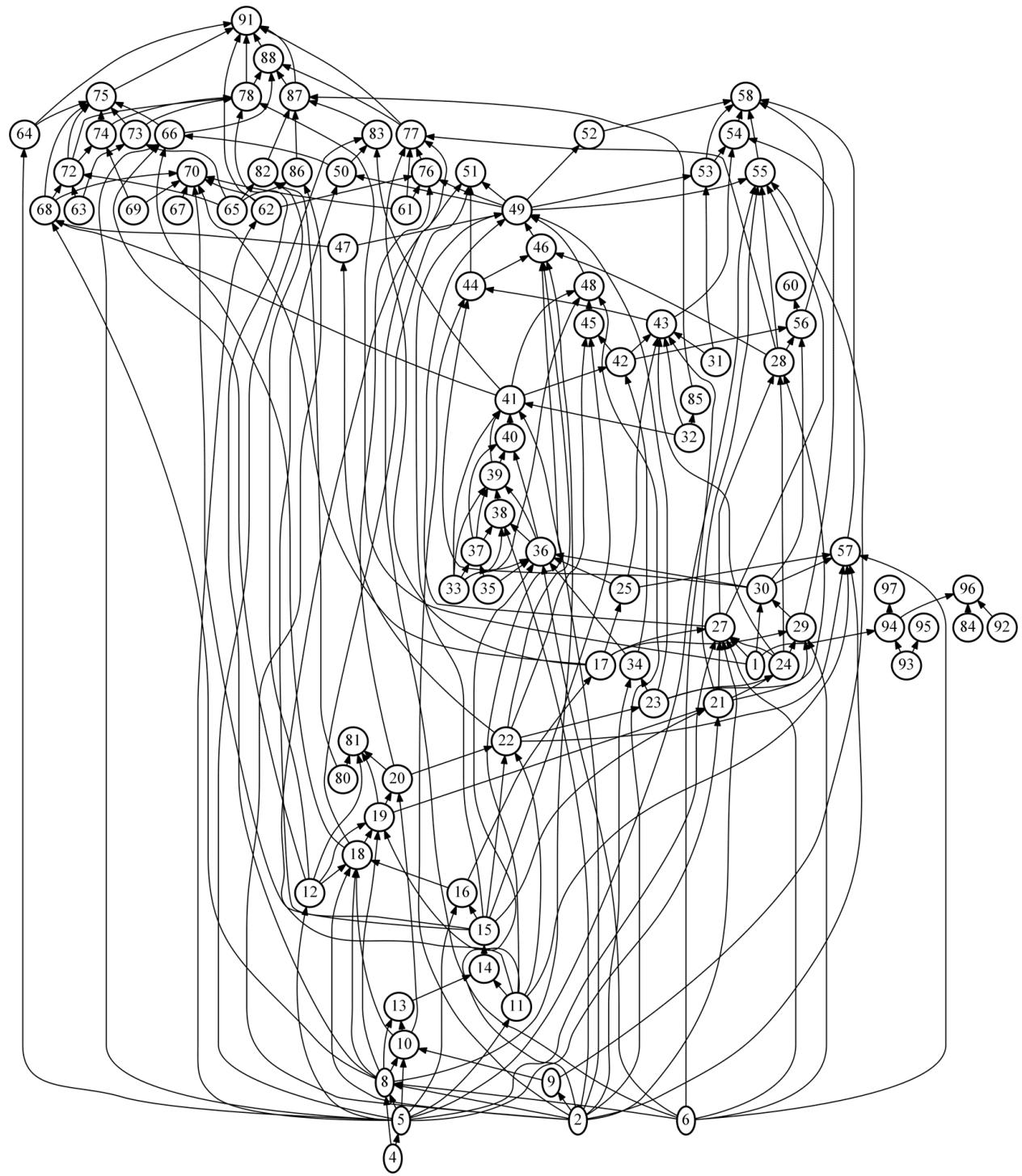


Figure 3.6: Dependency graph for the formal correctness proof.

3.4.1 Basic properties

We begin with some elementary properties that are needed to prove basic lemmas about freezing. In particular, we show that the *info* field of a Data-record cannot experience an ABA problem.

Definition 3.1 Let I' be an invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$ or $\text{VLX}(V)$ by a process p , and r be a Data-record in V . We say an invocation I of $\text{LLX}(r)$ is **linked to** I' if and only if:

1. I returns a value different from FAIL or FINALIZED, and
2. no invocation of $\text{LLX}(r)$, $\text{SCX}(V', R', \text{fld}', \text{new}')$, or $\text{VLX}(V')$, where V' contains r , is performed by p between I and I' .

Observation 3.2 An SCX-record U created by an invocation S of SCX satisfies the following invariants.

1. $U.\text{fld}$ points to a mutable field f of a Data-record r' in $U.V$.
2. The value stored in $U.\text{old}$ was read at line 8 from f by the $\text{LLX}(r')$ linked to S .
3. For each r in $U.V$, the pointer indexed by r in $U.\text{infoFields}$ was read from $r.\text{info}$ at line 4 by the $\text{LLX}(r)$ linked to S .
4. For each r in $U.V$, the $\text{LLX}(r)$ linked to S must enter the if-block at line 7, and see $r.\text{info} = r.\text{info}$ at line 9.

Proof: None of the fields of an SCX-record except *state* change after they are initialized at line 21. The contents of the table pointed to by $U.\text{infoFields}$ do not change, either. Therefore, it suffices to show that these invariants hold when U is created. The proof of these invariants follows immediately from the precondition of SCX, the pseudocode of LLX, and the definition of an LLX linked to an SCX. ■

The following two definitions ease discussion of the important steps that access shared memory.

Definition 3.3 A process is said to be **helping** an invocation of SCX that created an SCX-record U whenever it is executing $\text{HELP}(\text{ptr})$, where ptr points to U . (For brevity, we sometimes say a process is “helping U ” instead of “helping the SCX that created U .”)

Note that, since HELP does not call itself directly or indirectly, a process cannot be helping two different invocations of SCX at the same time.

Definition 3.4 We say that a freezing CAS, update CAS, frozen step, mark step, abort step, commit step or frozen check step S belongs to an SCX-record U when S is performed by a process helping U . We say that a frozen step, mark step, abort step, commit step or frozen check step is **successful** if it changes the the field it modifies to a different value. A freezing CAS or update CAS is successful if the CAS succeeds. Any step is **unsuccessful** if it is not successful.

Lemma 3.5 No freezing CAS, update CAS, frozen step, mark step, abort step, commit step or frozen check step belongs to the dummy SCX-record.

Proof: According to Definition 3.4, we must simply show that no process ever helps the dummy SCX-record. An invocation H of $\text{HELP}(\text{scxPtr})$ is only invoked at line 21 or line 12. If H occurs at line 21, then scxPtr cannot be the dummy SCX-record, since the argument to H is a newly created SCX-record. So, suppose H occurs at line 12 in LLX. Then the LLX sees $\text{state} = \text{InProgress}$ just before performing H . Thus, scxPtr had $\text{state} = \text{InProgress}$ at some point before H . Since the dummy SCX-record initially has state Aborted, and InProgress is never written into any state field, scxPtr cannot be the dummy SCX-record. ■

Lemma 3.6 *Every update to the info field of a Data-record r changes $r.\text{info}$ to a value that has never previously appeared there. Hence, there is no ABA problem on info fields.*

Proof: We first note that $r.\text{info}$ can only be changed by a freezing CAS at line 26. When a freezing CAS attempts to change an info field f from x to y , $y.\text{infoFields}$ contains x (by line 25). Then, since $y.\text{infoFields}$ does not change after the SCX-record pointed to by y is created, the SCX-record pointed to by x was created before the SCX-record pointed to by y . So, letting a_1, a_2, \dots be the sequence of SCX-records ever pointed to by $r.\text{info}$, we know that a_1, a_2, \dots were created (at line 21) in that order. Since we have assumed memory allocations always receive new addresses, a_1, a_2, \dots are distinct. ■

Definition 3.7 *A freezing CAS **on** a Data-record r is one that operates on $r.\text{info}$. A mark step **on** a Data-record r is one that writes to $r.\text{marked}$.*

Lemma 3.8 *For each Data-record r in the V sequence of an SCX-record U , only the first freezing CAS belonging to U on r can succeed.*

Proof: Let ptr be a pointer to U , and fcas be the first freezing CAS belonging to U on r . Let $r.\text{info}$ be the old value used by fcas . By Definition 3.4, the new value used by fcas is ptr . Since fcas belongs to U , Lemma 3.5 implies that U is not the dummy SCX-record initially pointed to by each info field. Hence, U was created by an invocation of SCX, so Observation 3.2.3 implies that $r.\text{info}$ contained $r.\text{info}$ during the $\text{LLX}(r)$ linked to S . Since the $\text{LLX}(r)$ linked to S terminates before the start of S , and S creates U , the $\text{LLX}(r)$ linked to S must terminate before any invocation of $\text{HELP}(\text{ptr})$ begins. From the code of HELP , fcas occurs in an invocation of $\text{HELP}(\text{ptr})$. Thus, $r.\text{info}$ contains $r.\text{info}$ at some point before fcas . If fcas is successful, then $r.\text{info}$ contains $r.\text{info}$ just before fcas , and ptr just after. Otherwise, $r.\text{info}$ contains $r.\text{info}$ at some point before fcas , but contains some other value just before fcas . In either case, Lemma 3.6 implies that $r.\text{info}$ can never again contain $r.\text{info}$ after fcas . Finally, since each freezing CAS belonging to U on r uses $r.\text{info}$ as its old value (by line 25 and the fact that table $U.\text{infoFields}$ does not change after it is first created), there can be no successful freezing CAS belonging to U on r after fcas . ■

3.4.2 Changes to the info field of a Data-record and the state field of an SCX-record

We prove that freezing of nodes proceeds an orderly way. The first lemma shows that a process cannot freeze a node that is frozen by a different operation that is still in progress.

Lemma 3.9 *The info field of a Data-record r cannot be changed while $r.\text{info}$ points to an SCX-record with state InProgress .*

Proof: Suppose an *info* field of a Data-record r is changed while it points to an SCX-record U with $U.state = \text{InProgress}$. This change can only be performed by a successful freezing CAS $fcas$ whose old value is a pointer to U and whose new value is a pointer to W . Let S be the invocation of SCX that created W . From line 25, we can see that the old value for $fcas$ (a pointer to U) is stored in the table $W.infoFields$ and, by Observation 3.2.3, this value was read from $r.info$ (at line 4) by the LLX(r) linked to S . Hence, the LLX(r) linked to S reads $U.state$ at line 5. By Observation 3.2.4, the LLX(r) linked to S passes the test at line 7 and enters the if-block. This implies that, when $U.state$ was read at line 5, either it was Committed and r was unmarked, or it was Aborted. Thus, $U.state$ must be Aborted or Committed prior to $fcas$, and the claim follows from the fact that InProgress is never written to $U.state$. ■

It follows from Lemma 3.9 that if a node is frozen for an operation O , it remains frozen for O until O is committed or aborted.

Lemma 3.10 *If there is a successful freezing CAS $fcas$ belonging to an SCX-record U on a Data-record r , and some time t after the first freezing CAS belonging to U on r and before the first abort step or commit step belonging to U , then $r.info$ points to U at t .*

Proof: Since $fcas$ belongs to U , by Lemma 3.5, U cannot be the dummy SCX-record, so U is created at line 21, where $U.state$ is initially set to InProgress. Let t' be when the first abort step or commit step belonging to U on r occurs. By Lemma 3.8, $fcas$ must be the first freezing CAS belonging to U on r . Thus, t is after $fcas$ occurs, and before t' . Immediately following $fcas$, $r.info$ points to U . From the code, $U.state$ can only be changed by an abort step or commit step belonging to U . Therefore, $U.state = \text{InProgress}$ at all times after $fcas$ and before t' . By Lemma 3.9, $r.info$ cannot change after $fcas$, and before t' . Hence, $r.info$ points to U at t . ■

The following result proves that a frozen step occurs only after all Data-records are successfully frozen.

Lemma 3.11 *If a frozen step belongs to an SCX-record U then, for each r in $U.V$, there is a successful freezing CAS belonging to U on r that occurs before the first frozen step belonging to U .*

Proof: Suppose a frozen step belongs to U . Let $fstep$ be the first such frozen step and let H be the invocation of HELP that performs $fstep$. Since $fstep$ occurs at line 37, for each Data-record r in $U.V$, H must perform a successful freezing CAS belonging to U on r or see $otherPtr = scxPtr$ in the preceding loop. If H performs a successful freezing CAS belonging to U on r , then we are done. Otherwise, $r.info = scxPtr$ at some point before $fstep$. Since $fstep$ belongs to U , $scxPtr$ points to U . From Lemma 3.5, U is not the dummy SCX-record to which $r.info$ initially points. Hence, some process must have changed $r.info$ to point to U , which can only be done by a successful freezing CAS. ■

Finally, we show that Data-records are frozen in the correct order. (This is used below to prove that livelock cannot occur.)

Lemma 3.12 *Let U be an SCX-record, and $\langle r_1, r_2, \dots, r_l \rangle$ be the sequence of Data-records in $U.V$. For $i \geq 2$, a freezing CAS belonging to U on Data-record r_i can occur only after a successful freezing CAS belonging to U on r_{i-1} .*

Proof: Let $fcas$ be a freezing CAS belonging to U on r_i , for some $i \geq 2$. Let H be the invocation of HELP which performs $fcas$. The loop in H iterates over the sequence r_1, r_2, \dots, r_l , so H performs $fcas$ in iteration i of the loop. Since H reaches iteration i , H must perform iteration $i - 1$. Thus, by the code of HELP, H must perform a freezing CAS

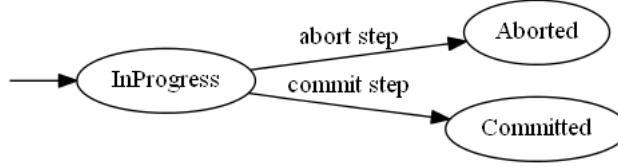


Figure 3.7: Possible transitions for the *state* field of an SCX-record (initially InProgress).

$fcas'$ belonging to U on r_{i-1} at line 26 before $fcas$. If $fcas'$ succeeds, then the claim is proved. Otherwise, H will check whether $r_{i-1}.info$ is equal to $scxPtr$ at line 27. Since HELP does not return in iteration $i - 1$, $r_{i-1}.info = scxPtr$. This can only be true if U is the dummy SCX-record, or there has already been a successful freezing CAS belonging to U on r_{i-1} . Since $fcas$ belongs to U , Lemma 3.5 implies that U cannot not be the dummy SCX-record. ■

3.4.3 Proving *state* and *info* fields change as described in Fig. 3.3

In this section, we first prove that an SCX-record's *state* transitions respect Figure 3.7, then we expand upon the information in Figure 3.7 by showing that frozen steps, abort steps, commit steps and successful freezing CAs proceed as illustrated in Figure 3.3.

We now prove an SCX-record U 's *state* transitions respect Figure 3.7 by noting that a $U.state$ is never changed to InProgress, and proving that it does not change from Aborted to Committed or vice versa. Since $U.state$ can only be changed by a commit step or abort step belonging to U , and each commit step is preceded by a frozen step (by the code of HELP), it suffices to show that there cannot be both a frozen step and an abort step belonging to an SCX-record.

Lemma 3.13 *Let U be an SCX-record. Suppose that there is a successful freezing CAS belonging to U on each Data-record in $U.V$. Then, a frozen check step belonging to U cannot occur until after a frozen step belonging to U has occurred.*

Proof: To derive a contradiction, suppose that the first frozen check step $fcstep$ belonging to U occurs before any frozen step belonging to U . Let H be the invocation of HELP in which $fcstep$ occurs. Before $fcstep$, H performs an unsuccessful freezing CAS at line 26 on one Data-record in $U.V$. The hypothesis of the lemma says that there is a successful freezing CAS, $fcas$, belonging to U on this same Data-record. By Lemma 3.8, $fcas$ must occur before H 's unsuccessful freezing CAS. Thus, $fcas$ occurs before $fcstep$, which occurs before any frozen step belonging to U . From the code of HELP, a frozen step belonging to U precedes the first commit step belonging to U , which implies that $fcas$ and $fcstep$ occur before the first commit step belonging to U . Further, the code of HELP implies that no abort step can occur before $fcstep$. Thus, $fcas$ and $fcstep$ occur strictly before the first commit step or abort step belonging to U . After $fcas$, and before $fcstep$, H sees $r.info \neq scxPtr$ at line 27. Since $scxPtr$ points to U , this implies that $r.info$ points to some SCX-record different from U . However, Lemma 3.10 implies that $r.info$ must point to U when H performs line 27, which is a contradiction. ■

Corollary 3.14 *If a frozen step belongs to an SCX-record U , then the first such frozen step must occur before any frozen check step belonging to U .*

Proof: Suppose a frozen step belongs to U . By Lemma 3.11, we know there is a successful freezing CAS belonging to U on r for each r in $U.V$. Thus, by Lemma 3.13, a frozen check step belonging to U cannot occur until a frozen step belonging to U has occurred. ■

Lemma 3.15 *There cannot be both a frozen step and an abort step belonging to the same SCX-record.*

Proof: Suppose a frozen step belongs to U . By Corollary 3.14, the first such frozen step precedes the first frozen check step, which, by the pseudocode of HELP, precedes the first abort step. This frozen step sets $U.allFrozen$ to TRUE and $U.allFrozen$ is never changed from TRUE to FALSE. Therefore, any process that performs a frozen check step belonging to U will immediately return TRUE, without performing an abort step. Thus, there can be no abort step belonging to U . ■

Corollary 3.16 *An SCX-record U 's state cannot change from Committed to Aborted or from Aborted to Committed.*

Proof: Suppose $U.state = \text{Committed}$. Then a commit step belonging to U must have occurred. Since each commit step is preceded by a frozen step, Lemma 3.15 implies that no abort step belongs to U . Thus, $U.state$ can never be set to Aborted.

Now suppose that $U.state = \text{Aborted}$. Then either U is the dummy SCX-record or an abort step belongs to U . If U is the dummy SCX-record then, by Lemma 3.5, no commit step belongs to U , so $U.state$ never changes to Committed. Otherwise, $U.state$ is initially InProgress, so there must have been an abort step belonging to U . Hence, by Lemma 3.15, no frozen step can belong to U . If there is no frozen step belonging to U , then there can be no commit step belonging to U (by the pseudocode of HELP). Therefore $U.state$ can never be set to Committed. ■

Corollary 3.17 *The changes to the state field of an SCX-record respect Figure 3.7.*

Proof: Immediate from Corollary 3.16 and the fact that an SCX-record's state field cannot change to InProgress from any other state. ■

The next five lemmas prove that any successful freezing CASs belonging to an SCX-record U must occur prior to the first frozen step or abort step belonging to U . This result allows us to fill in the gaps between Figure 3.7 and Figure 3.3(a).

Lemma 3.18 *Let U be an SCX-record, and let $\langle r_1, r_2, \dots, r_l \rangle$ be the sequence of Data-records in $U.V$. Suppose an abort step belongs to U and let $astep$ be the first such abort step. Then, there is a $k \in \{1, \dots, l\}$ such that*

1. *a freezing CAS belonging to U on r_k occurs prior to $astep$,*
2. *there is no successful freezing CAS belonging to U on r_k ,*
3. *for each $i \in \{1, \dots, k-1\}$, a successful freezing CAS belonging to U on r_i occurs prior to $astep$, and no successful freezing CAS belonging to U on r_i occurs after $astep$, and*
4. *$r_k.info$ changes after the LLX(r_k) linked to S reads $r_k.info$ at line 9 and before the first freezing CAS belonging to U on r_k .*

Proof: Let H be the invocation of HELP that performs $astep$ and k be the iteration of the loop in HELP during which H performs $astep$. The loop in H iterates over the sequence r_1, r_2, \dots, r_l of Data-records.

Claim 1. follows from the definition of k : before H performs $astep$, it performs a freezing CAS belonging to H on r_k at line 26.

To derive a contradiction, suppose Claim 2 is false, i.e., there is a *successful* freezing CAS belonging to U on r_k . By Claim 1, $fcas$ is before $astep$. From Corollary 3.16 and the fact that $astep$ occurs, we know that no commit step belongs to U . By Lemma 3.10, $r_k.info$ points to U at all times between the first freezing CAS belonging to U on r_k and $astep$. However, this contradicts the fact that $r_k.info$ does not point to U when H performs line 27 just before performing $astep$.

We now prove Claim 3.. By Claim 1., prior to $astep$, H performs a freezing CAS belonging to U on r_k . By Lemma 3.12, this can only occur after a successful freezing CAS belonging to U on r_i , for all $i < k$. By Lemma 3.8, there is no successful freezing CAS belonging to U on r_i after $astep$.

We now prove Claim 4.. By Claim 1. and Claim 2., an unsuccessful freezing CAS $fcas$ belonging to U on r_k occurs prior to $astep$. By line 25 and Observation 3.2.3, the old value for $fcas$ is read from $r_k.info$ and stored in $rinfo$ at line 4 by the $LLX(r_k)$ linked to S . By Observation 3.2.4, the $LLX(r_k)$ linked to S again sees $r_k.info = rinfo$ at line 9. Thus, since $fcas$ fails, $r_k.info$ must change after the $LLX(r_k)$ linked to S performs line 9 and before $fcas$ occurs. ■

Lemma 3.19 *No freezing CAS belonging to an SCX-record U is successful after the first frozen step or abort step belonging to U .*

Proof: First, suppose a frozen step belongs to U , and let $fstep$ be the first such frozen step. Then, by Lemma 3.11, there is a successful freezing CAS belonging to U on r for each r in $U.V$ that occurs before $fstep$. By Lemma 3.8, only the first freezing CAS belonging to U on r can be successful. Hence, no freezing CAS belonging to U is successful after $fstep$.

Now, suppose an abort step belongs to U , and let $astep$ be the first such abort step. Let $\langle r_1, r_2, \dots, r_l \rangle$ be the sequence of Data-records in $U.V$. By Lemma 3.18, there is a $k \in \{1, \dots, l\}$ such that no successful freezing CAS belonging to U is performed on any $r_i \in \{r_1, \dots, r_{k-1}\}$ after $astep$, and no successful freezing CAS belonging to U on r_k ever occurs. By Lemma 3.12, there is no freezing CAS belonging to U on any $r_i \in \{r_{k+1}, \dots, r_l\}$. ■

Corollary 3.20 *If there is a successful freezing CAS $fcas$ belonging to an SCX-record U on a Data-record r , then $fcas$ occurs before time t , when the first abort step or commit step belonging to U occurs. Moreover, $r.info$ points to U at all times after $fcas$ occurs, and before t .*

Proof: By Lemma 3.19, $fcas$ occurs before t . The claim then follows from Lemma 3.10. ■

Lemma 3.21 *Changes to the state and allFrozen fields of an SCX-record, as well as frozen steps, abort steps, commit steps, mark steps and successful freezing CAs can only occur as depicted in Figure 3.2.*

Proof: Initially, the dummy SCX-record has $state = \text{Aborted}$ and $allFrozen = \text{FALSE}$ and, by Lemma 3.5, they never change.

Every other SCX-record U initially has $state = \text{InProgress}$ and $allFrozen = \text{FALSE}$. Only abort steps, frozen steps, and commit steps can change $state$ or $allFrozen$. From the code of HELP, each transition shown in Figure 3.3(a), results in the indicated values for $state$ and $allFrozen$. A commit step on line 41 must be preceded by a frozen step on line 37. Therefore, from $[\text{InProgress}, \text{FALSE}]$, the only outgoing transitions are to $[\text{Aborted}, \text{FALSE}]$ and $[\text{InProgress}, \text{TRUE}]$. By Lemma 3.15, there cannot be both a frozen step and an abort step belonging to U . Hence, from $[\text{Aborted}, \text{FALSE}]$, there cannot be a frozen step or commit step and there cannot be an abort step from $[\text{InProgress}, \text{TRUE}]$ or $[\text{Committed}, \text{TRUE}]$.

	$r.info.state$	$r.marked$
Frozen	Committed InProgress	TRUE $\{\text{TRUE}, \text{FALSE}\}$
Unfrozen	Committed Aborted	FALSE $\{\text{TRUE}, \text{FALSE}\}$

Figure 3.8: When a Data-record r is frozen, in terms of $r.info.state$ and $r.marked$.

By Lemma 3.19, successful freezing CASs can only occur when $state = \text{InProgress}$ and $allFrozen = \text{FALSE}$. From the code of HELP, for each r in $U.R$, the first mark step belonging to U on r must occur after the first frozen step belonging to U and before the first commit step belonging to U . Since each r in $U.R$ initially has $r.marked = \text{FALSE}$, and $r.marked$ is only changed at line 38, where it is set to TRUE, only the first mark step belonging to U on r can be successful. ■

3.4.4 The period of time over which a Data-record is frozen

We now prove several lemmas which characterize the period of time over which a Data-record is frozen for an SCX-record. We first use the fact that the $state$ of an SCX-record cannot change from Aborted to Committed to extend Lemma 3.9 to prove that the $info$ field of a Data-record cannot be changed while the Data-record is frozen for an SCX-record. In the following, we often say “after X and before the first time Y happens.” In the event that Y never happens, this phrasing should be interpreted to mean simply “after X.”

Lemma 3.22 *If a frozen step belongs to an SCX-record U then, for each r in $U.V$, a freezing CAS belonging to U on r precedes the first frozen step belonging to U , and r is frozen for U at all times after the first freezing CAS belonging to U on r and before the first commit step belonging to U .*

Proof: Fix any r in $U.V$. If a frozen step belongs to U then, by Lemma 3.11, it is preceded by a successful freezing CAS belonging to U on r . Further, by Lemma 3.15, no abort step belongs to U . Thus, by Corollary 3.20, $r.info$ points to U at all points between time t_0 , when the first freezing CAS belonging to U on r occurs, and time t_1 , when the first commit step belonging to U occurs (after the first frozen step). Since no abort step belongs to U , $U.state = \text{InProgress}$ at all times before t_1 . Hence, by the definition of freezing (see Figure 3.8), r is frozen for U at all times between t_0 and t_1 . ■

Corollary 3.23 *If a frozen step belongs to an SCX-record U , then each r in $U.V$ is frozen for U at all times between the first frozen step belonging to U and the first commit step belonging to U .*

Proof: Suppose there is a frozen step belonging to U . By Lemma 3.22, each r in $U.V$ is frozen for U at all times between the first freezing CAS belonging to U on r and the first commit step belonging to U . It then follows directly from the pseudocode of HELP that the first frozen step belonging to U must follow the first freezing CAS belonging to U on r , for each r in $U.V$, and precede the first commit step belonging to U . ■

Corollary 3.24 *A successful mark step belonging to U can occur only while r is frozen for U .*

Proof: Immediate from Lemma 3.21 and Corollary 3.23. ■

Lemma 3.25 A Data-record can only be changed from unfrozen to frozen by a change in its *info* field (which can only be the result of a freezing CAS).

Proof: Let r be a Data-record whose *info* field points to an SCX-record U . According to the definition of a frozen Data-record (see Figure 3.8), if $r.info$ does not change, then r can only become frozen if $U.state$ changes from Committed or Aborted to InProgress, or from Aborted to Committed (provided r is marked). However, both cases are impossible by Corollary 3.17. ■

Definition 3.26 A Data-record r is called **permafrozen** for SCX-record U if r is marked, $r.info$ points to U and the $U.state$ is Committed. Notice that a Data-record that is permafrozen for U is also frozen for U .

Lemma 3.27 Once a Data-record r is permafrozen for SCX-record U , it remains permafrozen for U thereafter.

Proof: By definition, when r is permafrozen for U , it is frozen for U , $U.state$ is Committed and $r.marked = \text{TRUE}$. Once $r.marked$ is set to TRUE , it can never be changed back to FALSE . By Corollary 3.17, $U.state$ was never Aborted, $U.state$ will remain Committed forever, and r will be frozen for U as long as $r.info$ points to U . It remains only to prove that $r.info$ cannot change while r is permafrozen for U . Note that $r.info$ can be changed only by a successful freezing CAS.

To obtain a contradiction, suppose a freezing CAS $fcas$ changes $r.info$ from U to W while r is permafrozen for U . By Lemma 3.5, W is not the dummy SCX-record. Let S be the invocation of SCX that created W . From the code of HELP, r is in $W.V$. So, by the precondition of SCX, there is an invocation of $\text{LLX}(r)$ linked to S . By Observation 3.2.3 and line 25, the old value for $fcas$ (a pointer to U) was read at line 4 of the $\text{LLX}(r)$ linked to S . Let I be the invocation of $\text{LLX}(r)$ linked to S . Since we have argued that $U.state$ is never Aborted, $U.state \in \{\text{InProgress}, \text{Committed}\}$ when I reads $state$ from $U.state$ at line 5.

If $state = \text{InProgress}$ then I does not enter the if-block at line 7, and returns FAIL or FINALIZED, which contradicts Definition 3.1.1.

Now, consider the case where $state = \text{Committed}$. If we can argue that r is marked when I performs line 6, then we shall obtain the same contradiction as in the previous case. Since $state = \text{Committed}$, a commit step belonging to U occurs before I performs line 5. By Lemma 3.21, any successful mark step belonging to U occurs prior to this commit step. Therefore, if r is in $U.R$, then r will be marked when I performs line 6, and we obtain the same contradiction. The only remaining possibility is that r is not in $U.R$, and r is marked by a successful mark step $mstep$ belonging to some other SCX-record U' after I performs line 6, and before $fcas$ occurs (which is while r is permafrozen for U). Since $r.info$ points to U when I performs line 4, and again when $fcas$ occurs, Lemma 3.6 implies that $r.info$ points to U throughout this time. However, this contradicts Corollary 3.24, which states that $mstep$ can only occur while $r.info$ points to U' . ■

Lemma 3.28 Suppose a successful mark step $mstep$ belonging to an SCX-record U on r occurs. Then, r is frozen for U when $mstep$ occurs, and forever thereafter.

Proof: By Corollary 3.24, $mstep$ must occur while r is frozen for U . From the code of HELP, a frozen step belonging to U must precede $mstep$, and r must be in V (since it is marked at line 38). Thus, Corollary 3.23 implies that r is frozen for U at all times between $mstep$ and the first commit step belonging to U . Since $r.marked$ is never changed from TRUE to FALSE , $mstep$ must be the first mark step that ever modifies $r.marked$. From the code of HELP, $mstep$

must precede the first commit step belonging to U . If any commit step belonging to U occurs after $mstep$, immediately after the first such commit step, r will be marked, and $r.info.state$ will be Committed, so r will become permafrozen for U . By Lemma 3.27, r will remain frozen for U , thereafter. ■

Lemma 3.29 *Suppose I is an invocation of $\text{LLX}(r)$ that returns a value different from FAIL or FINALIZED. Then, r is not frozen at any time in $[t_0, t_1]$, where t_0 is when I reads $r.info.state$ at line 5, and t_1 is when I reads $r.info$ at line 9.*

Proof: We prove that r is not frozen at any time between t_0 and t_1 . Since I returns a value different from FAIL or FINALIZED, it enters the if-block at line 7, and sees $r.info = rinfo$ at line 9. Therefore, it sees either $state = \text{Committed}$ and $r.marked = \text{FALSE}$, or $state = \text{Aborted}$ at line 7. In each case, Corollary 3.17 guarantees that $rinfo.state$ will never change again after time t_0 . Thus, if $state = \text{Aborted}$, then r is not frozen at any time between t_0 and t_1 . Now, suppose $state = \text{Committed}$. We prove that $r.marked$ does not change between t_0 and t_1 . A pointer to an SCX-record W is read from $r.info$ and stored in the local variable $rinfo$ at line 4, before t_0 . At line 9, $r.info$ still contains a pointer to W . By Lemma 3.6, $r.info$ must not change between line 4 and line 9. Therefore, $r.info$ points to W at all times between t_0 and t_1 . By Corollary 3.24, a successful mark step can occur between t_0 and t_1 only if it belongs to W . Since $state = \text{Committed}$, a commit step belonging to W must have occurred before t_0 . By Lemma 3.21, any successful mark step belonging to W must have occurred before t_0 . Therefore, $W.state = \text{Committed}$ and $r.marked = \text{FALSE}$ throughout $[t_0, t_1]$. ■

Corollary 3.30 *Let S be an invocation of SCX, and r be any Data-record in the V sequence of S . Then, r is not frozen at any time in $[t_0, t_1]$, where t_0 is when the $\text{LLX}(r)$ linked to S reads $r.info.state$ at line 5, and t_1 is when the $\text{LLX}(r)$ linked to S reads $r.info$ at line 9.*

Proof: Immediate from Definition 3.1.1 and Lemma 3.29. ■

3.4.5 Properties of update CAS steps

Observation 3.31 *An immutable field of a Data-record cannot change from its initial value.*

Proof: This observation follows from the facts that Data-records can only be changed by SCX and an invocation of SCX can only accept a pointer to a mutable field as its fld argument (to modify). ■

Observation 3.32 *Each mutable field of a Data-record can be modified only by a successful update CAS.*

Observation 3.33 *Each update CAS belonging to an SCX-record U is of the form $\text{CAS}(U.fld, U.old, U.new)$. Invariant: $U.fld$ and $U.new$ contain the arguments fld and new , respectively, that were passed to the invocation of $\text{SCX}(V, R, fld, new)$ that created U .*

Proof: An update CAS occurs at line 39 in an invocation of $\text{HELP}(scxPtr)$, where it operates on $scxPtr.fld$, using $scxPtr.old$ as its old value, and $scxPtr.new$ as its new value. The fields of $scxPtr$ do not change after $scxPtr$ is created at line 21. At this line, the arguments fld and new that were passed to the invocation of $\text{SCX}(V, R, fld, new)$ are stored in $scxPtr.fld$ and $scxPtr.new$, respectively. ■

Lemma 3.34 *The first update CAS belonging to an SCX-record U on a Data-record r occurs while r is frozen for U .*

Proof: Let $upcas$ be the first update CAS belonging to U . By line 39, such an update CAS will modify $U.fld$ which, by Observation 3.2.1, is a mutable field of a Data-record r in $U.V$. Since $upcas$ is preceded by a frozen step in the pseudocode of HELP, a frozen step belonging to U must precede $upcas$. Hence, Corollary 3.23 applies, and each r in $U.V$ is frozen for U at all times between the first frozen step belonging to U and the first commit step $cstep$ belonging to U . From the code of HELP, if $cstep$ exists, then it must occur after $upcas$. Thus, when $upcas$ occurs, r is frozen for U . ■

In Section 3.3 we described a constraint on the use of SCX that allows us to implement an optimized version of SCX (which avoids the creation of a new Data-record to hold each value written to a mutable field), and noted that the correctness of the unoptimized version follows trivially from the correctness of the optimized version. In order to prove the next few lemmas, we must invoke this constraint. In fact, we assume a weaker constraint, and are still able to prove what we would like to. We now give this weaker constraint, and remark that it is automatically satisfied if the constraint in Section 3.3 is satisfied.

Constraint 3.35 *Let fld be a mutable field of a Data-record r . If an invocation S of $\text{SCX}(V, R, fld, new)$ is linearized, then:*

- *new is not the initial value of fld , and*
- *no invocation of $\text{SCX}(V', R', fld, new)$ is linearized before the $\text{LLX}(r)$ linked to S is linearized.*

We prove the following six lemmas solely to prove that only the first update CAS belonging to an SCX-record can succeed. This result is eventually used to prove that exactly one successful update CAS belongs to any SCX-record which is helped to *successful* completion.

We need to know about the linearization of SCXs and linked LLXs to prove the next lemma, which uses Constraint 3.35. Let S be an invocation of SCX, and U be the SCX-record that it creates. As we shall see in Section 3.4.8, we linearize S if and only if there is an update CAS belonging to U , and S is linearized at its first update CAS. Each invocation of LLX linked to an invocation of SCX is linearized at line 9.

Lemma 3.36 *No two update CASs belonging to different SCX-records can attempt to change the same field to the same value.*

Proof: Suppose, to derive a contradiction, that update CASs belonging to two different SCX-records U and U' attempt to change the same (mutable) field of some Data-record r to the same value. Let $upcas$ and $upcas'$ be the first update CAS belonging to U and U' , respectively. Let S and S' be the invocation of SCX that created U and U' , respectively. From Observation 3.33 and the fact that $upcas$ and $upcas'$ attempt to change the same field to the same value, we know that S and S' must have been passed the same fld and new arguments. Note that S and S' are linearized at $upcas$ and $upcas'$, respectively. Without loss of generality, suppose S is linearized after S' . By Constraint 3.35, S' is linearized after the invocation I of $\text{LLX}(r)$ linked to S is linearized.

By Lemma 3.30, r is not frozen when I is linearized. By Lemma 3.34, r is frozen for U' when $upcas'$ occurs (which is after I is linearized). By Lemma 3.25, r can become frozen for U' only by a successful freezing CAS belonging to U' on r . Therefore, a successful freezing CAS $fcas'$ belonging to U' on r occurs after I is linearized, and before $upcas'$. By Lemma 3.34, r is frozen for U when $upcas$ occurs (which is after $upcas'$), which implies that a successful freezing CAS $fcas$ belonging to U on r occurs after $upcas'$, and before $upcas$. To recap, I is linearized before $fcas'$,

which is before $upcas'$, which is before $fcas$, which is before $upcas$. By line 25 and Observation 3.2.3, the old value old for $fcas$ is read from $r.info$ and stored in $r.info$ at line 4 by I . Since I performs line 4 before it is linearized, old is read from $r.info$ before $fcas'$. Since $fcas'$ changes $r.info$ to point to U' , Lemma 3.6 implies that $r.info$ does not point to U' at any time before $fcas'$. Therefore, old is not U' . Since $fcas$ is successful, $r.info$ must be changed to old at some point after $fcas'$, and before $upcas$. However, this contradicts Lemma 3.6, since $r.info$ had already contained old before $fcas'$. ■

Lemma 3.37 *An update CAS never changes a field back to its initial value.*

Proof: By Observation 3.33, each update CAS belonging to an SCX-record U attempts to change a field to the value new that was passed as an argument to the invocation of SCX that created U . Since Constraint 3.35 implies that new cannot be the initial value of the field, we know that no update CAS can change the field to its initial value. ■

Lemma 3.38 *No update CAS has equal old and new values.*

Proof: Let $upcas$ be an update CAS and let U be the SCX-record to which it belongs. By Observation 3.33, the old value used by $upcas$ is $U.old$, and the new value used by $upcas$ is $U.new$. Let f be the field of a Data-record pointed to by $U.fld$; this is the field to which $upcas$ is applied. By Lemma 3.37, $U.new$ cannot be the initial value of f . If $U.old$ is the initial value of f , then we are done. So, suppose $U.old$ is not the initial value of f . Since a mutable field can only be changed by a successful update CAS, there exists a successful update CAS $upcas'$ which changed f to $U.old$ prior to $upcas$. By Observation 3.2.2, $U.old$ was read from f prior to the start of the invocation S of SCX that created U and, therefore, prior to $upcas$. Hence, when $upcas'$ occurs, U has not yet been created. Note that $upcas'$ must occur in an invocation of $\text{HELP}(ptr')$ where ptr' points to some SCX-record U' different from U . However, by Lemma 3.36, $upcas$ and $upcas'$ use different new values, so $U.old$ (the new value for $upcas'$) must be different from $U.new$ (the new value for $upcas$). ■

Lemma 3.39 *At most one successful update CAS can belong to an SCX-record.*

Proof: We prove this lemma by contradiction. Consider the earliest point in the execution when the lemma is violated. Let $upcas'_U$ be the earliest occurring second successful update CAS belonging to any SCX-record, and U be the SCX-record to which it belongs, and let $upcas_U$ be the preceding successful update CAS belonging to U . Further, let f be the field upon which $upcas'_U$ operates, and let old and new be the old and new values used by $upcas_U$, respectively. (By Observation 3.33, $upcas_U$ and $upcas'_U$ attempt to change the same field from the same old value to the same new value.) By Lemma 3.38, we know that $old \neq new$. Then, since $upcas'_U$ is successful, there must be a successful update CAS $upcas'_W$ belonging to some SCX-record W which changes f to old between $upcas_U$ and $upcas'_U$. By Lemma 3.37, old is not the initial value of f . Hence, there must be another successful update CAS $upcas_W$ which changes f to old before $upcas_U$. By Lemma 3.36, $upcas_W$ must belong to W , so $upcas_W$ and $upcas'_W$ both precede $upcas'_U$. This contradicts the definition of $upcas'_U$. ■

Lemma 3.40 *An update CAS never changes a field to a value that has already appeared there. (Hence, there is no ABA problem on mutable fields.)*

Proof: Suppose a successful update CAS $upcas$ belonging to an SCX-record U changes a field f to have value new . By Lemma 3.37, new is not the initial value of f . By Lemma 3.39, all successful update CASs that change f must belong to different SCX-records. Hence, Lemma 3.36 implies that no update CAS other than $upcas$ can change f to new . ■

Lemma 3.39 proved that at most one update CAS of each SCX-record can succeed. Now we prove that such a successful update CAS must be the *first* one belonging to SCX-record.

Lemma 3.41 *Only the first update CAS belonging to an SCX-record U can succeed.*

Proof: Let $upcas$ be the first update CAS belonging to U , and f be the field that $upcas$ attempts to modify. If $upcas$ succeeds then, by Lemma 3.39, there can be no other successful update CAS belonging to U . So, suppose $upcas$ fails. By Observation 3.33, each update CAS belonging to U uses the same old value $U.old$. By Observation 3.2.2, $U.old$ was read from f prior to the start of the invocation S of SCX that created U and, therefore, prior to $upcas$. Then, since $upcas$ fails, f must change between when $U.old$ is read from f and when $upcas$ occurs. By Observation 3.32, f can only be changed by an update CAS. By Lemma 3.40, each update CAS applied to f changes it to a value that it has not previously contained. Therefore, f will never again be changed to $U.old$. Hence, every subsequent update CAS belonging to U will fail. ■

3.4.6 Freezing works

In addition to being used to prove the remaining lemmas of this section, the following two results are used to prove linearizability in Section 3.4.8. Intuitively, they allow us to determine whether a Data-record has changed simply by looking at its *info* field, and whether it is frozen.

Corollary 3.42 *An update CAS belonging to an SCX-record U on a Data-record r can succeed only while r is frozen for U .*

Proof: Suppose a successful update CAS $upcas$ belongs to an SCX-record U . By Lemma 3.41, it is the first update CAS belonging to U . The claim follows from Lemma 3.34. ■

By Observation 3.32, a mutable field of r can only change while r is frozen.

Lemma 3.43 *If a Data-record r is not frozen at time t_0 , $r.info$ points to an SCX-record U at or before time t_0 , and $r.info$ points to U at time $t_1 > t_0$, then no field of r is changed during $[t_0, t_1]$.*

Proof: Since $r.info$ points to U at or before time t_0 , and again at time t_1 , Lemma 3.6 implies that $r.info$ must point to U at all times in $[t_0, t_1]$. Further, from Lemma 3.25, r can only be changed from unfrozen to frozen by a change to $r.info$. Therefore, at all times in $[t_0, t_1]$, r is not frozen. By Corollary 3.42, and Observation 3.32, each mutable field of r can change only while r is frozen. By Corollary 3.24, $r.marked$ can change only while r is frozen. Finally, by Observation 3.31, immutable fields do not ever change. Hence, no field of r changes during $[t_0, t_1]$. ■

The remaining results of this section describe intervals over which certain fields of a Data-record do not change. Suppose U is an SCX-record created by an invocation S of SCX, r is a Data-record in $U.V$, and I is the invocation of $LLX(r)$ linked to S . Intuitively, we use the preceding lemma to prove, over the next two lemmas, that no field of r changes between when I last reads $r.info$, and when r becomes frozen for U . We then use this result in Section 3.4.7 to prove that S succeeds if and only if this holds for each r in V . The remaining results of this section are used primarily to prove that exactly one successful update CAS belongs to U if a frozen step belongs to U (and S does not crash, or some process helps it complete).

Corollary 3.44 *Let U be an SCX-record, and S be the invocation of SCX that created U . If there is a successful freezing CAS belonging to U on r , then no field of r changes after the $\text{LLX}(r)$ linked to S reads $r.info.state$ at line 5, and before this freezing CAS occurs.*

Proof: Let $fcas$ be a successful freezing CAS belonging to U on r . Note that the $\text{LLX}(r)$ linked to S terminates before S begins. Since S creates U and $fcas$ changes $r.info$ to point to U , S begins before $fcas$. We now check that Lemma 3.43 applies. By Corollary 3.30, r is not frozen when the $\text{LLX}(r)$ linked to S executes line 5. From line 25 of HELP and Observation 3.2.3, we know the old value $r.info$ for $fcas$ is read from $r.info$ at line 4 by the $\text{LLX}(r)$ linked to S . Further, since $fcas$ succeeds, $r.info$ contains $r.info$ just prior to $fcas$. Thus, Lemma 3.43 applies, and proves the claim. ■

Lemma 3.45 *If an update CAS belongs to an SCX-record U then, for each r in $U.V$, there is a successful freezing CAS belonging to U on r , and no mutable field of r changes during $[t_0(r), t_1]$, where $t_0(r)$ is when the first such freezing CAS occurs, and t_1 is when the first update CAS belonging to U occurs.*

Proof: Suppose an update CAS belongs to an SCX-record U . Let $upcas$ be the first such update CAS. Since each update CAS is preceded in the code by a frozen step, a frozen step also belongs to U . Fix any r in $U.V$. By Lemma 3.15, there is a successful freezing CAS belonging to U on r . By Lemma 3.22, r is frozen for U at all times in $[t_0(r), t_2]$, where t_2 is when the first commit step belonging to U occurs. Since an update CAS belonging to an SCX-record W can modify r only while r is frozen for W (by Corollary 3.42), any update CAS that modifies r during $[t_0(r), t_2]$ must belong to U . From the code of HELP, $t_0(r) < t_1 < t_2$. However, since the first update CAS belonging to U occurs at t_1 , no update CAS belonging to U can occur during $[t_0(r), t_1]$. ■

Lemma 3.46 *Let U be an SCX-record created by an invocation S of SCX, and r be a Data-record in $U.V$. Let t_0 be when the $\text{LLX}(r)$ linked to S reads $U.state$ at line 5, and t_2 be when the first commit step belonging to U occurs. If an update CAS belongs to U then, for each r in $U.V$, between t_0 and t_2 , $r.marked$ can be changed (from FALSE to TRUE, by the first mark step belonging to U on r) only if r is in $U.R$.*

Proof: Fix any r in $U.V$. The fact that $r.marked$ can be changed only from FALSE to TRUE follows immediately from the fact that $r.marked$ is initially FALSE, and is only changed at line 38. It also follows that a successful mark step on r must be the first mark step on r . The rest of the claim is more subtle. Suppose a successful mark step $mstep$ belonging to an SCX-record W on r occurs during (t_0, t_2) . By Lemma 3.28, r is frozen for W when $mstep$ occurs. From the code of HELP, a frozen step $fstep$ belonging to U precedes the first update CAS belonging to U , and a freezing CAS belonging to U on r precedes $fstep$. By Lemma 3.11, there is a successful freezing CAS belonging to U on r . By Lemma 3.8, it must be the first freezing CAS $fcas$ belonging to U on r . Let t_1 be when $fcas$ occurs. Note that $t_0 < t_1 < t_2$. By Corollary 3.44, r does not change during $[t_0, t_1]$. Thus, $mstep$ cannot occur in $[t_0, t_1]$. By Lemma 3.22, r is frozen for U at all times during $[t_1, t_2]$. This implies $U = W$, so $mstep$ is the first mark step belonging to U on r . Finally, since there is a mark step belonging to U on r , we obtain from line 38 that r is in $U.R$. ■

3.4.7 Correctness of HELP

The following lemma shows that a helper of an SCX-record cannot return TRUE until after the SCX-record is Committed. We shall use this to ensure that the SCX does not return TRUE until after the SCX has taken effect.

Lemma 3.47 *An invocation of $\text{HELP}(\text{scxPtr})$ where scxPtr points to an SCX-record U cannot return from line 31 before the first commit step belonging to U .*

Proof: Suppose an invocation H of $\text{HELP}(\text{scxPtr})$ returns at line 31. Before returning, H sees $r.info \neq \text{scxPtr}$ at line 27, which implies that $r.info$ does not point to U . Prior to this, H performs a freezing CAS belonging to U at line 26. By line 29, a frozen step belongs to U . Then, since Lemma 3.22 states that $r.info$ points to U at all times between the first freezing CAS belonging to U and the first commit step belonging to U , a commit step belonging to U must occur before H returns. ■

Next, we obtain an exact characterization of the update CAS steps that succeed.

Lemma 3.48 *If there is an update CAS belonging to an SCX-record U , then the first update CAS belonging to U is successful and changes the mutable field pointed to by $U.fld$ from $U.old$ to $U.new$. No other update CAS belonging to U is successful.*

Proof: Let $t_0(r)$ be when the $\text{LLX}(r)$ linked to S reads $r.info.state$ at line 5, and t_1 be when the first update CAS belonging to U occurs. Since an update CAS belongs to U , Lemma 3.45 implies that, for each r in $U.V$, no mutable field of r changes between $t_0(r)$ and t_1 . From Observation 3.2.2 and the code of LLX , we see that the value stored in $U.old$ is read from the field pointed to by $U.fld$ after time $t_0(r)$. Further, since the $\text{LLX}(r)$ linked to S terminates before S begins (by the definition of an LLX linked to an SCX) and, in turn, before U is created, we know the value stored in $U.old$ was read before any update CAS belonging to U occurred. Then, since $U.fld$ is a mutable field of a Data-record in $U.V$, this field does not change between $t_0(r)$ and t_1 . By Observation 3.33, any update CAS belonging to U will attempt to change $U.fld$ from $U.old$ to $U.new$, so the first update CAS belonging to U will succeed. Lemma 3.41 completes the proof. ■

Our next lemma shows that the SCXs that are linearized have the desired effect.

Lemma 3.49 *Let U be an SCX-record created by an invocation S of SCX , and ptr point to U . If either*

- a frozen step belongs to U and some invocation of $\text{HELP}(\text{ptr})$ terminates, or
- S or any invocation of $\text{HELP}(\text{ptr})$ returns TRUE,

then the following claims hold.

1. *Every invocation of $\text{HELP}(\text{ptr})$ that terminates returns TRUE.*
2. *Exactly one successful update CAS belongs to U , and it is the first update CAS belonging to U . It changes the mutable field pointed to by $U.fld$ from $U.old$ to $U.new$.*
3. *A frozen step of U and a commit step of U occur before any invocation of $\text{HELP}(\text{ptr})$ returns.*
4. *At all times after the first commit step for U , each r in $U.R$ is permafrozen for U .*

Proof: We first simplify the lemma's hypothesis. If S returns TRUE, then S 's invocation of $\text{HELP}(\text{ptr})$ has returned TRUE. If some invocation of $\text{HELP}(\text{ptr})$ returns TRUE, a commit step belongs to U by Lemma 3.47, and that commit step is preceded by a frozen step of U . So, for the remainder of the proof, we can assume that a frozen step belongs to U and some invocation of $\text{HELP}(\text{ptr})$ terminates.

Proof of Claim 1: Since a frozen step belongs to U , Lemma 3.15 implies that no abort step belongs to U . Thus, H cannot return at line 35, which implies that H must return TRUE.

Proof of Claim 2 and Claim 3: By Claim 1, H must return TRUE. If H returns at line 42, then it does so after performing an update CAS and a commit step, each belonging to U . Otherwise, H returns at line 31. However, by Lemma 3.47, no invocation of $\text{HELP}(ptr)$ can return at line 31 until the first commit step belonging to U , which is necessarily preceded by an update CAS for U (by inspection of HELP). Thus, Claim 3 is proved. Lemma 3.48 proves Claim 2.

Proof of Claim 4: By Corollary 3.23, every r in $U.V$ is frozen for U from the first frozen step belonging to U until the first commit step belonging to U . From the code of HELP , each r in $U.R$ is marked before the first commit step belonging to U , and Lemma 3.46 implies that they are still marked when the first commit step belonging to U occurs. Further, immediately after the first commit step belonging to U (which must exist by Claim 3), $U.\text{state}$ will be Committed, so each r that is in both $U.V$ and $U.R$ will be permafrozen for U . Since R is a subsequence of V , and $U.R$ and $U.V$ do not change after they are obtained at line 21 from R and V , respectively, it follows from Lemma 3.27 that each r in $U.R$ remains permafrozen for U forever. ■

Now we show that SCXs that are not linearized do not modify any mutable fields, and do not return TRUE.

Lemma 3.50 *Let U be an SCX-record created by an invocation S of SCX, and ptr be a pointer to U . If S or any invocation H of $\text{HELP}(ptr)$ returns FALSE, then the following claims hold.*

1. Every invocation of $\text{HELP}(ptr)$ that terminates returns FALSE.
2. An abort step belonging to U occurs before any invocation of $\text{HELP}(ptr)$ returns.
3. No update CAS belongs to U .

Proof: Note that, if S returns FALSE, then its invocation of $\text{HELP}(ptr)$ returns FALSE, so an invocation H exists and returns FALSE. By Lemma 3.49.1, if any invocation of $\text{HELP}(ptr)$ returned TRUE, then H would have to return TRUE. Since H returns FALSE, every terminating invocation of $\text{HELP}(ptr)$ must return FALSE, which proves Claim 1. We now prove Claim 2 and Claim 3. Consider the invocation H' of $\text{HELP}(ptr)$ that returns earliest. By Claim 1, H' returns FALSE. Before H' returns FALSE, an abort step belonging to U is performed at line 34. Thus, Lemma 3.15 implies that no frozen step belongs to U . By the code of HELP , each update CAS belonging to U follows a frozen step belonging to U . ■

Now that we have proved each invocation of HELP that returns TRUE or FALSE has its expected effect, we must prove that the return value is always correct. (Otherwise, for example, two invocations of SCX with overlapping V sequences could interfere with one another, but still return TRUE.)

Lemma 3.51 *Let U be an SCX-record created by an invocation S of SCX, and ptr be a pointer to U . Any invocation H of $\text{HELP}(ptr)$ that terminates returns TRUE if no r in $U.V$ changes from when the $\text{LLX}(r)$ linked to S reads $r.\text{info}$ at line 9 at time $t_0(r)$ to when the first freezing CAS belonging to U on r at time $t_1(r)$. Otherwise, H returns FALSE.*

Proof: Since H terminates, it must return TRUE or FALSE. Hence, it suffices to prove H returns TRUE if and only if no r in $U.V$ changes between $t_0(r)$ and $t_1(r)$.

Case I: Suppose H returns TRUE. By Lemma 3.49.3, a frozen step belongs to U . Hence, by Lemma 3.11, there is a successful freezing CAS belonging to U for each r in $U.V$. Then, by Corollary 3.44, no field of r changes between $t_0(r)$ and $t_1(r)$.

Case II: Suppose H returns FALSE. We show that some r in $U.V$ changes between $t_0(r)$ and $t_1(r)$. Since H can only return FALSE at line 35, immediately before H returns, it performs an abort step belonging to U . Thus, Lemma 3.18.4. applies and the claim is proved. ■

3.4.8 Linearizability of LLX/SCX/VLX

As described in Section 3.2, we linearize all reads, all invocations of LLX that do not return FAIL, all invocations of VLX that return TRUE, and all invocations of SCX that modify the sequential data structure (all that return TRUE, and some that do not terminate). In our implementation, subtle interactions with a concurrent invocation of SCX can cause an invocation of LLX to return FAIL. Since this cannot occur in a linearized execution, we do not linearize any such invocation of LLX. Similarly, we allow some invocations of SCX and VLX to return FALSE because of interactions with concurrent invocations of SCX. Rather than distinguishing between the invocations of SCX or VLX that return FALSE because of an earlier linearized invocation of SCX (which is allowed by the sequential specification), and those that return FALSE because of contention, we simply opt not to linearize any invocation of SCX or VLX that returns FALSE. Alternatively, we could have accounted for the invocations that returns FALSE because of contention by allowing spurious failures in the sequential specification of the operations. However, this would unnecessarily complicate the sequential specification. Intuitively, an algorithm designer using LLX, SCX and VLX is most likely to be interested in invocations of SCX and VLX that return TRUE since these, respectively, change the sequential data structure, and indicate that a set of Data-record have not changed since they were last passed to successful invocations of LLX by this process. Knowing whether an invocation of SCX or VLX was unsuccessful because of a change to the sequential data structure, or because of contention, is less likely to be useful.

Before we give the linearization points, we state precisely which invocations of SCX we shall linearize. Let S be an invocation of SCX, and U be the SCX-record it creates. We linearize S if and only if there is an update CAS belonging to U . By Lemma 3.49, every successful invocation of SCX will be linearized. (By Lemma 3.50, no unsuccessful invocation of SCX will be linearized.)

We first give the linearization points of the operations, then we prove our LLX/SCX/VLX implementation respects the correctness specification given in Section 3.2.

Linearization points:

- An $\text{LLX}(r)$ that returns values at line 11 is linearized at line 9. We linearize an $\text{LLX}(r)$ that returns FINALIZED at line 14.
- Let U be an SCX-record created by an invocation S of SCX. Suppose there is an update CAS belonging to U . We linearize S at the first such update CAS (which is the unique successful update CAS belonging to U , by Lemma 3.48).
- An invocation I of VLX that returns TRUE is linearized at the first execution of line 47.
- We assume reads are atomic. Hence, a read is simply linearized when it occurs.

Lemma 3.52 *The linearization point of each linearized operation occurs during the operation.*

Proof: This is trivial to see for reads, and invocations of LLX and VLX. Let U be an SCX-record created by a linearized invocation S of SCX, and ptr be a pointer to U . This claim is not immediately obvious for S , since the first update CAS belonging to U may be performed by a process helping U to complete (not the process performing S). Since S is linearized, there is an update CAS $upcas$ belonging to U , which can only occur in an invocation of $\text{HELP}(ptr)$. Since ptr points to U , which is created by S , $upcas$ must occur after the start of S . If S does not terminate, then we are done. Otherwise, Lemma 3.49.3 implies that a commit step belongs to U , and the first such commit step occurs before any invocation of $\text{HELP}(ptr)$ returns. From the code of HELP , $upcas$ must occur before the first commit step belonging to U , so $upcas$ must occur before any invocation of $\text{HELP}(ptr)$ returns. Finally, since S invokes $\text{HELP}(ptr)$, $upcas$ must occur before S terminates. ■

We first show that each read returns the correct result according to its linearization point.

Lemma 3.53 *If a read R_f of a field f is linearized after a successful invocation of $\text{SCX}(V, R, fld, new)$, where fld points to f , then R_f returns the parameter new of the last such SCX. Otherwise, R_f returns the initial value of f .*

Proof: We proceed by cases.

Case I: f is an immutable field. In this case, a pointer to f cannot be the fld parameter of an invocation of SCX. Further, by Observation 3.31, f cannot be modified after its initialization, so R_f returns the initial value of f .

Case II: f is a mutable field. Suppose there is no successful invocation of $\text{SCX}(V, R, fld, new)$, where fld points to f , linearized before R_f . Since an invocation of SCX is linearized at its first update CAS, there can be no update CAS on f prior to R_f . Since f can only be modified by successful update CAS, R_f must return the initial value of f .

Now, suppose there is a successful invocation of $\text{SCX}(V, R, fld, new)$, where fld points to f , linearized before R_f . Let S be the last such invocation of SCX linearized before R_f , and U be the SCX-record it creates. By Lemma 3.49.2, there is exactly one successful update CAS $upcas$ belonging to U , occurring at the linearization point of S . Since each successful update CAS is the linearization point of some invocation of SCX, and no invocation of $\text{SCX}(V', R', fld, new')$ is linearized between S and R_f , no successful update CAS occurs between S and R_f . Since a mutable field can only be changed by update CAS, R_f returns the value stored by the successful update CAS $upcas$ belonging to U . By Lemma 3.49.2, $upcas$ changes f to $U.new$. Finally, since $U.new$ does not change after it is obtained from new at line 21, the claim is proved. ■

Next, we prove that an LLX that returns a snapshot does return the correct result according to its linearization point.

Corollary 3.54 *Let r be a Data-record with mutable fields f_1, \dots, f_y , and I be an invocation of $\text{LLX}(r)$ that returns a tuple of values $\langle m_1, \dots, m_y \rangle$ at line 11. For each mutable field f_i of r , if I is linearized after a successful invocation of $\text{SCX}(V, R, fld, new)$, where fld points to f_i , then m_i is the parameter new of the last such invocation of SCX. Otherwise, m_i is the initial value of f_i .*

Proof: Since I returns at line 11, the same value is read from $r.info$ on line 4 and line 9 at times t_0 and t_1 , respectively. By Lemma 3.29, r is unfrozen at line 7, which is between t_0 and t_1 . Thus, Lemma 3.43 implies that r does not change during $[t_0, t_1]$. Since the values returned by the LLX are read from the fields of r between t_0 and t_1 (at line 8), each read returns the same result as it would if it were executed atomically at the linearization point of the LLX (line 9). Finally, Lemma 3.53 completes the proof. ■

Next, we show that an $\text{LLX}(r)$ returns FINALIZED only if r really has been finalized.

Lemma 3.55 *Let I be an invocation of $\text{LLX}(r)$, and U be the SCX-record to which I reads a pointer at line 4. If I returns FINALIZED, then an invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$ that created U is linearized before I , and r is in R .*

Proof: Suppose I returns FINALIZED. Then, I is linearized at line 14. At line 12, I will either see $\text{state} \neq \text{InProgress}$ or will invoke $\text{HELP}(r.info)$ (where $r.info$ is a pointer to U). Thus, when I is linearized, $U.\text{state}$ is either Aborted or Committed. Since I sees $\text{marked}_1 = \text{TRUE}$, r is marked when I performs line 3. By Lemma 3.21, once r is marked, it can never again point to an SCX-record with state Aborted. Therefore, $U.\text{state}$ must be Committed when I is linearized. Since $U.\text{state} = \text{Committed}$, Lemma 3.5 implies that U is not the dummy SCX-record, so there must be an invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$ that created U . From the code of HELP, an update CAS belonging to U occurs before the first commit step belonging to U . Since S is linearized at its first update CAS, S is linearized before I .

It remains to show that r is in R . Since $U.R$ does not change after it is obtained from R at line 21, it suffices to show r is in $U.R$. By line 13, $\text{marked}_1 = \text{TRUE}$, which means that r is marked when I reads a pointer to U from $r.info$ at line 4. Let t_0 be when I performs line 4, and t_1 be when the first commit step belonging to U occurs. We consider two cases. Suppose $t_1 < t_0$. Then, when I performs line 4, r is marked, $r.info$ points to U , and $U.\text{state} = \text{Committed}$, which means that r is permafrozen for U . By Lemma 3.27, r is frozen for U at all times after t_0 . Now, suppose $t_0 < t_1$. In this case, Lemma 3.21 implies $U.\text{state} = \text{InProgress}$ when I performs line 4, and that $U.\text{state}$ will never be Aborted. By Lemma 3.9, $r.info$ must point to U at all times in $[t_0, t_1]$. Thus, at t_1 , r is marked and $r.info$ points to U , which means that r is permafrozen for U . By Lemma 3.27, r is frozen for U at all times after t_1 . In each case, r is frozen for U at all times in some (non-empty) suffix of the execution. Since r is marked, there must be a successful mark step belonging to some SCX-record W on r . By Lemma 3.28, r is frozen for W at all times after this mark step. Therefore, $U = W$, which means there is a mark step belonging to U on r . Finally, line 38 implies that r is in $U.R$. ■

Lemma 3.56 *Let r be a Data-record, I be an invocation of $\text{LLX}(r)$ that terminates, and S be a linearized invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$ with r in R . I returns FINALIZED if it is linearized after S , or begins after S is linearized. (This implies I will be linearized in both cases.)*

Proof: Let U be the SCX-record created by S .

Case I: I begins after S is linearized. In this case, S is linearized at a successful update CAS upcas belonging to U that occurs before I begins. From the code of HELP, a mark step on r belonging to U must occur before upcas . Consider the first mark step $mstep$ on r . Since $r.\text{marked}$ is initially FALSE, $mstep$ must be successful. Let W be the SCX-record to which $mstep$ belongs. By Lemma 3.28, r is frozen for W at all times after $mstep$. By Lemma 3.42, r is frozen for U when upcas occurs (which is after $mstep$). Since r can be frozen for only one SCX-record at a time, $W = U$. Therefore, $r.info$ points to U throughout I . So, when I performs line 7, it will see $\text{state} = \text{Committed}$ and $\text{marked}_2 = \text{TRUE}$. Moreover, when it subsequently performs line 13, it will see $\text{marked}_1 = \text{TRUE}$, so it will return FINALIZED.

Case II: I is linearized after S . I can either be linearized at line 9 or at line 14. If it is linearized at line 14, then it returns FINALIZED, and we are done. Suppose, in order to derive a contradiction, that I is linearized at line 9. Then, I returns at line 11 and, by Corollary 3.30, r is unfrozen at all times during $[t_0, t_1]$, where t_0 is when I performs line 5, and t_1 is when I performs line 9. Since I is linearized at time t_1 , S must be linearized at an update CAS upcas belonging to U that occurs before time t_1 . By Corollary 3.42, upcas can only occur while r is frozen for U . Since r is unfrozen at all times during $[t_0, t_1]$, upcas must occur at some point before t_0 . From the code of HELP, a mark step belonging to U

must occur before *upcas*. Consider the first mark step *mstep* belonging to any SCX-record *W* on *r*. As we argued in the previous case, *r* is frozen for *W* at all times after *mstep*. However, this contradicts our argument that *r* is unfrozen at all times during $[t_0, t_1]$ (since t_0 is after *mstep*). Thus, *I* cannot be linearized at line 9, so *I* must return FINALIZED.

■

The following lemma proves that an SCX succeeds only when it is supposed to, according to the correctness specification.

Lemma 3.57 *If an invocation *I* of VLX(*V*) or SCX(*V, R, fld, new*) is linearized then, for each *r* in *V*, no SCX(*V', R', fld', new'*) with *r* in *V'* is linearized between the LLX(*r*) linked to *I* and *I*.*

Proof: Fix any *r* in *V*. By the preconditions of SCX and VLX, there must be an LLX(*r*) linked to *I*. If *I* is an invocation of SCX, then let *U* be the SCX-record that it creates. Let *L* be the LLX(*r*) linked to *I*, t_0 be when *L* performs line 4, t_1 be when *L* performs line 5 and t_2 be when *L* is linearized (at line 9). Since *L* is linked to *I*, t_2 exists. Let t_3 be when *I* is linearized. For SCX, t_3 is when the first update CAS belonging to *U* occurs. For VLX, t_3 is when *I* first performs line 47. Since *I* is linearized, t_3 exists. Finally, we define time t_4 . For SCX, t_4 is when the first commit step belonging to *U* occurs, or the end of the execution if there is no commit step belonging to *U* (or ∞ if the execution is infinite). For VLX, t_4 is when *I* sees *rinfo = r.info* at line 47 (in the iteration for *r*). If *I* is an invocation of VLX then, since *I* is linearized, it returns TRUE. This implies that *I* must see *rinfo = r.info* at line 47 in the iteration for *r*, so t_4 exists. Clearly, t_4 exists if *I* is an invocation of SCX.

We now prove $t_0 < t_1 < t_2 < t_3 < t_4$. From the code of LLX, $t_0 < t_1 < t_2$. Suppose *I* is an invocation of VLX. Since *L* terminates before *I* begins, $t_2 < t_3$. Trivially, $t_3 < t_4$. Now, suppose *I* is an invocation of SCX. Since each update CAS belonging to *U* occurs in an invocation of HELP(*ptr*) where *ptr* points to *U*, and *U* is created during *I*, t_3 must occur after the start of *I*. Since *L* terminates before *I* begins, $t_2 < t_3$. From the code of HELP, the first update CAS belonging to *U* precedes the first commit step belonging to *U* (as well as the other options for t_4), so $t_3 < t_4$.

Next, we prove that, at all times in $[t_1, t_4]$, *r* is either frozen for *U*, or not frozen (i.e., *r* is not frozen for any SCX-record different from *U* at any point during $[t_1, t_4]$). We consider two cases.

Case I: Suppose *I* is an invocation of VLX. Then, *r.info* contains *rinfo* at t_0 , and again at t_4 . By Lemma 3.6, *r.info* must contain *rinfo* at all times in $[t_0, t_4]$. By Corollary 3.30, *r* is unfrozen at time t_1 . By Lemma 3.25, *r* can only be changed from unfrozen to frozen by a change to *r.info*. Since *r* does not change during $[t_0, t_4]$, *r* is unfrozen at all times in $[t_1, t_4]$.

Case II: Suppose *I* is an invocation of SCX. From the code of HELP, a frozen step belonging to *U* precedes the first update CAS belonging to *U*. By Lemma 3.11, a successful freezing CAS belonging to *U* on *r* precedes the first frozen step belonging to *U*. Let t'_2 be when the first successful freezing CAS *fcas* belonging to *U* on *r* occurs. It follows that $t'_2 < t_3$. Since each freezing CAS belonging to *U* occurs in an invocation of HELP(*ptr*) where *ptr* points to *U*, and *U* is created during *I*, each freezing CAS belonging to *U* must occur after the start of *I*. Recall that *L* terminates before the start of *I*. Hence, $t'_2 > t_2$. By Observation 3.2.3 and line 25, the old value for *fcas* is the value *rinfo* that was read from *r.info* at line 4 of *L* (at t_0). Since *fcas* is successful, *r.info* must contain *rinfo* just before *fcas*. Therefore, *r.info* contains *rinfo* at t_0 , and again at t'_2 . By the same argument we made in Case I (but with t'_2 instead of t_4), Lemma 3.6, Corollary 3.30, and Lemma 3.25 imply that *r* is unfrozen at all times in $[t_1, t'_2]$. By Lemma 3.22, *r* is frozen for *U* at all times in $(t'_2, t_4]$, which proves this case.

At last, we have assembled the results needed to obtain a contradiction. Suppose, to derive a contradiction, that an invocation *S* of SCX(*V', R', fld', new'*) with *r* in *V'* is linearized between *L* and *I* (i.e., in (t_2, t_3)). Let *W* be the

SCX-record created by S . S is linearized at the first update CAS $upcas$ belonging to W . From the code of HELP, a frozen step belonging to W precedes $upcas$, and $upcas$ precedes any commit step belonging to W . By Lemma 3.22, a successful freezing CAS $fcas$ belonging to W on r precedes $upcas$, and r is frozen for W at all times after $fcas$, and before the first commit step belonging to W . Therefore, r is frozen for W when $upcas$ occurs in (t_2, t_3) . Since, at all times in $[t_1, t_4]$, r is either frozen for U , or not frozen, we must have $W = U$. This is a contradiction, since $upcas$ occurs before the *first* update CAS belonging to U occurs (at t_3). Thus, S cannot exist. ■

Theorem 3.58 *Our implementation of LLX/SCX/VLX satisfies the correctness specification discussed in Section 3.2. That is, we linearize all successful LLXs, all successful SCXs, a subset of the SCXs that never terminate, all successful VLXs, and all reads, such that:*

1. *Each read of a field f of a Data-record r returns the last value stored in f by a linearized SCX (or f 's initial value, if no linearized SCX has modified f).*
2. *Each linearized LLX(r) that does not return FINALIZED returns the last value stored in each mutable field f of r by a linearized SCX (or f 's initial value, if no linearized SCX has modified f).*
3. *Each linearized LLX(r) returns FINALIZED if and only if it is linearized after an SCX(V, R, fld, new) with r in R .*
4. *If an invocation I of SCX(V, R, fld, new) or VLX(V) returns TRUE then, for all r in V , there has been no SCX(V', R', fld', new') with r in V' linearized since the LLX(r) linked to I .*

Proof: By Lemma 3.52, the linearization point of each operation occurs during that operation. Claim 1 follows immediately from Lemma 3.53. Claim 2 is immediate from Lemma 3.54. The only-if direction of Claim 3 follows from Lemma 3.55, and the if direction follows from Lemma 3.56. Claim 4 is immediate from Lemma 3.57. ■

3.4.9 Progress guarantees

Lemma 3.59 *LLX, SCX and VLX are wait-free*

Proof: The loop in H iterates over the elements of the finite sequence V and performs a constant amount of work during each iteration. If H does not return from the loop, then it performs a constant amount of work after the loop and returns. The claim then follows from the code. ■

Lemma 3.60 *Our implementation satisfies the first progress property in Section 3.2: Each terminating LLX(r) returns FINALIZED if it begins after the end of a successful SCX(V, R, fld, new) with r in R or after another LLX(r) has returned FINALIZED.*

Proof: Consider a terminating invocation I' of LLX(r). If I' begins after the end of a successful SCX(V, R, fld, new) with r in R , the claim follows from Lemma 3.56. If I' begins after another invocation I of LLX(r) has returned FINALIZED, then I is linearized after an invocation S of SCX(V, R, fld, new) with r in R . Since I precedes I' , I' starts after S is linearized. By Lemma 3.56, I' returns FINALIZED. ■

We now begin to prove the non-blocking progress properties. First, we describe how we assign blame to an SCX for each failed invocation of LLX, VLX or SCX.

Definition 3.61 Let I be an invocation of LLX that returns FAIL. If I enters the if-block at line 7, then let U be the SCX-record pointed to by $r.info$ when I performs line 9. Otherwise, let U be the SCX-record pointed to by $r.info$ when I performs line 4. We say I **blames** the invocation S of SCX that created U . (We prove below that S exists.)

Lemma 3.62 If an invocation I of LLX returns FAIL, then it blames some invocation of SCX.

Proof: Suppose I enters the if-block at line 7. Then, I must see $r.info \neq r.info$ at line 9. Let U be the SCX-record pointed to by $r.info$ when I performs line 9. Since I reads $r.info$ from $r.info$ at line 4, $r.info$ must change to point to U between when I performs line 4 and line 9. Thus, there must be a successful freezing CAS belonging to U on r between these two times. Since a successful freezing CAS belongs to U , Lemma 3.5 implies that U cannot be the dummy SCX-record. Therefore, U must be created by an invocation of SCX.

Now, suppose I does not enter the if-block at line 7. Then, from the code of LLX, $r.info.state$ cannot be Aborted when I performs line 5, so the SCX-record pointed to by $r.info$ is not the dummy SCX-record. Since I reads the value stored in $r.info$ from $r.info$ at line 4, the SCX-record pointed to by $r.info$ when I performs line 4 must have been created by an invocation of SCX. ■

Definition 3.63 Let I be an invocation of VLX(V) that returns FALSE, and r be the Data-record in V for which I sees $r.info \neq r.info$ at line 47. Consider the first successful freezing CAS on r between when the $LLX(r)$ linked to I reads $r.info$ at line 4, and when I sees $r.info \neq r.info$ at line 47. Let U be the SCX-record to which this freezing CAS belongs, and S be the invocation of SCX that created U . (We prove below that S exists.) We say I **blames** S for r .

Lemma 3.64 If an invocation I of VLX returns FALSE, then it blames some invocation of SCX.

Proof: Since I returns FALSE, it sees $r.info \neq r.info$ at line 47, for some r in V . Let p be the process that performs I . By line 46, $r.info$ is a copy of r 's $info$ value in p 's local table of LLX results. By the precondition of VLX and the definition of an $LLX(r)$ linked to I , this value is read from $r.info$ at line 4 by the $LLX(r)$ linked to I . Therefore, $r.info$ must change between when the $LLX(r)$ linked to I performs line 4 and when I sees $r.info \neq r.info$ at line 47. Thus, there must be a successful freezing CAS belonging to some SCX-record U on r between these two times. Since a freezing CAS belongs to U , Lemma 3.5 implies that U is not the dummy SCX-record, so U must have been created by an invocation of SCX. ■

Definition 3.65 Let U be an SCX-record created by an invocation S of SCX that returns FALSE, and U' be an SCX-record created by an invocation S' of SCX. Consider the Data-records r that are in both $U.V$ and $U'.V$, and for which there is no successful freezing CAS belonging to U on r . Let r' be the Data-record among these which occurs earliest in $U.V$. We say S **blames** S' for r' if and only if there is a successful freezing CAS on r' belonging to U' , and this freezing CAS is the earliest successful freezing CAS on r' to occur between when the $LLX(r')$ linked to S reads $r'.info$ at line 4 and the first freezing CAS belonging to U on r' .

Lemma 3.66 Let U be an SCX-record created by an invocation S of SCX. If S returns FALSE, then it blames some other invocation of SCX.

Proof: Since S returns FALSE, Lemma 3.50.2 implies that an abort step belongs to U . By Lemma 3.18, there is a Data-record r_k in $U.V$ such that there is a freezing CAS belonging to U on r_k , but no successful one. Moreover, $r_k.info$ changes after time t_1 , when the $LLX(r_k)$ linked to S reads $r_k.info$ at line 9, and before time t_2 , when the first

freezing CAS belonging to U on r_k occurs. Since the $\text{LLX}(r_k)$ linked to S terminates before U is created (at line 21 of S), and a freezing CAS belonging to U can only occur after U is created, we know $t_1 < t_2$. Let t_0 be the time when the $\text{LLX}(r_k)$ performs line 4. Note that $t_0 < t_1 < t_2$. Since $r_k.\text{info}$ can only be changed by a successful freezing CAS, there must be a successful freezing CAS on r_k during (t_1, t_2) . Let $fcas$ be the earliest successful freezing CAS on r_k during (t_0, t_2) , and let U' be the SCX-record to which it belongs. Since $fcas$ occurs before the first freezing CAS belonging to U on r_k , we know that $U \neq U'$. Let

$$\rho = \{r \mid r \text{ is in } U.V \text{ and } r \text{ is in } U'.V \text{ and } \# \text{ successful freezing CAS belonging to } U \text{ on } r\}.$$

By the code of HELP , a freezing CAS belonging to U' can only modify a Data-record in $U'.V$. Thus, $r_k \in \rho$.

We now show r_k is the element of ρ that occurs earliest in $U.V$. Suppose, to derive a contradiction, that some $r_i \in \rho$ comes before r_k in $U.V$. By Lemma 3.18.1. and Lemma 3.18.2., there is an unsuccessful freezing CAS belonging to U on r_k . By Lemma 3.12, before this unsuccessful freezing CAS, there must be a successful freezing CAS belonging to U on r_i . However, this implies $r_i \notin \rho$, which is a contradiction.

Let S' be the invocation of SCX that creates U' . Thus far, we have shown that S blames S' . It remains to show that $S \neq S'$. By Lemma 3.5, a freezing CAS or abort step cannot belong to the dummy SCX-record. Therefore, neither U nor U' can be the dummy SCX-record. Since $U \neq U'$, U and U' must be created by different invocations of SCX. ■

We now prove that an invocation of $\text{LLX}(r)$ can return FAIL only under certain circumstances.

Definition 3.67 Let U be an SCX-record created by an invocation S of SCX. The **threatening section** of S begins with the first freezing CAS belonging to U , and ends with the first commit step or abort step belonging to U .

Lemma 3.68 Let U be an SCX-record created by an invocation S of SCX. The threatening section of S lies within S , and every successful freezing CAS or update CAS belonging to U occurs during S 's threatening section.

Proof: Let ptr be a pointer to U , and t_0 and t_1 be the times when S 's threatening section begins and ends, respectively. Since each freezing CAS, update CAS, abort step or commit step belonging to U occurs in an invocation of $\text{HELP}(ptr)$, and S creates U , we know that these steps can only occur after S begins. Hence, t_0 is after S begins. Clearly every freezing CAS belonging to U occurs after t_0 . From the code of HELP , the first update CAS belonging to U occurs between t_0 and t_1 . By Lemma 3.41, this is the only update CAS belonging to U that can succeed. By Lemma 3.8, no freezing CAS belonging to U can succeed after the first frozen step or abort step belonging to U . From the code of help , the first frozen step belonging to U must occur before the first commit step belonging to U . Thus, every successful freezing CAS belonging to U occurs between t_0 and t_1 . From the code of SCX, S performs an invocation H of $\text{HELP}(ptr)$ before it returns, and H will perform either a commit step or abort step belonging to U , so long as it does not return from line 31. By Lemma 3.47, H cannot return from line 31 until after the first commit step belonging to U . Therefore, a commit step or abort step belonging to U must occur before S terminates, so t_1 is before S terminates. ■

Observation 3.69 Let S be an invocation of $\text{SCX}(V, R, fld, new)$, and U be the SCX-record it creates. If there is a freezing CAS belonging to U on r , then r is in V .

Proof: From the code of HELP , there will only be a freezing CAS belonging to U on r if r is in $U.V$, and line 21 implies that r is in V . ■

Lemma 3.70 *An invocation I of $\text{LLX}(r)$ can return FAIL only if it overlaps the threatening section of some invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$ with r in V .*

Proof: By Lemma 3.62, I blames an invocation S of SCX. Let U be the SCX-record created by S , and ptr be a pointer to U . By Definition 3.61, I reads a pointer to U from $r.\text{info}$. Since U is not the dummy SCX-record, $r.\text{info}$ can only point to U after a successful freezing CAS belonging to U on r . By Observation 3.69, r is in V . We now show that I overlaps the threatening section of S . Consider the two cases of Definition 3.61.

Case I: I enters the if-block at line 7, and reads a pointer to U from $r.\text{info}$ at line 9. In this case, from the code of LLX, we know that $r.\text{info}$ changes between when I performs line 4 and when I performs line 9. Since $r.\text{info}$ can only be changed to point to U by a successful freezing CAS belonging to U , there must be a successful freezing CAS belonging to U during I . By Lemma 3.68, I must overlap the threatening section of S .

Case II: I does not enter the if-block at line 7. Since I reads a pointer to U from $r.\text{info}$ at line 4, we know that $r.\text{info}$ is a pointer to U . By the test at line 7, either $\text{state} = \text{Committed}$ and $\text{marked}_2 = \text{TRUE}$, or $\text{state} = \text{InProgress}$.

Suppose $\text{state} = \text{InProgress}$. Then, $U.\text{state} = \text{InProgress}$ when I performs line 5. Since U is not the dummy SCX-record, a pointer to U can appear in $r.\text{info}$ only after a successful freezing CAS belonging to U . By Corollary 3.17, $U.\text{state}$ can only be InProgress before the first commit step or abort step belonging to U . Therefore, Definition 3.67 implies that I performs line 5 during the threatening section of S .

Now, suppose $\text{state} = \text{Committed}$ and $\text{marked}_2 = \text{TRUE}$. By Corollary 3.17, $U.\text{state} = \text{Committed}$ at all times after I performs line 5. We consider two sub-cases. If $\text{marked}_1 = \text{TRUE}$, then I will return FINALIZED if it reaches line 13. Since we have assumed that I returns FAIL, this case is impossible. Otherwise, a mark step $mstep$ belonging to some SCX-record W changes $r.\text{marked}$ to TRUE between line 3 and line 6. It remains only to show that $mstep$ occurs during the threatening section of the invocation of SCX that created W . Since $r.\text{marked}$ is initially FALSE , and is never changed from TRUE to FALSE , $mstep$ must be the first mark step belonging to W on r . From the code of HELP, a frozen step belonging to W must precede $mstep$. Therefore, Lemma 3.15 implies that no abort step belonging to W ever occurs. From the code of HELP, $mstep$ must occur after the first freezing CAS belonging to W , and before the first commit step belonging to W . By Definition 3.67, $mstep$ occurs during the threatening section of the invocation of SCX that created W . ■

We now prove that an invocation of SCX or VLX can return FALSE only under certain circumstances.

Definition 3.71 *The **vulnerable interval** of an invocation I of SCX or VLX begins at the earliest starting time of an $\text{LLX}(r)$ linked to I , and ends when I ends.*

Lemma 3.72 *Let I be an invocation of SCX or VLX, and U be an SCX-record created by an invocation S of SCX. If I blames S for a Data-record r , then a successful freezing CAS belonging to U on r occurs during I 's vulnerable interval.*

Proof: Suppose I is an invocation of SCX. Let U_I be the SCX-record created by I . By Definition 3.65, a successful freezing CAS belonging to U on r occurs between when the $\text{LLX}(r)$ linked to I performs line 4, and the first freezing CAS $fcas$ belonging to U_I on r . By Lemma 3.68, $fcas$ occurs during I . Now, suppose I is an invocation of VLX. Then, by Definition 3.63, a successful freezing CAS belonging to U on r occurs between when the $\text{LLX}(r)$ linked to I performs line 4, and when I sees $r.\text{info} \neq r.\text{info}$ at line 9. ■

Observation 3.73 If an invocation I of $\text{SCX}(V, R, \text{fld}, \text{new})$ or $\text{VLX}(V)$ blames an invocation of SCX for a Data-record r , then r is in V .

Proof: Suppose I is an invocation of SCX . Let U be the SCX -record created by I . By Definition 3.65, r is in $U.V$. Since $U.V$ does not change after U is created at line 21 of I , r is in V . Now, suppose I is an invocation of VLX . In this case, the claim is immediate from Definition 3.63. ■

Observation 3.74 If an invocation I of SCX or VLX blames an invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$ for a Data-record r , then r is in V .

Proof: Let U be the SCX -record created by S . By Lemma 3.72, there is a successful freezing CAS belonging to U on r . The claim then follows from Observation 3.69. ■

Lemma 3.75 An invocation I of $\text{SCX}(V, R, \text{fld}, \text{new})$ or $\text{VLX}(V)$ ending at time t can return FALSE only if its vulnerable interval overlaps the threatening section of some other $\text{SCX}(V', R', \text{fld}', \text{new}')$, where some Data-record appears in both V and V' .

Proof: Suppose I returns FALSE. By Lemma 3.64 and Lemma 3.66, I blames an invocation S of $\text{SCX}(V', R', \text{fld}', \text{new}')$, where $I \neq S$, for a Data-record r . Let U be the SCX -record created by S , and U_I be the SCX -record created by I . By Lemma 3.72, a successful freezing CAS f_{cas} belonging to U on r occurs during I 's vulnerable interval. By Lemma 3.68, f_{cas} occurs during the threatening section of S . Therefore, I 's vulnerable interval overlaps the threatening section of S . By Observation 3.73 and Observation 3.74, r is in both V and V' . ■

We now prove bounds on the number of invocations of LLX , SCX and VLX that can blame an invocation of SCX .

Lemma 3.76 Let I be an invocation of $\text{LLX}(r)$ that returns FAIL, and U be the SCX -record created by the invocation of SCX that is blamed by I . A commit step or abort step belonging to U occurs before I returns.

Proof: By Definition 3.61, we know that I reads a pointer to U from $r.info$. By Lemma 3.62, U is not the dummy SCX -record. Thus, U can have state Committed or Aborted only after a commit step or abort step belonging to U has occurred. If I returns at line 11, then it saw $state \in \{\text{Aborted}, \text{Committed}\}$ at line 7, and we are done. So, suppose I returns at line 14 or line 16. Then, before returning, I performs line 12, where it either sees $state \in \{\text{Aborted}, \text{Committed}\}$ or invokes $\text{HELP}(r.info)$. If I sees $state \in \{\text{Aborted}, \text{Committed}\}$, then we are done. So, suppose I invokes $\text{HELP}(r.info)$. From the code of HELP , if I 's invocation of HELP returns FALSE, then I performs an abort step belonging to U during its invocation of HELP . Otherwise, by Lemma 3.49.3, a commit step belonging to U occurs before I 's invocation of HELP returns. ■

Lemma 3.77 Each invocation of SCX can be blamed by at most two invocations of LLX per process.

Proof: Let S be an invocation of SCX . To derive a contradiction, suppose there is some process p that blames S for three failed invocations of LLX : I_1, I_2 and I_3 (which are performed by p in this order). By Definition 3.61, I_1, I_2 and I_3 each read a pointer to U from $r.info$, either at line 4 or at line 9. Since $r.info$ points to U at some point during I_1 , and again at or after the time I_2 performs line 4, we know from Lemma 3.6 that $r.info$ points to U when I_2 performs line 4.

By the same argument, $r.info$ points to U when I_3 performs line 4. Thus, in both I_2 and I_3 , the local variable $r.info$ is a pointer to U .

We first argue that I_2 cannot enter the if-block at line 7. Suppose, to obtain a contradiction, that I_2 enters the if-block. Since we know that $r.info$ points to U when I_2 performs line 4 and when I_3 performs line 4, Lemma 3.6 implies that $r.info$ points to U when I_2 performs line 9. Thus, I_2 will see $r.info = r.info$ at line 9, and will return at line 11, which contradicts our assumption that I_2 returns FAIL at line 16.

Since I_2 does not enter the if-block at line 7, either $state = \text{Committed}$ and $marked_2 = \text{TRUE}$ or $state = \text{InProgress}$ in I_2 . By Lemma 3.76, a commit step or abort step belonging to U occurs prior to the termination of I_1 , which is before the start of I_2 . By Corollary 3.17, $r.info.state$ does not change after it is set to Committed or Aborted by this commit step or abort step, so $state \neq \text{InProgress}$ in I_2 . Therefore, $state = \text{Committed}$ and $marked_2 = \text{TRUE}$ in I_2 , which means that the Data-record r is marked at some point during I_2 . By inspection of the code, once a Data-record is marked, it remains marked forever. Thus, when I_3 performs line 3, it will see $r.marked = \text{TRUE}$, so $marked_1 = \text{TRUE}$ in I_3 . Since an invocation of LLX can return FAIL only if $marked_1 = \text{FALSE}$, I_3 cannot return FAIL, which is a contradiction. ■

Lemma 3.78 *Each invocation of $\text{SCX}(V, R, fld, new)$ can be blamed by at most $|V|$ invocations of SCX or VLX per process.*

Proof: By Observation 3.74, if an invocation of SCX or VLX blames an invocation of $\text{SCX}(V, R, fld, new)$ for r , then r is in V . Thus, it suffices to prove that an invocation S of $\text{SCX}(V, R, fld, new)$ cannot be blamed for any r in V by more than one invocation of SCX or VLX performed by process p .

Let I and I' be invocations of SCX or VLX performed by process p , and U, U' and U_S be the SCX-records created by I, I' and S , respectively. Without loss of generality, let I' occur after I . Suppose, in order to derive a contradiction, that I and I' both blame S for the same Data-record r . Let $t_0 (t'_0)$ be the time when the $\text{LLX}(r)$ linked to $I (I')$ performs line 4, and $t_1 (t'_1)$ be the time when $I (I')$ finishes. By Lemma 3.72, a successful freezing CAS belonging to U_S occurs between t_0 and t_1 , and a successful freezing CAS belonging to U_S occurs between t'_0 and t'_1 . If we can show $t_0 < t_1 < t'_0 < t'_1$, then we shall have demonstrated that there must be two such freezing CASs, which contradicts Lemma 3.8.

Since the $\text{LLX}(r)$ linked to $I (I')$ terminates before $I (I')$, we know $t_0 < t_1 (t'_0 < t'_1)$. By Observation 3.73, r is in the V sequences of invocations I and I' . Hence, Definition 3.1.2 implies that $t_1 \notin [t'_0, t'_1]$, and $t'_1 \notin [t_0, t_1]$. Since I' occurs after I , $t_1 < t'_1$. Therefore, $t_0 < t_1 < t'_0 < t'_1$. Therefore, $t_0 < t_1 < t'_0 < t'_1$. ■

We now define the blame graph and prove a number of its properties.

Definition 3.79 *We define the **blame graph** for an execution to be a directed graph whose nodes are the invocations of LLX, VLX and SCX, with an edge from an invocation I to another invocation I' if and only if I blames I' . (Note that only the nodes corresponding to invocations of SCX can have incoming edges.)*

The next property we prove is that, for each execution, there is a bound on the length of the longest path in the blame graph. As mentioned in Section 3.2, we require the following constraint in order to prove this bound exists.

Constraint 3.80 *If there is a configuration C after which the value of no field of any Data-record changes, then there is a total order \prec on all Data-records created during the execution such that, if Data-record r_1 appears before data Data-record r_2 in the sequence V passed to an invocation of SCX whose linked LLXs begin after C , then $r_1 \prec r_2$.*

Lemma 3.81 *Let U be an SCX-record created by an invocation of SCX whose linked LLXs begin after the configuration C that is specified in Constraint 3.80. Immediately after a successful freezing CAS belonging to U on r , $r.info$ points to U and, for each r' in $U.V$, where $r' \prec r$, $r'.info$ points to U and a successful freezing CAS belonging to U on r' has occurred.*

Proof: Let $fcas$ be a successful freezing CAS belonging to U on r , and let r' be any Data-record in $U.V$ that satisfies $r' \prec r$. By Constraint 3.80, r' must occur before r in the sequence $U.V$. By Lemma 3.12, a successful freezing CAS $fcas'$ belonging to U on r' occurs prior to $fcas$. Thus, Corollary 3.20 implies that $r'.info$ points to U at all times after $fcas'$ and before the first commit step or abort step belonging to U . Similarly, $r.info$ points to U at all times after $fcas$ and before the first commit step or abort step belonging to U . By Lemma 3.19, $fcas$ must precede the first frozen step or abort step belonging to U . From the code of HELP, the first frozen step belonging to U must precede the first commit step belonging to U . Hence, $fcas$ and $fcas'$ both precede the first commit step or abort step belonging to U . Therefore, immediately after $fcas$, the $info$ fields of r and r' both point to U . ■

Lemma 3.82 *Let U_1 , U_2 and U_3 be SCX-records respectively created by invocations S_1 , S_2 and S_3 of SCX whose linked LLXs begin after the configuration C that is specified in Constraint 3.80, and r and r' be Data-records. If S_1 blames S_2 for r , and S_2 blames S_3 for r' , then $r \prec r'$.*

Proof: Since S_1 blames S_2 for r , we know from Definition 3.65 that r is in $U_2.V$. Similarly, since S_2 blames S_3 for r' , we know r' is in $U_3.V$. Furthermore, a successful freezing CAS belonging to U_2 on r occurs, and no successful freezing CAS belonging to U_2 on r' occurs. By Lemma 3.12, r must occur before r' in the sequence $U_2.V$. Thus, Constraint 3.80 implies $r \prec r'$. ■

Lemma 3.83 *There can be only as many successful update CAs as there are invocations of SCX that either return TRUE, or do not terminate.*

Proof: From the code, an update CAS can only occur in an invocation of $\text{HELP}(ptr)$, where ptr points to an SCX-record U . Further, from the code of HELP, there is at least one freezing CAS belonging to U or frozen step belonging to U . Hence, Lemma 3.5 implies that U is not the dummy SCX-record. Thus, U is created by an invocation of SCX at line 21. By Lemma 3.41, only the first update CAS belonging to an SCX-record can succeed. By Lemma 3.50.3, no update CAS belongs to an SCX-record created by an unsuccessful invocation of SCX. ■

We think of processes as accessing such a data structure via a fixed number of special Data-records called *entry points*, each of which has a single mutable pointer to a Data-record. We assume there is always some Data-record reachable by following pointers from an entry point that is not finalized. (This assumption that entry points cannot be finalized is not crucial, but it simplifies the statement of some progress guarantees.)

Definition 3.84 *A Data-record is *initiated* at all times after it first becomes reachable by following Data-record pointers from an entry point.*

Observation 3.85 *The only step in an execution that can cause a Data-record to become initiated is a successful update CAS.*

Lemma 3.86 *Let S_1 and S_2 be invocations of SCX, and let r be a Data-record. If S_1 blames S_2 for r , then r was initiated before the start of S_1 , and before the start of S_2 .*

Proof: Let U_1 and U_2 be the SCX-records created by S_1 and S_2 , respectively. By Definition 3.65, r is in both $U.V$ and $U'.V$. By Observation 3.2.3, there are invocations of $\text{LLX}(r)$ linked to S_1 and S_2 , respectively. By the precondition of LLX , r must be initiated before the $\text{LLX}(r)$ linked to S_1 , and before the $\text{LLX}(r)$ linked to S_2 . Finally, the $\text{LLX}(r)$ linked to S_1 must terminate before S_1 begins, and the $\text{LLX}(r)$ linked to S_2 must terminate before S_2 begins. \blacksquare

Lemma 3.87 *If no SCX is linearized after some time t , then the following hold.*

1. *A finite number N of Data-records are ever initiated in the execution.*
2. *Let σ be the set of invocations of SCX in the execution whose vulnerable intervals start at or before t . The longest path in the blame graph consisting entirely of invocations of LLX, SCX, and VLX that are not in σ has length at most $N + 2$*

Proof: Claim 1 follows immediately from Observation 3.85 and Lemma 3.83.

We now prove claim 2. Suppose, in order to derive a contradiction, that there is a path of length at least $N + 3$ in the blame graph consisting entirely of invocations of LLX, SCX, and VLX that are not in σ . Since only invocations of SCX can be blamed, at least $N + 2$ of the nodes on this path must correspond to invocations of SCX. Let S_1, S_2, \dots, S_{N+2} be invocations of SCX corresponding to any $N + 2$ consecutive nodes on this path, and let U_1, U_2, \dots, U_{N+2} be the SCX-records they created, respectively. For each $i \in \{1, 2, \dots, N + 1\}$, let r_i be the Data-record for which S_i blames S_{i+1} . Since no invocation of SCX is linearized after t , and the vulnerable sections of S_1, S_2, \dots, S_{N+2} all start after t , no invocation of SCX is linearized after the first $\text{LLX}(r)$ linked to any of these invocations of SCX. Therefore, from Lemma 3.82 and the fact that, for each $i \in \{1, 2, \dots, N\}$, S_i blames S_{i+1} for r_i and S_{i+1} blames S_{i+2} for r_{i+1} , we obtain $r_i \prec r_{i+1}$. By Lemma 3.86, before any invocation of SCX in $\{S_1, S_2, \dots, S_{N+2}\}$ begins, r_1, r_2, \dots, r_{N+1} have all been initiated. Therefore, some Data-record r appears twice in $\{r_1, r_2, \dots, r_{N+1}\}$. Since the \prec relation is transitive, we obtain $r \prec r$, which is a contradiction. \blacksquare

We now prove the main progress property for SCX.

Lemma 3.88 *If invocations of SCX complete infinitely often, then invocations of SCX succeed infinitely often.*

Proof: Suppose, to derive a contradiction, that after some time t' , invocations of SCX are performed infinitely often, but no invocation of SCX is successful. Then, since we only linearize successful SCXs, and a subset of the non-terminating SCXs, there is a time $t \geq t'$ after which no invocation of SCX is linearized. Let σ be the set of invocations of SCX in the execution whose vulnerable intervals start at or before t . By Lemma 3.77 and Lemma 3.78, the in-degree of each node in the blame graph is bounded. Since σ is finite, and the in-degree of each node in σ is bounded, only a finite number of invocations of SCX can blame invocations in σ . Now, consider any maximal path π consisting entirely of invocations of LLX, SCX, and VLX that are *not* in σ . By Lemma 3.87.2, π has length at most $N + 3$. Since no invocation of SCX is successful after t , the invocation S of SCX corresponding to the last node on path π must be unsuccessful. By Lemma 3.66, S must blame some other invocation of SCX. Since π is maximal, S must blame an invocation of SCX in σ . Thus, there can be only finitely many of these paths (of bounded length). However, this contradicts our assumption that invocations of SCX occur infinitely often. \blacksquare

Unfortunately, since SCX and VLX cannot be invoked unless a sequence of invocations of LLX (linked to the SCX or VLX) return snapshots, the previous result is not strong enough unless we can guarantee that processes can invoke SCX and/or VLX infinitely often. We first give two definitions that help clarify the progress guarantees for SCX and VLX.

Definition 3.89 An SCX-UPDATE algorithm performs LLXs on a sequence V of Data-records and invokes $\text{SCX}(V, R, \text{fld}, \text{new})$ if all of these LLXs return snapshots. A successful SCX-UPDATE is one in which the SCX returns TRUE.

Definition 3.90 A VLX-QUERY algorithm performs LLXs on a sequence V of Data-records and invokes $\text{VLX}(V)$ if all of these LLXs return snapshots. A successful VLX-QUERY is one in which the VLX returns TRUE.

Theorem 3.91 Our implementation of LLX/SCX/VLX satisfies the following progress properties.

1. If operations (LLX, SCX, VLX) are performed infinitely often, then operations succeed infinitely often.
2. Suppose that (a) there is always some non-finalized Data-record reachable by following pointers from an entry point, (b) for each Data-record r , each process performs finitely many invocations of $\text{LLX}(r)$ that return FINALIZED, and (c) processes perform infinitely many executions of SCX-UPDATE and/or VLX-QUERY algorithms. Then, infinitely many SCX or VLX operations succeed.

Proof: Both claims have similar proofs, by cases.

Proof of Claim 1. Suppose operations are performed infinitely often.

Case I: invocations of SCX are performed infinitely often. In this case, Lemma 3.88 implies that invocations of SCX will succeed infinitely often, and the claim is proved.

Case II: after some time t , no invocation of SCX is performed. In this case, the blame graph contains a finite number of invocations of SCX. By Lemma 3.77 and Lemma 3.78, the in-degree of each node in the blame graph is bounded. By Lemma 3.62 and Lemma 3.64, each unsuccessful invocation of LLX or VLX blames an invocation of SCX. Therefore, only finitely many invocations of LLX and VLX can be unsuccessful. Thus, eventually, every invocation of LLX or VLX succeeds.

Proof of Claim 2. Suppose the antecedent holds, and processes perform infinitely many executions of SCX-UPDATE or VLX-QUERY algorithms.

Case I: invocations of SCX are performed infinitely often. In this case, Lemma 3.88 implies that invocations of SCX will succeed infinitely often, and the claim is proved.

Case II: eventually, no invocation of SCX is performed. Then, as we argued in Case II of the proof of Claim 1, only finitely many invocations of LLX can be unsuccessful. This implies that, after some time t , every invocation of LLX is successful. Furthermore, by antecedent (b) of this claim, for each Data-record r , each process will perform finitely many invocations of $\text{LLX}(r)$ that return FINALIZED. By Lemma 3.87, there are a finite number of Data-records that are ever initiated in the execution. Therefore, each process can perform only finitely many invocations of LLX that return FINALIZED. Consequently, after some time t' , every invocation of LLX by any process will return a snapshot. Any SCX-UPDATE that starts after t' will perform LLXs that return snapshots, and will invoke SCX, violating our assumption in this case. Thus, after t' , processes execute infinitely many VLX-QUERY algorithms, but no SCX-UPDATE algorithms. Any VLX-QUERY algorithm that starts after t' will perform LLXs that return snapshots, and will invoke VLX. Thus, infinitely many invocations of VLX are performed after t' . Since none of these invocations are concurrent with any invocation of SCX, Lemma 3.75 implies that all of these invocations of VLX must succeed.

■

3.5 Additional properties of LLX/SCX/VLX

In this section we prove some additional properties of LLX/SCX/VLX that are intended to simplify the design of certain data structures. At this level, a *configuration* consists of the state of each process, and a collection of Data-records (which have only mutable and immutable fields). A *step* is either a READ, or a linearized invocation of LLX, SCX or VLX.

Definition 3.92 A Data-record r is **in the data structure** in some configuration C if and only if r is reachable by following pointers from an entry point. We say a Data-record r is **removed (from the data structure)** by some step s if and only if r is in the data structure immediately before s , and r is not in the data structure immediately after s . We say a Data-record r is **added (to the data structure)** by some step s if and only if r is not in the data structure immediately before s , and r is in the data structure immediately after s .

Note that a Data-record can be removed from or added to the data structure only by a linearized invocation of SCX.

If the following constraint is satisfied, then the results of this section apply.

Constraint 3.93 For each linearized invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$, R contains precisely the Data-records that are removed from the data structure by S .

Lemma 3.94 If a Data-record r is removed from the data structure for the first time by step s , then no linearized invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$, where fld is a mutable field of r , occurs at or after s . (Hence, r does not change at or after s .)

Proof: The only step that can change r is a linearized invocation of SCX. The invocation S' of $\text{SCX}(V', R', \text{fld}', \text{new}')$ that removes r modifies a mutable field of some Data-record different from r . Thus, fld' is not a field of r . Since this is the only change to the data structure when s occurs, r does not change when s occurs. Suppose, to derive a contradiction, that an invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$, where fld is a mutable field of r , occurs after s . Then, since r is in V , the precondition of SCX implies that an invocation I of $\text{LLX}(r)$ linked to S must occur before S . By Constraint 3.93, r is in R' . Thus, if I occurs after S' , then it returns FINALIZED, which contradicts Definition 3.1. Otherwise, S' occurs between I and S , so S cannot be linearized, which contradicts our assumption. ■

Lemma 3.95 If an invocation I of $\text{LLX}(r)$ returns a value different from FAIL or FINALIZED, then r is in the data structure just before I is linearized.

Proof: By the precondition of LLX, r is initiated and, hence, in the data structure, at some point before I . Suppose, to derive a contradiction, that r is not in the data structure just before I is linearized. Then, some linearized invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$ must remove r before I is linearized. By Constraint 3.93, r is in R . However, this implies that I must return FINALIZED, which is a contradiction. ■

Lemma 3.96 If S is a linearized invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$, where new is a Data-record, then new is in the data structure just after S .

Proof: Note that fld is a mutable field of a Data-record r in V . We first show that r is in the data structure at some point before S . By the precondition of SCX, before S , there is an $\text{LLX}(r)$ linked to S . By the precondition of LLX, r

must be initiated when this linked LLX occurs. Thus, Definition 3.84 and Definition 3.92 imply that r is in the data structure at some point before S . Suppose, to derive a contradiction, that r is not in the data structure just after S . Then, r must either be removed by S , or by some previous step. However, this directly contradicts Lemma 3.94. ■

Let C_1 and C_2 be configurations in the execution. We use $C_1 < C_2$ to mean that C_1 precedes C_2 in the execution. We say $C_1 \leq C_2$ precisely when $C_1 = C_2$ or $C_1 < C_2$. We denote by $[C_1, C_2]$ the set of configurations $\{C \mid C_1 \leq C \leq C_2\}$.

Lemma 3.97 *Let r_1, r_2, \dots, r_l be a sequence of Data-records, where r_1 is an entry point, and C_1, C_2, \dots, C_{l-1} be a sequence of configurations satisfying $C_1 < C_2 < \dots < C_{l-1}$. If, for each $i \in \{1, 2, \dots, l-1\}$, a field of r_i points to r_{i+1} in configuration C_i , then r_{i+1} is in the data structure in some configuration in $[C_1, C_i]$. Additionally, if a mutable field f of r_l contains a value v in some configuration C_l after C_{l-1} then, in some configuration in $[C_1, C_l]$, r_l is in the data structure and f contains v .*

Proof: We prove the first part of this result by induction on i .

Since each entry point is always in the data structure, and r_1 points to r_2 in configuration C_1 , r_2 is in the data structure in C_1 . Thus, the claim holds for $i = 1$.

Suppose the claim holds for i , $1 \leq i \leq l-2$. We prove it holds for $i+1$. If r_i is in the data structure when it points to r_{i+1} in C_i , then r_{i+1} is in the data structure in C_i , and we are done. Suppose r_i is not in the data structure in C_i . By the inductive hypothesis, r_i is in the data structure in some configuration in $[C_1, C_{i-1}]$. Let $s, C_1 < s < C_i$, be the first step such that r_i is removed from the data structure by s . In the configuration C just before s , r_i is in the data structure. By Lemma 3.94, r_i does not change at or after s . Thus, r_i does not change after C . Since C occurs before C_i , and r_i points to r_{i+1} in C_i , r_i must point to r_{i+1} in C . Therefore, in C (which satisfies $C_1 \leq C < C_i$), r_i is in the data structure and points to r_{i+1} .

The second part of the proof is quite similar to the inductive step we just finished. Suppose f contains v in C_l . If r_l is in the data structure in C_l , then we are done. Suppose r_l is not in the data structure in C_l . We have shown above that r_l is in the data structure in some configuration in $[C_1, C_l]$. Let $s', C_1 < s' < C_l$, be the first step such that r_l is removed from the data structure by s' . In the configuration C' just before s' , r_l is in the data structure. By Lemma 3.94, r_l does not change at or after s' . Thus, r_l does not change after C' . Since C' occurs before C_l , and f contains v in C_l , f must contain v in C' . Therefore, in C' (which satisfies $C_1 \leq C' < C_l$), r_l is in the data structure and f contains v . ■

3.6 Modifications to enable memory reclamation

Each invocation of SCX creates a new SCX-record, which must eventually be reclaimed. In managed languages such as Java and C#, automatic garbage collection can be used to reclaim SCX-records. In unmanaged languages, more specialized techniques must be used. We start by describing a modification to the SCX algorithm that enables garbage collection in managed languages, and then describe how memory can be reclaimed in unmanaged languages.

3.6.1 Garbage collection

At a high level, garbage collection reclaims records only once all processes can no longer access them. Thus, as long as any process can reach a record by following pointers, the record can not be reclaimed. The vanilla SCX algorithm presented above is poorly suited for automatic garbage collection, because it allows garbage to remain reachable for an arbitrarily long time.

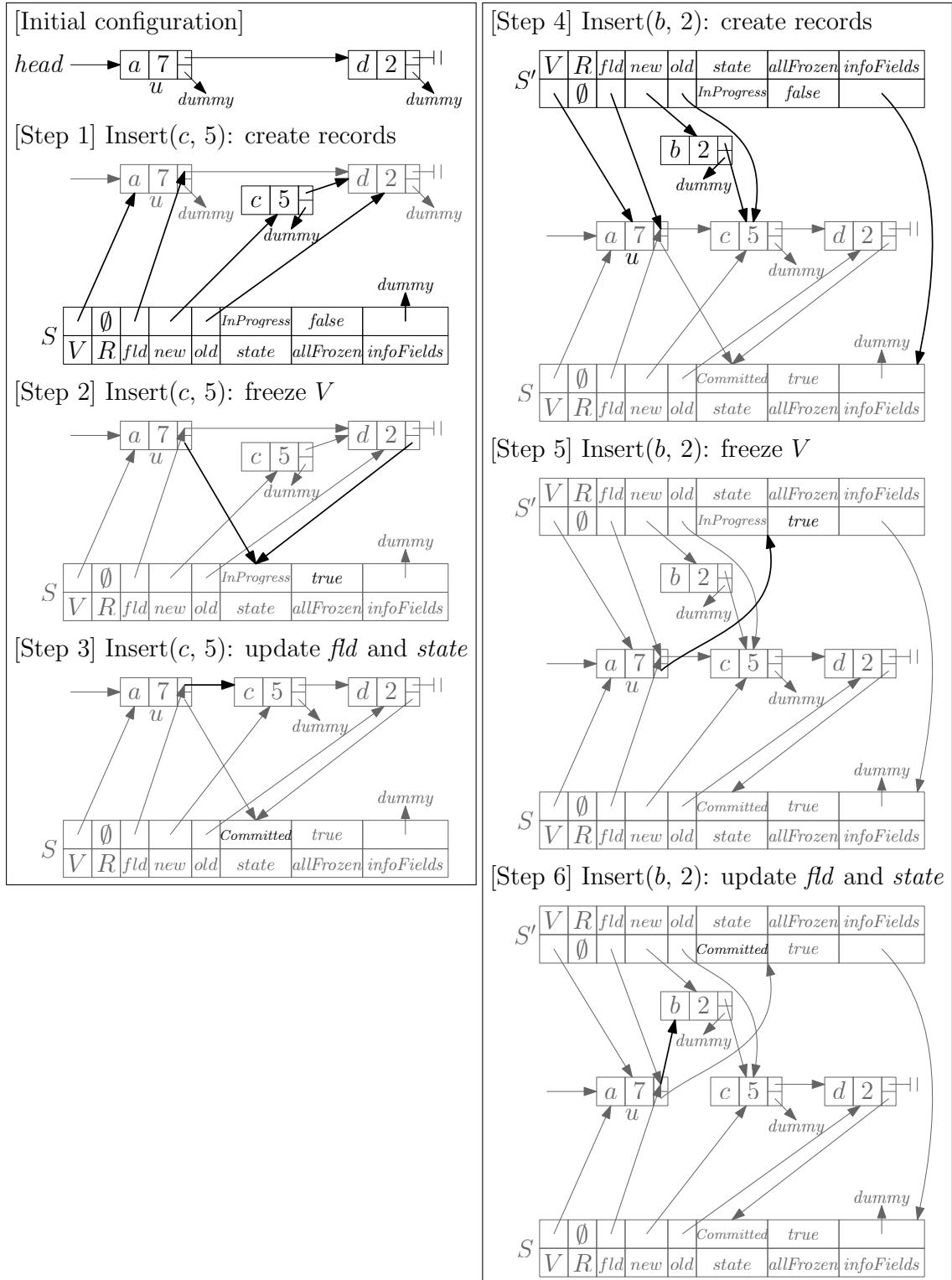


Figure 3.9: Example execution of a multiset in which SCX operations create reachable garbage.

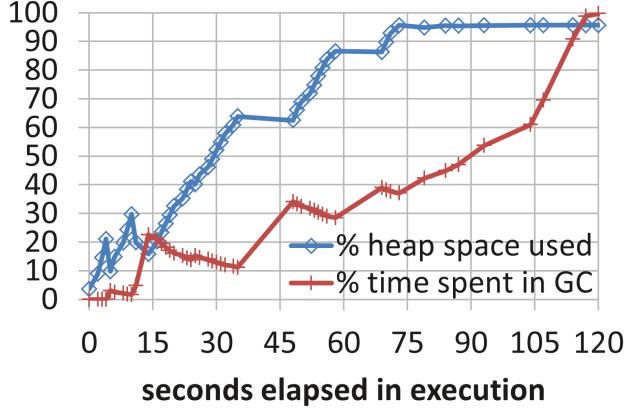


Figure 3.10: Experiment showing total memory usage and time spent in garbage collection for a Java microbenchmark on a balanced binary search tree implemented from LLX and SCX.

To show this experimentally, we produced a Java implementation of a lock-free balanced binary search tree (discussed in Chapter 5), and ran a microbenchmark wherein 8 processes perform 50% INSERT and 50% DELETE operations on keys drawn uniformly randomly from $[0, 10^6]$ for 120 seconds. During this experiment, we used the NetBeans 7.2 profiler to collect statistics on the total memory usage of the Java virtual machine (JVM), and on the fraction of the execution time that was spent performing garbage collection. We ran the experiment on an Intel i7-2600k with 4 cores and 2 hyperthreads per core and 16GB of RAM, running Windows 7, with the Java 1.7.0_07 64-bit server VM. The Java heap was automatically sized by the JVM. (I.e., we did not specify explicit minimum or maximum sizes for the heap.) Measurements showed that it was approximately 3GB.

The results of the experiment appear in Figure 3.10. Garbage collection performed reasonably well until approximately 15 seconds into the experiment. As the execution continued, more and more of the total heap space was occupied, and the fraction of execution time spent performing garbage collection increased (in response to the increased memory pressure). Eventually, nearly 100% of the execution time was spent in garbage collection, and insertions and deletions become virtually impossible, because garbage collection could not free enough memory. Note that gaps in the measurements (e.g., between 35 and 48 seconds) were caused by long garbage collection delays.

The problem We explain the poor performance of garbage collection using an example, which is illustrated in Figure 3.9. There, newly added or changed elements are drawn in black, and elements unchanged from the previous step are drawn in gray. Nodes are drawn with a key, number of copies, and pointers to the next node (the upper pointer) and to an SCX-record (the lower pointer). The dummy SCX-record is denoted simply with *dummy*.

In this example, $\text{INSERT}(c, 5)$ performs an SCX that creates an SCX-record S (Step 1), freezes the node u containing key a (Step 2), then updates the next pointer of u to perform the insertion and sets $S.\text{state}$ to *Committed* (Step 3). Note that u will continue to point to S until it is next frozen by an SCX. Next, $\text{INSERT}(b, 2)$ performs an SCX that creates an SCX-record S' (Step 4), freezes u (Step 5), then updates $u.\text{next}$ and $S'.\text{state}$ (Step 6). Since u pointed to S when $\text{INSERT}(b, 2)$ performed its SCX, S' has a pointer to S in its *infoFields*. (Recall that *infoFields* contains the old values used for freezing CAS steps.) Thus, S continues to be reachable (through S'). Note that u will continue to point to S' until it is next frozen by an SCX. Furthermore, the next time u is frozen, the SCX that freezes it will have a pointer to S' in its SCX-record (in the *infoFields* field), so S' will continue to be reachable (and, hence, so will S). In

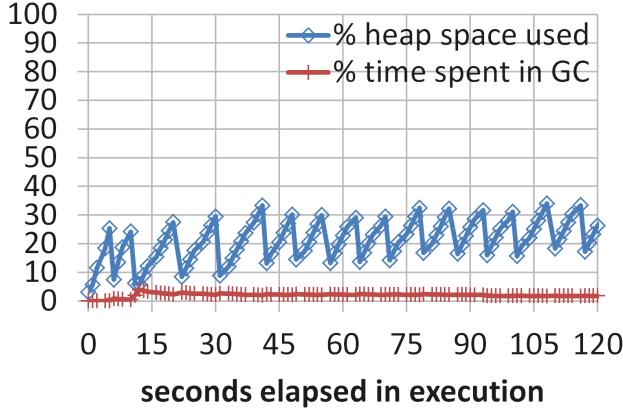


Figure 3.11: Experiment showing the result of modifying the HELP procedure to break chains of pointers once they are no longer necessary.

this way, long chains of pointers between SCX-records can prevent them from being reclaimed. Similarly, as long as an SCX-record remains reachable, so do the nodes that it points to (and the SCX-records they each point to, etc.).

A solution We solve this problem by modifying the HELP procedure so that it breaks these chains of pointers once they are no longer necessary. Specifically, after a process finishes helping an SCX-record S and sets $S.state$ to *Committed* or *Aborted*, it then replaces all pointers in $S.V$, $S.R$, $S.fld$, $S.new$, $S.old$ and $S.infoFields$ with NIL. Note that this change makes it possible for helpers to see NIL in some or all of these fields. However, these fields can contain NIL only after *state* is *Committed* or *Aborted* (which means helping is no longer necessary). Therefore, each invocation of $\text{HELP}(ptr)$, where ptr is a pointer to an SCX-record S , begins by making a copy of S , and then checking whether $S.state$ contains *Committed* or *Aborted*. If so, the invocation of HELP simply returns, without helping. Otherwise, it proceeds as usual, referring to the copy of S , instead of reading the fields of S directly. Figure 3.11 shows the result of this fix on the experiment described above.

3.6.2 Unmanaged languages

In unmanaged languages, where automatic garbage collection is unavailable, memory must be reclaimed manually. In order to safely reclaim an SCX-record S , a process must know that it can no longer be accessed by any other process. This requires knowing S is not reachable (by following pointers) from any node in the data structure, or from any pointer stored in the private memory of a process. Most state of the art lock-free memory reclamation schemes (e.g., [8, 28, 97]) require an algorithm to invoke a *retire* procedure when a Data-record is no longer reachable starting from any node in the data structure. The memory reclamation scheme then determines when Data-records are no longer reachable starting from any pointer in the private memory of a process, and subsequently frees them.

3.6.3 Reachability from nodes in the data structure

We modify LLX and SCX to use a (very) limited form of reference counting. This allows us to determine whether S is reachable from a node in the data structure. To do this, we augment the *state* field of each SCX-record with a natural number k (so the *state* field now contains $\langle s, k \rangle$, where $s \in \{\text{InProgress}, \text{Committed}, \text{Aborted}\}$). LLX ignores the new

k component of the $state$ field and continues to use only the s component. We modify HELP as follows. Just before the loop at line 24, we initialize a new variable k to zero. At the end of each iteration of this loop, we increment k . Instead of setting $S.state$ to *Aborted* at line 34, we use CAS to change $S.state$ from *InProgress* to $\langle \text{Aborted}, k \rangle$. We call this an *abort CAS*, and say it *belongs to S* if it attempts to modify $S.state$. Similarly, instead of setting $S.state$ to *Committed* at line 41, we use CAS to change $S.state$ from *InProgress* to $\langle \text{Committed}, k \rangle$. We call this a *commit CAS*, and say it *belongs to S* if it attempts to modify $S.state$. Every time a process performs a successful freezing CAS belonging to S (at line 26) that changes the $info$ field of a Data-record from old to S , the process now also performs an atomic fetch and decrement (just after the freezing CAS) to change $old.state$ from $\langle s, k \rangle$ to $\langle s, k - 1 \rangle$ (where s is *Committed* or *Aborted*). We claim that, if $k - 1 = 0$, then old is no longer reachable from any node in the data structure.

Correctness More specifically, we claim the following. If any process performs an abort CAS (resp., commit CAS) belonging to S , then the first abort CAS (resp., commit CAS) belonging to S will succeed, and it will change $S.state$ to $\langle \text{Aborted}, k \rangle$ (resp., $\langle \text{Committed}, k \rangle$), where k is the number of successful freezing CAS steps that belong to S . (Observe that k is then equal to the number of pointers to S that were stored in Data-records by freezing CAS steps.) Furthermore, $S.state$ will not change after this successful abort CAS (resp., commit CAS).

In the following, we refer to the original LLX and SCX algorithm presented in Section 3.3 as the *original algorithm*, and we refer to the version that performs the reference counting described above as the *modified algorithm*.

In the original algorithm, $state$ was changed in two places: at line 41, where it was changed to *Committed* by a write, and at line 34, where it was changed to *Aborted* by a write. Furthermore, $state$ could change only from *InProgress* to *Aborted* or *Committed*. Therefore, whenever line 34 was executed, $state$ could contain *InProgress* or *Aborted*, but not *Committed*. Similarly, whenever line 41 was executed, $state$ could contain *InProgress* or *Committed*, but not *Aborted*. This implies that, if a process helping S executed line 34 (resp., line 41), then no process helping S could execute line 41 (resp., line 34). It follows that, in the modified algorithm, if a process helping S executes an abort CAS at line 34 (resp., commit CAS at line 41), then no process helping S executes a commit CAS (resp., abort CAS).

Suppose some process performs an abort CAS belonging to S . Then, by the argument above, there is no commit CAS belonging to S . Moreover, $state$ must contain *InProgress* or $\langle \text{Aborted}, k \rangle$ (for some k) whenever an abort CAS is performed. The first time a process performs an abort CAS belonging to S , $S.state$ must contain *InProgress*, so the abort CAS will succeed and change $S.state$ to $\langle \text{Aborted}, k \rangle$. After this abort CAS, $S.state$ will never again contain *InProgress*, so no subsequent abort CAS belonging to S can succeed. Thus, $S.state$ will not change after the successful abort CAS. Finally, Lemma 3.18 implies that there is a successful freezing CAS belonging to S on each of the first k Data-records in V , and there are no other successful freezing CAS steps belonging to S . Consequently, k is the number of successful freezing CAS steps that belong to S .

Now, suppose some process performs a commit CAS belonging to S . Then, by the argument above, there is no abort CAS belonging to S . Moreover, $state$ must contain *InProgress* or $\langle \text{Committed}, k \rangle$ (for some k) whenever a commit CAS is performed. The first time a process performs a commit CAS belonging to S , $S.state$ must contain *InProgress*, so the commit CAS will succeed and change $S.state$ to $\langle \text{Committed}, k \rangle$. After this commit CAS, $S.state$ will never again contain *InProgress*, so no subsequent commit CAS belonging to S can succeed. Thus, $S.state$ will not change after the successful commit CAS. Finally, we prove that k is the number of successful freezing CAS steps that belong to S . Any process that performs a commit CAS belonging to S at line 41 must also perform a frozen step belonging to S at line 37. Therefore, Lemma 3.11 and Lemma 3.8 imply that there is exactly one successful freezing CAS belonging

to S for each Data-record in V . In other words, there are $|V|$ successful freezing CAS steps belonging to S . Since a commit CAS is executed only if a process exits the loop at line 24 after iterating through each Data-record in V , every commit CAS attempts to change *state* to $\langle \text{Committed}, |V| \rangle$. Thus, $k = |V|$.

3.6.4 Reachability from private pointers

Suppose all invocations of LLX and SCX are performed by *operations*, and processes do not keep any pointers in private memory between operations. (That is, each time a process finishes an operation, it “forgets” all of the pointers in its private memory.) Then, we can use a distributed epoch-based reclamation scheme called DEBRA (described in Chapter 11) to reclaim SCX-records once they are no longer reachable from any pointer in the private memory of a process. Using DEBRA entails invoking a pair of procedures *leaveQstate()* and *enterQstate()* at the beginning and end of each operation, respectively, and invoking another procedure *retire(S)* whenever an SCX-record S is no longer reachable from any node in the tree (but may be reachable from the private memory of some process). In the absence of process failures, every SCX-record is eventually reclaimed by DEBRA. However, DEBRA is not fault tolerant. For lock-free algorithms that satisfy some additional properties, a fault tolerant version called DEBRA+ can be used.

For lock-free algorithms that cannot use DEBRA, it may be possible to use other lock-free reclamation schemes such as Hazard Pointers [97], Beware and Cleanup [57] or ThreadScan [8]. The details of the algorithm implemented using LLX and SCX can affect whether it is possible to use each of these reclamation schemes (see Chapter 11).

Chapter 4

Multiset implemented with LLX/SCX

4.1 Implementation

We now give a detailed description of the implementation of a multiset using LLX and SCX. We assume that keys stored in the multiset are drawn from a totally ordered set and $-\infty < k < \infty$ for every key k in the multiset. As described in Section 3.2, we use a singly-linked list of nodes, sorted by key. To avoid special cases, it always has a sentinel node, $head$, with key $-\infty$ at its beginning and a sentinel node, $tail$, with key ∞ at its end. The definition of Node, the Data-record used to represent a node, and the pseudocode are presented in Figure 4.1.

$\text{SEARCH}(key)$ traverses the list, starting from $head$, by reading $next$ pointers until reaching the first node r whose key is at least key . This node and the preceding node p are returned. $\text{GET}(key)$ performs $\text{SEARCH}(key)$, outputs r 's count if r 's key matches key , and outputs 0, otherwise.

An invocation I of $\text{INSERT}(key, count)$ starts by calling $\text{SEARCH}(key)$. Using the nodes p and r that are returned, it updates the data structure. It decides whether key is already in the multiset (by checking whether $r.key = key$) and, if so, it invokes LLX(r) followed by an SCX linked to r to increase $r.count$ by $count$, as depicted in Figure 3.5(b). Otherwise, I performs the update depicted in Figure 3.5(a): It invokes LLX(p), checks that p still points to r , creates a node, new , and invokes an SCX linked to p to insert new between p and r . If p no longer points to r , the LLX returns FAIL or FINALIZED, or the SCX returns FALSE, then I restarts.

An invocation I of $\text{DELETE}(key, count)$ also begins by calling $\text{SEARCH}(key)$. It invokes LLX on the nodes p and r and then checks that p still points to r . If r does not contain at least $count$ copies of key , then I returns FALSE. If r contains exactly $count$ copies, then I performs the update depicted in Figure 3.5(c) to remove node r from the list. To do so, it invokes LLX on the node, $rnext$, that $r.next$ points to, makes a copy $rnext'$ of $rnext$, and invokes an SCX linked to p, r and $rnext$ to change $p.next$ to point to $rnext'$. This SCX also finalizes the nodes r and $rnext$, which are thereby removed from the data structure. The node $rnext$ is replaced by a copy to avoid the ABA problem in $p.next$. If r contains more than $count$ copies, then I replaces r by a new copy r' with an appropriately reduced count using an SCX linked to p and r , as shown in Figure 3.5(d). This SCX finalizes r . If an LLX returns FAIL or FINALIZED, or the SCX returns FALSE then I restarts.

4.2 Correctness and progress

This section gives a detailed proof of correctness for the multiset implementation. We start with a high-level sketch.

4.2.1 Proof sketch

The proof begins by showing that this multiset implementation satisfies some basic invariants. Notably, (1) $head$ always points to a node, (2) if a node has key ∞ then its $next$ pointer is nil , and (3) if a node's key is not ∞ then its $next$ pointer points to a node with a strictly larger key. These invariants imply that the data structure is always a sorted list.

Next, we prove that the Data-records removed from the data structure by a linearized invocation of $\text{SCX}(V, R, fld, new)$ are exactly the Data-records in R . This allows us to apply the proposition in Section 3.3.3 to prove that there is a time during each SEARCH when the nodes r and p that it returns are both in the list and $p.next = r$.

Each GET and each DELETE that returns FALSE is linearized at the linearization point of the SEARCH it performs. Every other INSERT or DELETE is linearized at its successful SCX. Linearizability of all operations then follows from the following invariant: At every time t , the multiset of keys in the data structure is equal to the multiset of keys that

```

type Node
  ▷ Fields from sequential data structure
  key      ▷ key (immutable)
  count    ▷ occurrences of key (mutable)
  next     ▷ next pointer (mutable)
  ▷ Fields defined by LLX/SCX algorithm
  info     ▷ a pointer to an SCX-record
  marked   ▷ a Boolean value

shared Node tail := new Node( $\infty$ , 0, NIL)
shared Node head := new Node( $-\infty$ , 0, tail)


---


1  GET(key)
2     $\langle r, - \rangle$  := SEARCH(key)
3    if key = r.key then return r.count
4    else return 0


---


5  SEARCH(key)
6    ▷ Postcondition: p and r point to Nodes with p.key < key  $\leq$  r.key.
7    p := head
8    r := p.next
9    while key > r.key do
10       p := r
11       r := r.next
12    return  $\langle r, p \rangle$ 


---


13 INSERT(key, count)      ▷ Precondition: count > 0
14   while TRUE do
15      $\langle r, p \rangle$  := SEARCH(key)
16     if key = r.key then
17       localr := LLX(r)
18       if localr  $\notin$  {FAIL, FINALIZED} then
19         if SCX( $\langle r \rangle, \langle \rangle, \&r.count, localr.count + count$ ) then return
20     else
21       localp := LLX(p)
22       if localp  $\notin$  {FAIL, FINALIZED} and r = localp.next then
23         if SCX( $\langle p \rangle, \langle \rangle, \&p.next, \text{new Node}(key, count, r)$ ) then return


---


24 DELETE(key, count)      ▷ Precondition: count > 0
25   while TRUE do
26      $\langle r, p \rangle$  := SEARCH(key)
27     localp := LLX(p)
28     localr := LLX(r)
29     if localp, localr  $\notin$  {FAIL, FINALIZED} and r = localp.next then
30       if key  $\neq$  r.key or localr.count < count then return FALSE
31       else if localr.count > count then
32         if SCX( $\langle p \rangle, \langle r \rangle, \&p.next, \text{new Node}(r.key, localr.count - count, localr.next)$ ) then return TRUE
33       else ▷ assert: localr.count = count
34       if LLX(localr.next)  $\notin$  {FAIL, FINALIZED} then
35         if SCX( $\langle p, r, localr.next \rangle, \langle r, localr.next \rangle, \&p.next, \text{new copy of } localr.next$ ) then return TRUE

```

Figure 4.1: Pseudocode for a multiset, implemented with a singly linked list.

would result from the atomic execution of the sequence of operations linearized up to time t . Next, we prove that all operations (INSERT, DELETE and SEARCH) return the correct values. Finally, we prove progress by showing the number of invocations of LLX that return FINALIZED is bounded (a requirement for using LLX and SCX), and then using the progress properties of LLX and SCX.

4.2.2 Complete proof

In the following, we define the **response** of a SEARCH to be a step at which a value is returned. Note that we specify Lemma 4.1.3, instead of directly proving the considerably simpler statement in Constraint 3.93, so that we can reuse the intermediate results when proving linearizability.

Lemma 4.1 *The multiset algorithm satisfies the following properties.*

1. *Every invocation of LLX or SCX has valid arguments, and satisfies its preconditions.*
2. *Every invocation of SEARCH satisfies its postconditions.*
3. *Let S be an invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$ performed by an invocation I of INSERT or DELETE, and p, r and r_{next} refer to the local variables of I . If I performs S at line 19, then no Data-record is added or removed by S , and $R = \emptyset$. If I performs S at line 23, then only new is added by S , no Data-record is removed by S , and $R = \emptyset$. If I performs S at line 32, then only new is added by S , only r is removed by S , and $R = \{r\}$. If I performs S at line 35, then only new is added by S , only r and r_{next} are removed by S , and $R = \{r, r_{\text{next}}\}$.*
4. *The head entry point always points to a Node, the next pointer of each Node with $\text{key} \neq \infty$ points to some Node with a strictly larger key, and the next pointer of each Node with $\text{key} = \infty$ is NIL.*

Proof: We prove these claims by induction on the sequence of steps taken in the execution. The only steps that can affect these claims are invocations of LLX and SCX, and responses of SEARCHes. **Base case.** Clearly, Claim 1, Claim 2 and Claim 3 hold before any such step occurs. Before the first SCX, the data structure is in its initial configuration. Thus, Claim 4 holds before any step occurs. **Inductive step.** Suppose these claims hold before some step s . We prove they hold after s .

Proof of Claim 1. The only steps that can affect this claim are invocations of LLX and SCX.

Suppose s is an invocation of LLX. By inductive Claim 3 and Observation 4.3, Constraint 3.93 is satisfied at all times before s occurs. The only places in the code where s can occur are at lines 17, 21, 27, 28 and 34. Suppose s occurs at line 17, 21, 27 or 28. Then, by inductive Claim 2, argument to s is non-NIL. We can apply Lemma 3.97 to show that the argument to s is in the data structure and, hence, initiated, at some point during the last SEARCH before s . Now, suppose s occurs at line 34 (so $\text{localr}.next$ is the argument to s). Then, $\text{key} = r.\text{key}$ when line 30 is performed so, by the precondition of DELETE, $r.\text{key} \neq \infty$. By inductive Claim 4, $r.next \neq \text{NIL}$ when LLX(r) is performed at line 28, so $\text{localr}.next \neq \text{NIL}$. We can apply Lemma 3.97 to show that $\text{localr}.next$ is in the data structure and, hence, initiated, at some point between the start of the last SEARCH before s and the last LLX(r) before s (which reads $\text{localr}.next$ from $r.next$).

Suppose s is a step that performs an invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$. Then, the only places in the code where s can occur are at lines 19, 23, 32 and 35. It is a trivial exercise to inspect the code of INSERT and DELETE, and argue that the process that performs s has done an LLX(r) linked to S for each $r \in V$, that $R \subseteq V$, and that fld points

to a mutable field of a Data-record in V . It remains to prove that Precondition (2) and Precondition (3) of SCX are satisfied. Let I be the invocation of $\text{LLX}(r)$ linked to S . Suppose s occurs at line 19. The only step that can affect the claim is a linearized invocation s of $\text{SCX}(V, R, \text{fld}, \text{new})$ performed at line 19. From the code, fld is $r.\text{count}$. Since s is linearized, no invocation of $\text{SCX}(V'', R'', \text{fld}'', \text{new}'')$ with $r \in V''$ is linearized between I and s . Thus, $r.\text{count}$ does not change between when I and s are linearized. From the code, new is count plus the value read from $r.\text{count}$ by I . Therefore, new is strictly larger than $r.\text{count}$ was when I was linearized, which implies that new is strictly larger than $r.\text{count}$ when s is linearized. This immediately implies Precondition (2) and Precondition (3) of SCX. Now, suppose s occurs at line 23, 32 or 35. Then, new is a pointer to a Node that was created after I . Thus, no invocation of $\text{SCX}(V', R', \text{fld}, \text{new})$ can even begin before I . We now prove that new is not the initial value of the field pointed to by fld . From the code, fld is $p.\text{next}$. If $p.\text{next}$ is initially NIL, then we are done. Otherwise, $p.\text{next}$ initially points to some Node r' . Clearly, r' must be created before p . Hence, r' must be created before the invocation of SEARCH followed a pointer to p . Since new is a pointer to a Node that is created after this invocation of SEARCH, $\text{new} \neq r'$.

Proof of Claim 2. To affect this claim, s must be the response of an invocation of $\text{SEARCH}(\text{key})$. We prove a loop invariant that states r is a Node, and either p is a Node and $p.\text{key} < \text{key}$ or $p = \text{head}$. Before the loop, $p = \text{head}$ and $r = \text{head}.\text{next}$. By inductive Claim 4 $\text{head}.\text{next}$ is always a Node, so the claim holds before the loop. Suppose the claim holds at the beginning of an iteration. Let r and p be the respective values of local variables r and p at the beginning of the iteration, and r' and p' be their values at the end of the iteration. From the code, $p' = r$ and r' is the value read from $r.\text{next}$ at line 11. By the inductive hypothesis, p' is a Node. Since the loop did not exit before this iteration, $\text{key} > p'.\text{key}$. Further, since $\text{SEARCH}(\text{key})$ is invoked only when $\text{key} < \infty$ (by inspection of the code and preconditions), $p'.\text{key} < \infty$. By inductive Claim 4, $p'.\text{next} = r.\text{next}$ always points to a Node, so r' is a Node, and the inductive claim holds at the end of the iteration. Finally, the exit condition of the loop implies $\text{key} \leq r'.\text{key}$, so SEARCH satisfies its postcondition.

Proof of Claim 3. Since a Data-record can be removed from the data structure only by a change to a mutable field of some other Data-record, this claim can be affected only by linearized invocations of SCX. Suppose s is a linearized invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$. Then, s can occur only at line 19, 23, 32 or 35. Let I be the invocation of INSERT or DELETE in which s occurs. We proceed by cases.

Suppose s occurs at line 19. Then, fld is a pointer to $r.\text{count}$. Thus, s changes a count field, not a next pointer. Since this is the only change that is made by s , no Data-record is removed by s , and no Data-record is added by s . Since $R = \emptyset$, the claim holds.

Suppose s occurs at line 23. Then, fld is a pointer to $p.\text{next}$, and new is a pointer to a new Node. Before performing s , I performs an invocation L_1 of $\text{LLX}(p)$, which returns a value different from FAIL, or FINALIZED, at line 27. Just after performing L_1 , I sees that $\text{local } p.\text{next} = r$. Note that L_1 is linked to s . Since s is linearized, and $p \in V$, $p.\text{next}$ does not change in between when L_1 and s are linearized. Therefore, s changes $p.\text{next}$ from r to point to a new Node whose next pointer points to r . Since s is linearized, Lemma 3.94 implies that p must be in the data structure just before s (and when its change occurs). Since this is the only change that is made by s , no Data-record is removed by s , and new points to the only Data-record that is added by s . Since $R = \emptyset$, the claim holds.

Suppose s occurs at line 32 or line 35. Then, fld is a pointer to $p.\text{next}$, and new is a pointer to a new Node. Before performing s , I performs invocations L_1 and L_2 of $\text{LLX}(p)$ and $\text{LLX}(r)$, respectively, which each return a value different from FAIL, or FINALIZED. Note that L_1 and L_2 are linked to s . Just after performing L_1 , I sees that $\text{local } p.\text{next} = r$. Since s is linearized, and $p \in V$, $p.\text{next}$ does not change in between when L_1 and s are linearized. Similarly, since $r \in V$, $r.\text{next}$ does not change between when L_1 and s are linearized. Before s , I sees $\text{key} = r.\text{key}$ at

line 30. By the precondition of DELETE, $r.key \neq \infty$. Thus, inductive Claim 4 (and the fact that keys do not change) implies that $r.next$ points to some Node $rnext = localr.next$ at all times between when L_2 and s are linearized. We consider two sub-cases.

Case I: s occurs at line 32. Therefore, s changes $p.next$ from r to point to a new Node whose $next$ pointer points to $rnext$ and, when this change occurs, $r.next$ points to $rnext$. Since s is linearized, Lemma 3.94 implies that p must be in the data structure just before s (and when its change occurs). Since this is the only change that is made by s , r points to the only Data-record that is removed by s , and new points to the only Data-record that is added by s . Since $R = \{r\}$, the claim holds.

Case II: s occurs at line 35. Since $rnext \in V$, $rnext.next$ does not change between when the LLX at line 34 and s are linearized. Thus, $rnext.next$ contains the same value v throughout this time. Therefore, s changes $p.next$ from r to point to a new Node whose $next$ pointer contains v and, when this change occurs, $p.next$ points to r , $r.next$ points to $rnext$, and $rnext.next$ contains v . Since s is linearized, Lemma 3.94 implies that p must be in the data structure just before s (and when its change occurs). Since this is the only change that is made by s , r and $rnext$ point to the only Data-records that are removed by s , and new points to the only Data-record that is added by s . Since $R = \{r, rnext\}$, the claim holds.

Proof of Claim 4. This claim can be affected only by a linearized invocation of SCX that changes a $next$ pointer. Suppose s is a linearized invocation of $SCX(V, R, fld, new)$. Then, s can occur only at line 23, 32 or 35. We argued in the proof of Claim 3 that, in each of these cases, s changes $p.next$ from r to point to a new Node, and that this is the only change that it makes. Let I be the invocation of INSERT or DELETE in which s occurs.

Suppose s occurs at line 32. We argued in the proof of Claim 3 that, at all times between when the LLX(r) at line 28 and s are linearized, $p.next$ points to r and $r.next$ points to some Node $rnext$. Therefore, $new.key = r.key$ and $new.next$ points to $rnext$. We show r is a Node (and not the *head* entry point), and $r.key \neq \infty$. Since $r.next$ points to a Node $rnext \neq NIL$, $r \neq NIL$ and $r.key \neq \infty$ (by the inductive hypothesis). Similarly, since $p.next$ points to r , either $r = NIL$ or r is a Node, so we are done. Since $r.next$ points to $rnext$ just before s is linearized, setting $new.next$ to point to $rnext$ does not violate the inductive hypothesis. Since $p.next$ points to r , the inductive hypothesis implies that either p is the *head* entry point or $p.key < r.key$. Clearly, setting $p.next$ to point to new does not violate the inductive hypothesis in either case.

Suppose s occurs at line 23. Then, $new.key = key$ and $new.next$ points to r . Before s , I invokes $SEARCH(key)$ at line 15, and then sees $key \neq r.key$ at line 16. By inductive Claim 2, this invocation of $SEARCH$ satisfies its post-conditions, which implies that r points to a Node which satisfies $key < r.key$. Since $SEARCH(key)$ is invoked only when $key < \infty$ (by inspection of the code and preconditions), $key < \infty$. Thus, setting $new.next$ to point to r does not violate the inductive hypothesis. The post conditions of $SEARCH$ also imply that either p is a Node and $p.key < key$ or $p = head$. Therefore, setting $p.next$ to point to new does not violate the inductive hypothesis.

Suppose s occurs at line 35. We argued in the proof of Claim 3 that, at all times between when the LLX(r) at line 28 and s are linearized, $p.next$ points to r , $r.next$ points to some Node $rnext$ (pointed to by $localr.next$) and $rnext.next$ points to some Node $rnext'$. Thus, $new.key = rnext.key$ and $new.next$ points to $rnext'$. Since $rnext.next$ points to a Node $rnext'$, $rnext.key < \infty$ (by the inductive hypothesis). Therefore, since $rnext.next$ points to $rnext'$ just before s , setting $new.next$ to point to $rnext'$ does not violate the inductive hypothesis. By the inductive hypothesis, either p is the *head* entry point, or $p.key < r.key < rnext.key = new.key < \infty$. Clearly, setting $p.next$ to point to new does not violate the inductive hypothesis in either case. ■

Corollary 4.2 *The head entry point always points to a sorted list with strictly increasing keys.*

Proof: Immediate from Lemma 4.1.4. ■

Observation 4.3 *Lemma 4.1.3 implies Constraint 3.93.*

We now argue that the multiset algorithm satisfies a constraint placed on the use of LLX and SCX. This constraint is used to guarantee progress for SCX.

Observation 4.4 *Consider any execution with a configuration C after which no field of any Data-record changes. There is a total order on all Data-records created during this execution such that, if Data-record r_1 appears before Data-record r_2 in the sequence V passed to an invocation S of SCX whose linked LLXs begin after C , then $r_1 < r_2$.*

Proof: Since the LLXs linked to S begin after C , it follows immediately from the multiset code that V is a subsequence of nodes in the list. By Corollary 4.2, they occur in order of strictly increasing keys, so r_1 before r_2 in V implies $r_1.key < r_2.key$. Thus, we take the total order on keys to be our total order. ■

Definition 4.5 *The number of occurrences of $key \neq \infty$ in the data structure at time t is $count$ if there is a Data-record r in the data structure at time t such that $r.key = key$ and $r.count = count$, and zero, otherwise.*

We call an invocation of INSERT or DELETE **effective** if it performs a linearized invocation of SCX (which either returns TRUE, or does not terminate). From the code of INSERT and DELETE, each **effective** invocation of INSERT or DELETE performs exactly one linearized invocation of SCX, each invocation of INSERT that returns is effective, and each invocation of DELETE that returns TRUE is effective. We linearize each effective invocation of INSERT or DELETE at its linearized invocation of SCX. The linearization point for an invocation I of $\text{DELETE}(key, count)$ that returns FALSE is subtle. Suppose I returns FALSE after seeing $r.key \neq key$. Then, we must linearize it at a time when the nodes p and r returned by its invocation I' of SEARCH are both in the data structure and $p.next$ points to r . By Observation 4.3, Constraint 3.93 is satisfied. This means we can apply Lemma 3.97 to show that there is a time during I' when p is in the data structure and $p.next = r$ (so r is also in the data structure). We linearize I at the last such time. Now, suppose I returns FALSE after seeing $r.count < count$. Then, we must linearize it at a time when the node r returned by its invocation I' of SEARCH is both in the data structure, and satisfies $r.count < count$. As in the previous case, we can apply Lemma 3.97 to show that there is a time after the start of I' , and at or before when I reads a value v from $r.count$ at line 30, such that r is in the data structure and $r.count = v$. We linearize I at the last such time. Similarly, we linearize each GET at the last time after the start of the SEARCH in GET, and at or before when the GET reads a value v from $r.count$, such that r is in the data structure and $r.count = v$. Clearly, each operation is linearized during that operation.

Lemma 4.6 *At all times t , the multiset σ of keys in the data structure is equal to the multiset σ_L of keys that would result from the atomic execution of the sequence of operations linearized up to time t .*

Proof: We prove this claim by induction on the sequence of steps taken in the execution. Since $next$ pointers and $count$ fields can be changed only by linearized invocations of SCX (and key fields do not change), we need only consider linearized invocations of SCX when reasoning about σ . Thus, invocations of INSERT and DELETE that are not effective cannot change the data structure. Since invocations of GET do not invoke SCX, they cannot change the

data structure. Therefore, we need only consider effective invocations of `INSERT` and `DELETE` when reasoning about σ_L . Since each effective invocation of `INSERT` or `DELETE` is linearized at its linearized invocation of `SCX`, the steps that can affect σ and σ_L are exactly the same. **Base case.** Before any linearized `SCX` has occurred, no *next* pointer has been changed. Thus, the data structure is in its initial configuration, which implies $\sigma = \emptyset$. Since no effective invocation of `INSERT` or `DELETE` has been linearized, $\sigma_L = \emptyset$. **Inductive step.** Let s be a linearized invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$, I be the (effective) invocation of `INSERT` or `DELETE` that performs S , and p, r and r_{next} refer to the local variables of I . Suppose $\sigma = \sigma_L$ before s . Let σ' denote σ after s , and σ'_L denote σ_L after s . We prove $\sigma' = \sigma'_L$.

Suppose S is performed at line 19. Then, I is an invocation of `INSERT(key, count)`, and $\sigma'_L = \sigma_L + \{\text{count copies of } key\}$. By Lemma 4.1.3, no Data-record is added or removed by S . Before I performs S , I performs an invocation L of `LLX(r)` linked to S at line 17. Since S is linearized, no mutable field of r changes between when L and S are linearized. Therefore, the value $localr.count$ that L reads from $r.count$ is equal to the value of $r.count$ at all times between when L and S are linearized, and line 19 implies that S changes $r.count$ from $localr.count$ to $localr.count + count$. Since S is linearized, Lemma 3.94 implies that r must be in the data structure just before S is linearized. By Lemma 4.1.4, r is the only Node in the data structure with key key , so σ contains exactly v copies of key just before S is linearized. Since this is the only change made by S , $\sigma' = \sigma + \{\text{count copies of } key\}$, and the inductive hypothesis implies $\sigma' = \sigma'_L$.

Suppose S is performed at line 23. Then, I is an invocation of `INSERT(key, count)`, and $\sigma'_L = \sigma_L + \{\text{count copies of } key\}$. By Lemma 4.1.3, no Data-record is removed by S , and only new is added by S . From the code of `INSERT`, $new.key = key$ and $new.count = count$. Therefore, $\sigma' = \sigma + \{\text{count copies of } key\}$, and the inductive hypothesis implies $\sigma' = \sigma'_L$.

Suppose S is performed at line 32. Then, I is an invocation of `DELETE(key, count)`. Before I performs S , I performs an invocation L of `LLX(r)` linked to S at line 28. Since S is linearized, no mutable field of r changes between when L and S are linearized. Thus, the value $localr.count$ that L reads from $r.count$ is equal to the value of $r.count$ at all times between when L and S are linearized. This implies that I sees $r.key = key$ and $r.count \geq count$ at line 30. By Lemma 4.1.3, r is the only Data-record removed by S , and new is the only Data-record added by S . By Definition 3.92, r must be in the data structure just before S is linearized. By Lemma 4.1.4, r is the only Node in the data structure with key key . Hence, σ contains exactly $localr.count$ copies of key just before S is linearized. From the code of `DELETE`, $new.key = r.key$ and $new.count = localr.count - count$. Therefore, $\sigma' = \sigma - \{\text{count copies of } key\}$. By the inductive hypothesis, $\sigma = \sigma_L$. Thus, there are $localr.count \geq count$ copies of key in σ_L . Therefore, if I is performed atomically at its linearization point, it will enter the if-block at line 31, so $\sigma'_L = \sigma_L - \{\text{count copies of } key\} = \sigma'$.

Suppose S is performed at line 35. Then, I is an invocation of `DELETE(key, count)`. Before I performs S , I performs an invocation L of `LLX(r)` linked to S at line 28. Since S is linearized, no mutable field of r changes between when L and S are linearized. Thus, the value $localr.count$ that L reads from $r.count$ is equal to the value of $r.count$ at all times between when L and S are linearized. This implies that I sees $r.key = key$ and $r.count \geq count$ at line 30, and $count \geq r.count$ at line 31. Hence, $r.count = count$ at all times between when L and S are linearized. Let r_{next} be the Node pointed to by I 's local variable $localr.next$. (We know r_{next} is a Node, and not NIL, from $r.key = key < \infty$ and Lemma 4.1.4.) After L , I performs an invocation L' of `LLX(r_{\text{next}})` linked to S at line 34. By the same argument as for $r.count$, the value v that L' reads from $r_{\text{next}}.count$ is equal to the value of $r_{\text{next}}.count$ at all times between when L' and S are linearized. By Lemma 4.1.3, r and r_{next} are the only Data-records removed by S , and new is the only Data-record added by S . By Definition 3.92, r and r_{next} must be in the data structure just before S . By Lemma 4.1.4, r is the only Node in the data structure with key key , and r_{next} is the only Node in the data structure with its key. Hence, σ contains exactly $r.count = count$ copies of key , and exactly v copies of $r_{\text{next}}.key$. From the code of `DELETE`,

$new.key = rnext.key$ and $new.count = rnext.count = v$. Therefore, $\sigma' = \sigma - \{count \text{ copies of } key\}$ By the inductive hypothesis, $\sigma = \sigma_L$. Thus, there are exactly $count$ copies of key just before I in the linearized execution. From the code of DELETE, in the linearized execution, I will enter the else block at line 33, so $\sigma'_L = \sigma_L - \{count \text{ copies of } key\} = \sigma'$. \blacksquare

Lemma 4.7 *Each invocation of GET(key) that terminates returns the number of occurrences of key in the data structure just before it is linearized.*

Proof: Consider any invocation I of GET(key). Let I' be the invocation of SEARCH(key) performed by GET(key), and p and r refer to the local variables of I' . By Lemma 4.1.2, I' satisfies its postcondition, which means that $key \leq r.key$, and either $p.key < key$ or $p = head$. We proceed by cases. Suppose $key = r.key$. Then, after I' , I reads a value v from $r.count$ and returns v . By Observation 4.3, Constraint 3.93 is satisfied. By Lemma 3.97, there is a time after the start of I' , and at or before when I reads $r.count$, such that r is in the data structure and $r.count = v$. I is linearized at the last such time. By Corollary 4.2, r is the only Data-record in the list that contains key key . Suppose that either $key < r.key$ and $p = head$, or $key < r.key$ and $p.key < key$. Then, I returns zero. By Lemma 3.97, at sometime during I' , p was in the data structure and $p.next$ pointed to r . I is linearized at the last such time. By Corollary 4.2, the data structure contains no occurrences of key when I is linearized. \blacksquare

Lemma 4.8 *Each invocation I of DELETE(key, count) that terminates returns TRUE if the data structure contains at least count occurrences of key just before I is linearized, and FALSE otherwise.*

Proof: Case I: I returns FALSE. In this case, I satisfies $key \neq r.key$ or $localr.count < count$ at line 30. Suppose $key \neq r.key$. Then, by the postcondition of SEARCH, $key < r.key$, and either $p.key < key$ or $p = head$. By Observation 4.3, Constraint 3.93 is satisfied. By Lemma 3.97, there is a time during the preceding invocation I' of SEARCH, when p was in the data structure and $p.next$ pointed to r . I is linearized at the last such time. Corollary 4.2 implies that there are no occurrences of key in the data structure when I is linearized. By the precondition of DELETE, $count > 0$, so the claim is satisfied.

Now, suppose $localr.count < count$ at line 30. By Lemma 3.97, there is a time after the start of I' , and before I 's LLX(r) reads $localr.count$ from $r.count$, such that r is in the data structure and $r.count = localr.count$. I is linearized at the last such time. By Corollary 4.2, r is the only Data-record in the list that contains key key , so there are $r.count < count$ occurrences of $r.key = key$ in the data structure when I is linearized.

Case II: I returns TRUE. In this case, I satisfies $key = r.key$ and $localr.count \geq count$ at line 30, and I is linearized at an invocation S of SCX at line 32 or 35. In each case, Lemma 4.1.3 implies that r is removed by S , so r is in the data structure just before S is linearized. Hence, r is in the data structure just before I is linearized. Before I performs S , I performs an invocation L of LLX(r) linked to S at line 28 that reads $localr.count$ from $r.count$. Since S is linearized, no mutable field of r changes between when L and S are linearized. Therefore, the value of $localr.count$ is equal to the value of $r.count$ at all times between when L and S are linearized. Thus, just before I is linearized, r is in the data structure and $r.count \geq count$. Finally, Corollary 4.2 implies that r is the only Data-record in the list that contains key key , so the claim holds. \blacksquare

In Section 3.2.2, we explained that a process cannot invoke $SCX(V, R, fld, new)$ until it performs successful invocations of LLX on each $r \in V$, so an LLX(r) that returns FINALIZED can prevent a process from invoking $SCX(V, R, fld, new)$. Thus, if a process can repeatedly perform LLX(r) on a finalized Data-record r , then doing so

may prevent it from performing any successful SCX operations. As we discussed there, one way to prevent this from happening is to have each process explicitly keep track of all FINALIZED Data-records it has seen (so it can avoid performing LLX on them again). We now prove that this kind of explicit bookkeeping is unnecessary for the multiset algorithm, because processes will not repeatedly invoke $\text{LLX}(r)$ on a finalized Data-record r .

Lemma 4.9 *No process performs more than one invocation of $\text{LLX}(r)$ that returns FINALIZED, for any Data-record r .*

Proof: Let r be a Data-record. Suppose, to derive a contradiction, that a process p performs two invocations L and L' of $\text{LLX}(r)$ that return FINALIZED. Without loss of generality, let L occur before L' . From the code of INSERT and DELETE , p must perform an invocation of SEARCH , L , another invocation I of SEARCH , and then L' . Since L returns FINALIZED, it is linearized after an invocation S of $\text{SCX}(V, R, \text{fld}, \text{new})$ with $r \in R$. By Lemma 4.1.3, r is removed from the data structure by S . We now show that r cannot be added back into the data structure by any subsequent invocation of SCX . From the code of INSERT and DELETE , each invocation of $\text{SCX}(V', R', \text{fld}', \text{new}')$ that changes a next pointer is passed a newly created Node, that is not known to any other process, as its new' argument. This implies that new' is not initiated, and cannot have previously been removed from the data structure. Therefore, r is not in the data structure at any point during I . By Observation 4.3, Constraint 3.93 is satisfied. By Lemma 3.97, r is in the data structure at some point during I , which is a contradiction. ■

Finally, we use the progress guarantees provided by LLX and SCX to prove that the multiset is lock-free.

Lemma 4.10 *If operations (INSERT , DELETE and GET) are invoked infinitely often, then operations complete infinitely often.*

Proof: Suppose, to derive a contradiction, that operations are invoked infinitely often but, after some time t , no operation completes. If SCXs are performed infinitely often, then they will succeed infinitely often and, hence, operations will succeed infinitely often. Thus, there must be some time $t' \geq t$ after which no SCX is performed. Then, after t' , the data structure does not change, and only a finite number of nodes with keys different from ∞ are ever added to the data structure. Consider an invocation I of $\text{SEARCH}(\text{key})$ that is executing after t' . Each time I performs line 11, it reads a Node r_{next} from $r.\text{next}$, and $r_{\text{next}}.\text{key} > r.\text{key}$. Therefore, by Corollary 4.2, I will eventually see $r.\text{key} = \infty$ at line 9. This implies that every invocation of GET eventually completes. Therefore, INSERT and DELETE must be invoked infinitely often after t' . From the code of INSERT (DELETE), in each iteration of the while loop, a SEARCH is performed, followed by a sequence of LLXs . If these LLXs all return values different from FAIL or FINALIZED, then an invocation of SCX is performed. Since every invocation of SEARCH eventually completes, Definition 3.89 implies that invocations of SCX are set up infinitely often. Thus, invocations of SCX succeed infinitely often. From the code of INSERT and DELETE , after performing a successful invocation of SCX , an invocation of INSERT or DELETE will immediately return. ■

Chapter 5

A template for implementing trees

The binary search tree (BST) is among the most important data structures. Previous concurrent implementations of balanced BSTs without locks either used coarse-grained transactions, which limit concurrency, or lacked rigorous proofs of correctness. In this Chapter, we describe a general technique for producing a lock-free implementation of *any* data structure based on a down-tree (a directed acyclic graph of indegree one), with updates that modify any connected subgraph of the tree atomically. Our approach drastically simplifies the task of proving correctness. This makes it feasible to develop provably correct implementations of non-blocking balanced BSTs with updates that synchronize on a small constant number of nodes.

As with all concurrent implementations, the implementations obtained using our technique are more efficient if each update to the data structure involves a small number of nodes near one another. We call such an update *localized*. We use *operation* to denote an operation of the abstract data type (ADT) being implemented by the data structure. Operations that cannot modify the data structure are called *queries*. For some data structures, such as Patricia tries and leaf-oriented BSTs, operations modify the data structure using a single localized update. In some other data structures, operations that modify the data structure can be split into several localized updates that can be freely interleaved.

A particularly interesting application of our technique is to implement *relaxed-balance* versions of sequential data structures efficiently. Relaxed-balance data structures decouple updates that rebalance the data structure from operations, and allow updates that accomplish rebalancing to be delayed and freely interleaved with other updates. For example, a chromatic tree is a relaxed-balance version of a red-black tree (RBT) which splits up the insertion or deletion of a key and any subsequent rotations into a sequence of localized updates. There is a rich literature of relaxed-balance versions of sequential data structures [82], and several papers (e.g., [86]) have described general techniques that can be used to easily produce them from large classes of existing sequential data structures. The small number of nodes involved in each update makes relaxed-balance data structures perfect candidates for efficient implementation using our technique.

Our Contributions

- We provide a simple template that can be filled in to obtain an implementation of any update for a data structure based on a down-tree. We prove that any data structure that follows our template for all of its updates will automatically be linearizable and non-blocking. The template takes care of all process coordination, so the data structure designer is able to think of updates as atomic steps.
- To demonstrate the use of our template, we provide a complete, provably correct, non-blocking linearizable implementation of a chromatic tree [104], which is a relaxed-balanced version of a RBT. To our knowledge, this is the first provably correct, non-blocking balanced BST in which updates synchronize on a small constant number of nodes. Our chromatic trees always have height $O(c + \log n)$, where n is the number of keys stored in the tree and c is the number of insertions and deletions that are in progress (proved in Section 6.5).
- We show that sequential implementations of some queries are linearizable, even though they completely ignore concurrent updates. For example, an ordinary BST search (that works when there is no concurrency) also works in our chromatic tree. Ignoring updates makes searches very fast. We also describe how to perform successor queries in our chromatic tree, which interact properly with updates that follow our template.
- We show experimentally that our Java implementation of a chromatic tree rivals, and often significantly outperforms, known highly-tuned concurrent dictionaries, over a variety of workloads, contention levels and thread

counts. For example, with 128 threads, our algorithm outperforms Java’s non-blocking skip-list by 13% to 156%, the lock-based AVL tree of Bronson et al. by 63% to 224%, and a RBT that uses software transactional memory (STM) by 13 to 134 times (Section 6.7).

5.1 Related work

There are many lock-based implementations of search tree data structures. (See [3, 27] for state-of-the-art examples.) Here, we focus on implementations that do not use locks. Valois [121] sketched an implementation of non-blocking node-oriented BSTs from CAS. Fraser [55] gave a non-blocking BST using 8-word CAS, but did not provide a full proof of correctness. He also described how multi-word CAS can be implemented from single-word CAS instructions. Ellen et al. [52] gave a provably correct, non-blocking implementation of leaf-oriented BSTs directly from single-word CAS. A similar approach was used for k -ary search trees [31] and Patricia tries [113]. All three used the cooperative technique originated by Turek, Shasha and Prakash [120] and Barnes [17]. Howley and Jones [70] used a similar approach to build node-oriented BSTs. They tested their implementation using a model checker, but did not prove it correct. Natarajan and Mittal [101] give another leaf-oriented BST implementation, together with a sketch of correctness. Instead of marking nodes, it marks edges. This enables insertions to be accomplished by a single CAS, so they do not need to be helped. It also combines deletions that would otherwise conflict. All of these trees are not balanced, so the height of a tree with n keys can be $\Theta(n)$.

Tsay and Li [119] gave a general approach for implementing trees in a wait-free manner using LL and SC operations (which can, in turn be implemented from CAS, e.g., [9]). However, their technique requires every process accessing the tree (even for read-only operations such as searches) to copy an entire path of the tree starting from the root. Concurrency is severely limited, since every operation must change the root pointer. Moreover, an extra level of indirection is required for every child pointer.

Red-black trees [19, 59] are well known BSTs that have height $\Theta(\log n)$. Some attempts have been made to implement RBTs without using locks. It was observed that the approach of Tsay and Li could be used to implement wait-free RBTs [102] and, furthermore, this could be done so that only updates must copy a path; searches may simply read the path. However, the concurrency of updates is still very limited. Herlihy et al. [67] and Fraser and Harris [56] experimented on RBTs implemented using software transactional memory (STM), which only satisfied obstruction-freedom, a weaker progress property. Each insertion or deletion, together with necessary rebalancing is enclosed in a single large transaction, which can touch all nodes on a path from the root to a leaf.

Some researchers have attempted fine-grained approaches to build non-blocking balanced search trees, but they all use extremely complicated process coordination schemes. Spiegel and Reynolds [115] described a non-blocking data structure that combines elements of B-trees and skip lists. Prior to this paper, it was the leading implementation of an ordered dictionary. However, the authors provided only a brief justification of correctness. Braginsky and Petrank [24] described a B+tree implementation. Although they have posted a correctness proof, it is very long and complex.

In a balanced search tree, a process is typically responsible for restoring balance after an insertion or deletion by performing a series of rebalancing steps along the path from the root to the location where the insertion or deletion occurred. Chromatic trees, introduced by Nurmi and Soisalon-Soininen [104], decouple the updates that perform the insertion or deletion from the updates that perform the rebalancing steps. Rather than treating an insertion or deletion and its associated rebalancing steps as a single, large update, it is broken into smaller, localized updates that can be

interleaved, allowing more concurrency. This decoupling originated in the work of Guibas and Sedgewick [59] and Kung and Lehman [79]. We use the leaf-oriented chromatic trees by Boyar, Fagerberg and Larsen [22]. They provide a family of local rebalancing steps which can be executed in any order, interspersed with insertions and deletions. Moreover, an amortized *constant* number of rebalancing steps per INSERT or DELETE is sufficient to restore balance for any sequence of operations. We have also used our template to implement a non-blocking version of Larsen’s leaf-oriented relaxed AVL tree [83]. In such a tree, an amortized *logarithmic* number of rebalancing steps per INSERT or DELETE is sufficient to restore balance.

There is also a node-oriented relaxed AVL tree by Boug   et al. [21], in which an amortized *linear* number of rebalancing steps per INSERT or DELETE is sufficient to restore balance. Bronson et al. [27] developed a highly optimized fine-grained locking implementation of this data structure using optimistic concurrency techniques to improve search performance. Deletion of a key stored in an internal node with two children is done by simply marking the node and a later insertion of the same key can reuse the node by removing the mark. If all internal nodes are marked, the tree is essentially leaf-oriented. Crain et al. gave a different implementation using lock-based STM [37] and locks [38], in which *all* deletions are done by marking the node containing the key. Physical removal of nodes and rotations are performed by one separate thread. Consequently, the tree can become very unbalanced. Drachsler et al. [49] give another fine-grained lock-based implementation, in which deletion physically removes the node containing the key and searches are non-blocking. Each node also contains predecessor and successor pointers, so when a search ends at an incorrect leaf, sequential search can be performed to find the correct leaf. A non-blocking implementation of Boug  ’s tree has not appeared, but our template would make it easy to produce one.

5.2 LLX, SCX and VLX primitives

Our tree update template uses the LLX and SCX primitives described in Chapter 3. The benefit of using LLX and SCX is two-fold: the template can be described quite simply, and much of the complexity of its correctness proof is encapsulated in that of LLX and SCX. Recall that the implementation of the primitives from CAS in Chapter 3 is more efficient if the user of the primitives can guarantee that the following two constraints are satisfied. The first constraint prevents the ABA problem for the CAS steps that actually perform the updates, and the second prevents processes from encountering livelock due to freezing Data-records in different orders.

Constraint 1: Each invocation of $\text{SCX}(V, R, fld, new)$ tries to change fld to a value new that it never previously contained.

Constraint 2: Consider each execution that contains a configuration C after which the value of no field of any Data-record changes. There is a total order of all Data-records created during this execution such that, for every SCX whose linked LLXs begin after C , the V sequence passed to the SCX is sorted according to the total order.

It is easy to satisfy these two constraints using standard approaches (e.g., by attaching a version number to each field and sorting V sequences). However, we shall see that both constraints are *automatically* satisfied in a natural way when LLX and SCX are used according to our tree update template.

We assume there is a Data-record *entry* which acts as the entry point to the data structure and is never deleted. This Data-record points to the root of a down-tree. We represent an empty down-tree by a pointer to an empty Data-record. A Data-record is *in the tree* if it can be reached by following pointers from *entry*. A Data-record r is *removed from the tree* by an SCX if r is in the tree immediately prior to the linearization point of the SCX and is not in the tree

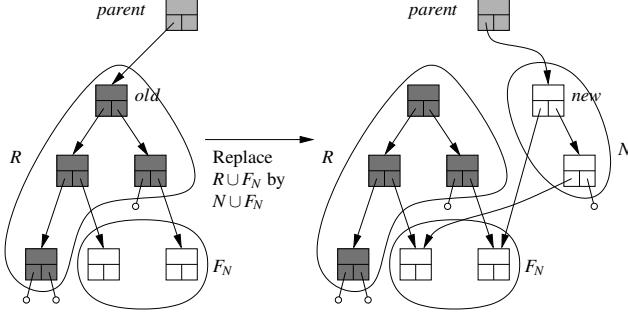


Figure 5.1: Example of the tree update template. R is the set of nodes to be removed, N is a tree of new nodes that have never before appeared in the tree, and F_N is the set of children of N (and of R). Nodes in F_N may have children. The shaded nodes (and possibly others) are in the sequence V of the SCX that performs the update. The darkly shaded nodes are finalized by the SCX.

immediately afterwards. Data structures produced using our template *automatically* satisfy one additional constraint:

Constraint 3: A Data-record is finalized when (and only when) it is removed from the tree.

Recall that, under this additional constraint, the implementation of LLX and SCX in Chapter 3 also guarantees the following three properties. These properties are useful for proving the correctness of our template.

- If $\text{LLX}(r)$ returns a snapshot, then r is in the tree just before the LLX is linearized.
- If an $\text{SCX}(V, R, \text{fld}, \text{new})$ is linearized and new is (a pointer to) a Data-record, then this Data-record is in the tree immediately after the SCX is linearized.
- If an operation reaches a Data-record r by following pointers read from other Data-records, starting from entry , then r was in the tree at some earlier time during the operation.

In the following, we sometimes abuse notation by treating the sequences V and R as sets, in which case we mean the set of all Data-records in the sequence.

The memory overhead introduced by the implementation of LLX and SCX is fairly low. Each node in the tree is augmented with a pointer to a descriptor and a bit. Every node that has had one of its child pointers changed by an SCX points to a descriptor. (Other nodes have a NIL pointer.) A descriptor can be implemented to use only three machine words after the update it describes has finished.

5.3 Tree update template

Our tree update template implements updates that atomically replace an old connected subgraph in a down-tree by a new connected subgraph. Such an update can implement any change to the tree, such as an insertion into a BST or a rotation used to rebalance a RBT. The old subgraph includes all nodes with a field (including a child pointer) to be modified. The new subgraph may have pointers to nodes in the old tree. Since every node in a down-tree has indegree one, the update can be performed by changing a single child pointer of some node parent . (See Figure 5.1.) However, problems could arise if a concurrent operation changes the part of the tree being updated. For example, nodes in the old subgraph, or even parent , could be removed from the tree before parent 's child pointer is changed. Our template takes care of the process coordination required to prevent such problems.

Each tree node is represented by a Data-record with a fixed number of child pointers as its mutable fields (but different nodes may have different numbers of child fields). Each child pointer points to a Data-record or contains NIL

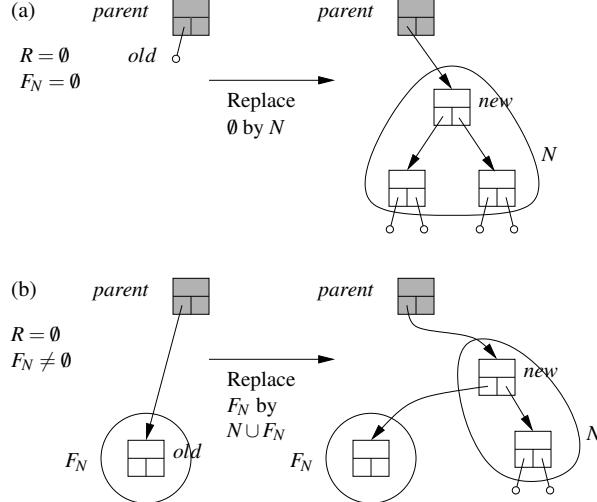


Figure 5.2: Examples of two special cases of the tree update template when no nodes are removed from the tree. (a) Replacing a NIL child pointer: In this case, $R = F_N = \emptyset$. (b) Inserting new nodes in the middle of the tree: In this case, $R = \emptyset$ and F_N consists of a single node.

(denoted by \circ in our figures). For simplicity, we assume that any other data in the node is stored in immutable fields. Thus, if an update must change some of this data, it makes a new copy of the node with the updated data.

At a high level, an update that follows the template proceeds in two phases: the *search phase* and the *update phase*. In the search phase, the update searches for a location where it should occur. Then, in the update phase, the update performs LLXs on a connected subgraph of nodes in the tree, including *parent* and the set R of nodes to be removed from the tree. Next, it decides whether the tree should be modified, and, if so, performs an SCX that atomically changes a child pointer, as shown in Figure 5.1, and finalizes the nodes in R . Figure 5.2 shows two special cases where R is empty. (As a minor note, search operations that do not perform any update can also be said to follow the template.)

Detailed description Figure 5.3 presents code for the template. An update is said to *follow* the template if it performs the sequence of steps in TEMPLATE.

Consider an update UP that follows the template. UP first invokes a SEARCHPHASE procedure to find the location in the tree where the update should occur. SEARCHPHASE reads a sequence of child pointers starting from *entry* to identify some node n_0 , and returns the *relevant* part m of the subtree rooted at n_0 that it observed. The node n_0 *must* be in the data structure at some time during SEARCHPHASE. (This requirement will be used to prove progress.) Intuitively, m contains only information that: (a) is observed by the search in the subtree rooted at n_0 , and (b) is relevant to the update. We require m to be a connected subgraph rooted at n_0 . (Consequently, m cannot be empty.) It is not necessary for m to contain a value for every field of a node. For instance, it may contain a value for the left child pointer of a node, but not for the right child pointer. As we will see, the update will modify the tree *only* if the fields of nodes in m agree with the values stored in m when the update is linearized. Next, UP invokes UPDATEPHASE(m).

In UPDATEPHASE, UP first determines whether it should modify the data structure by performing some deterministic local computation (denoted by UPDATENOTNEEDED in Figure 5.3) using the contents of m . If UP decides that an update is not needed (e.g., because UP is a deletion of a key that is not in the data structure), then it returns a result calculated locally by the RESULT function (and the update *succeeds*). Now, suppose UP decides that an update

```

1   TEMPLATE(args)
2     m := SEARCHPHASE(args)           ▷ search in the tree starting at entry to identify a location for the update
3     return UPDATEPHASE(m)
4
5   UPDATEPHASE(m)
6     if UPDATEREQUIRED(m) then return RESULT(m)
7     i := 0
8     ni := root of m
9     loop
10    si := LLX(ni)
11    if si ∈ {FAIL, FINALIZED} or CONFLICT(ni, si, m) then return FAIL
12    s'i := immutable fields of ni
13    exit loop when CONDITION(s0, s'0, ..., si, s'i)           ▷ CONDITION must eventually return TRUE
14    ni+1 := NEXTNODE(s0, s'0, ..., si, s'i)           ▷ returns a non-NIL child pointer from one of s0, ..., si
15    i := i + 1
16  end loop
17  if UPDATEREQUIRED(s0, s'0, ..., si, s'i) then return RESULT(s0, s'0, ..., si, s'i)
18  if SCX(SCX-ARGUMENTS(s0, s'0, ..., si, s'i)) then return RESULT(s0, s'0, ..., si, s'i)
19  else return FAIL

```

Figure 5.3: The tree update template. CONDITION, NEXTNODE, CONFLICT, SCX-ARGUMENTS and RESULT can be filled in with any locally computable functions that satisfy their postconditions. SEARCHPHASE should be filled in with a search procedure for the data structure.

is needed. Then, UP performs LLXs on a sequence $\sigma = \langle n_0, n_1, \dots \rangle$ of nodes starting with n_0 . This sequence must contain every node in m . For maximal flexibility of the template, the sequence σ can be constructed on-the-fly, as LLXs are performed. Thus, UP chooses a non-NIL child of one of the previous nodes to be the next node of σ by performing some deterministic local computation (denoted by NEXTNODE in Figure 5.3) using any information that is available locally, namely, the contents of m , snapshots of mutable fields returned by LLXs on the previous elements of σ , and values read from immutable fields of previous elements of σ . (This flexibility can be used, for example, to avoid unnecessary LLXs when deciding how to rebalance a BST.)

After each invocation of LLX, UP checks if the result of the LLX was FAIL or FINALIZED. If so, UP returns FAIL to indicate that it was aborted because of a concurrent update on an overlapping portion of the tree. Otherwise, UP performs another local computation, denoted by CONFLICT(n_i, s_i, m), which returns TRUE if the results of the LLX(n_i) disagree with m , and FALSE otherwise. For example, suppose UP is a deletion of a key k , and m contains a leaf l with key k and a node p that points to l . Then, if UP performs an LLX(p) which shows that p does not point to l , CONFLICT must return TRUE (and cause UP to return FAIL). Note that, if CONFLICT(n_i, s_i, m) returns TRUE, then a field of n_i changed since the value of that field in m was read in SEARCHPHASE. Thus, CONFLICT can cause UP to fail only if UP is concurrent with a successful update. This implies that UP can return FAIL without threatening the non-blocking progress property. (Similarly, whenever one can prove that a node has changed since UP performed LLX on it, UP can immediately return FAIL. Note that this possibility is not reflected explicitly in the template.)

Next, UP performs another local computation (denoted by CONDITION in Figure 5.3) to decide whether more LLXs should be performed. To avoid infinite loops, this function must eventually return TRUE in any execution of UP. (This condition is trivially satisfied if the loop in the template has a bounded number of iterations.) If all of the LLXs successfully return snapshots, then UP decides whether it should modify the tree by once again invoking

`UPDATENOTNEEDED`. If it decides to modify the tree, it invokes SCX. If UP invokes SCX and the SCX fails, then UP returns FAIL. In all other cases, UP *succeeds* and returns a result computed locally by the `RESULT` function.

Just before UP performs an SCX, it invokes a function `SCX-ARGUMENTS` which uses locally available information to construct the arguments V, R, fld and new for the SCX. The postconditions that must be satisfied by `SCX-ARGUMENTS` are somewhat technical, but intuitively, they are meant to ensure that the arguments produced describe an update as shown in Figure 5.1 or Figure 5.2. The update must remove a connected set R of nodes from the tree and replace it by a connected set N of newly-created nodes that is rooted at new by changing the child pointer stored in fld to point to new . In order for this change to occur atomically, we include R , the nodes in m , and the node containing fld , in V . This ensures that if any of these nodes has changed since it was last accessed by one of UP's LLXs, the SCX will fail. The sequence V may also include any other nodes in σ . Formally, we require `SCX-ARGUMENTS` to satisfy ten postconditions. The first three are basic requirements of SCX.

PC1: V is a subsequence of σ .

PC2: The node *parent* containing the mutable field *fld* is in V .

PC3: R is a subsequence of V .

PC4: All nodes in m are in V .

The next three postconditions guarantee that the R sequence of an SCX contains exactly the nodes that it removes from the tree. The first two are quite simple. Let G_N be the directed graph $(N \cup F_N, E_N)$, where E_N is the set of all child pointers of nodes in N when they are initialized, and $F_N = \{y : y \notin N \text{ and } (x, y) \in E_N \text{ for some } x \in N\}$. Let old be the value read from *fld* by the LLX on *parent*.

PC5: If $old = \text{NIL}$ then $R = \emptyset$ and $F_N = \emptyset$.

PC6: If $R = \emptyset$ and $old \neq \text{NIL}$, then $F_N = \{old\}$.

Stating the last of these three postconditions formally requires some care, since the tree may be changing while UP performs its LLXs. If $R \neq \emptyset$, let G_R be the directed graph $(R \cup F_R, E_R)$, where E_R is the union of the sets of edges representing child pointers read from each $r \in R$ when it was last accessed by one of UP's LLXs and $F_R = \{y : y \notin R \text{ and } (x, y) \in E_R \text{ for some } x \in R\}$. G_R represents UP's view of the nodes in R according to its LLXs, and F_R is the *fringe* of G_R . If other processes do not change the tree while UP is being performed, then F_R contains the nodes that should remain in the tree, but whose parents will be removed and replaced. Therefore, we must ensure that the nodes in F_R are reachable from nodes in N (so they are not accidentally removed from the tree). Let G_σ be the directed graph $(\sigma \cup F_\sigma, E_\sigma)$, where E_σ is the union of the sets of edges representing child pointers read from each $r \in \sigma$ when it was last accessed by one of UP's LLXs and $F_\sigma = \{y : y \notin \sigma \text{ and } (x, y) \in E_\sigma \text{ for some } x \in \sigma\}$. Since G_σ , G_R and G_N are not affected by concurrent updates, the following postcondition can be proved using purely sequential reasoning, ignoring the possibility that concurrent updates could modify the tree during UP.

PC7: If G_σ is a down-tree and $R \neq \emptyset$, then G_R is a non-empty down-tree rooted at old and $F_N = F_R$.

The next two postconditions are used to satisfy Constraint 1, which is used to prove there is no ABA problem.

PC8: G_N is a non-empty down-tree rooted at new .

PC9: UP allocates memory for all nodes in N , including new .

Note that there is no loss of generality in requiring new to be a newly-allocated node: If we wish to change a child y of node x to `NIL` (to chop off the entire subtree rooted at y) or to a descendant of y (to splice out a portion of the tree), then, instead, we can replace x by a new copy of x with an updated child pointer. Likewise, if we want to delete the entire tree, then *entry* can be changed to point to a new, empty Data-record.

The next postcondition is used to satisfy Constraint 2, which is used to prove progress.

PC10: The sequences V constructed by all updates that take place entirely during a period of time when no SCXs change the tree structure must be ordered consistently according to a fixed tree traversal algorithm (for example, an in-order traversal or a breadth-first traversal).

Properties of updates that follow the template Lock-free progress is guaranteed for all updates that follow the template. We briefly discuss the correctness properties provided by the template. We consider each of the different lines where an update UP that follows the template can return from UPDATEPHASE. Ad-hoc correctness arguments are needed if UP returns after an invocation of UPDATENOTNEEDED at line 6 or line 17, because the behaviour of UPDATENOTNEEDED depends on the semantics of the data structure being implemented. If UP returns FAIL at line 11 or line 19, then it has no effect on shared memory. The last place where UP can return is after performing a successful SCX at line 18. In this case, we prove that the *update phase* of UP (i.e., its invocation of UPDATEPHASE) is atomic, and that the fields of all nodes in m agree with their values in m when the update phase occurs. Furthermore, if SEARCHPHASE satisfies the following property, then the *entire template update is atomic* (including SEARCHPHASE).

Delayed traversal property (DTP): Suppose an invocation of SEARCHPHASE(args) terminates and returns m in configuration C , and an *atomic* invocation S' of SEARCHPHASE(args) is performed in a later configuration C' . If all of the nodes in m are in the tree in C' , and their fields agree with the values in m , then S' returns m .

5.4 Correctness proof

In the following, we use LLX and SCX as atomic primitives. Every invocation of SCX either succeeds or fails. A *successful* SCX modifies the data structure, and returns TRUE. A *failed* SCX does not modify the data structure, and returns FALSE. An LLX either succeeds or fails. A *successful* LLX returns FINALIZED or a snapshot. A *failed* LLX returns FAIL. Recall that an invocation of LLX is *linked to* an invocation I' of SCX($V, R, \text{fld}, \text{new}$) or VLX(V) by process p if r is in V , I returns a snapshot, and between I and I' , process p performs no invocation of LLX(r) or SCX($V', R', \text{fld}', \text{new}'$) and no unsuccessful invocation of VLX(V'), for any V' that contains r . We use the term *template operation* to refer to any operation that follows the tree update template. We call a template operation an *effective update* if it performs a successful SCX.

We now sketch the main ideas of the proof. Consider a data structure in which all updates follow the tree update template and SCX-ARGUMENTS satisfies postconditions PC1 to PC7. We linearize each *effective* update at its SCX. We prove, by induction on the sequence of steps in an execution, that the data structure is always a tree, each call to LLX and SCX satisfies its preconditions, Constraints 1 to 3 are satisfied, and each effective update atomically replaces a connected subgraph containing nodes $R \cup F_N$ with another connected subgraph containing nodes $N \cup F_N$ (finalizing and removing the nodes in R from the tree and adding the new nodes in N to the tree). Next, we prove that no node in the tree is finalized, every removed node is finalized, and removed nodes are never reinserted. Finally, we prove that effective updates have *atomic update phases*, and effective updates whose SEARCHPHASE procedures satisfy DTP are *entirely atomic* (including the search phase).

Additionally, a property was proved in Chapter 3 that allows some query operations to be performed very efficiently using only READs, for example, GET in Chapter 6.

QueryProp: If a process p follows child pointers starting from a node in the tree at time t and reaches a node r at time $t' \geq t$, then r was in the tree at some time between t and t' . Furthermore, if p reads v from a

mutable field of r at time $t'' \geq t'$ then, at some time between t and t'' , node r was in the tree and this field contained v .

Formal proof We now proceed with the formal proof. Since we refer to the preconditions of LLX and SCX in the following, we reproduce them here, for convenience.

- $\text{LLX}(r)$: r has been initiated (previously inserted into the tree)
- $\text{SCX}(V, R, \text{fld}, \text{new})$:
 1. for each $r \in V$, p has performed an invocation I_r of $\text{LLX}(r)$ linked to this SCX
 2. new is not the initial value of fld
 3. for each $r \in V$, no $\text{SCX}(V', R', \text{fld}, \text{new})$ occurred before I_r

The following lemma establishes Constraint 3, and some other properties that will be useful when proving linearizability.

Lemma 5.1 *The following properties hold in any execution of template operations.*

1. *Let S be a successful invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$, and G be the directed graph induced by the edges read by the LLXs linked to S . G is a sub-graph of the data structure at all times after the last LLX linked to S and before S , and no node in the N set of S is in the data structure before S .*
2. *Every LLX or SCX performed by a template operation has valid arguments, and satisfies its preconditions.*
3. *Let S be a successful invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$, where fld is a field of parent, and old is the value read from fld by the LLX(parent) linked to S . S changes fld from old to new , replacing a connected subgraph containing nodes $R \cup F_N$ with another connected subgraph containing nodes $N \cup F_N$. Further, the Data-records added by S are precisely those in N , and the Data-records removed by S are precisely those in R .*
4. *At all times, root is the root of a tree of Nodes. (We interpret \perp as the empty tree.)*

Proof: We prove these claims by induction on the sequence of steps taken in the execution. Clearly, these claims hold initially. Suppose they hold before some step s . We prove they hold after s . Let O be the operation that performs s .

Proof of Claim 1. To affect this claim, s must be a successful invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$. Since s is successful, the semantics of SCX imply that, for each $r \in V$, no successful $\text{SCX}(V', R', \text{fld}', \text{new}')$ with $r \in V'$ occurs between the invocation I of $\text{LLX}(r)$ linked to s and s . Thus, for each $r \in V$, no mutable field of r changes between I and s . We now show that all nodes and edges of G are in the data structure at all times after the last LLX linked to s , and before s . Fix any arbitrary $r \in V$. By inductive Claim 2, I satisfies its precondition, so r was initiated when I started and, hence, was in the data structure before I . By the semantics of SCX, since I returns a value different from FAIL or FINALIZED, no successful invocation of $\text{SCX}(V'', R'', \text{fld}'', \text{new}'')$ with $r \in R''$ occurs before I . By inductive Claim 3, Constraint 3 is satisfied at all times before s . Thus, r is not removed before I . Since s is successful, no successful invocation of $\text{SCX}(V'', R'', \text{fld}'', \text{new}'')$ with $r \in R''$ occurs between I and s . (If such an invocation were to occur then, since r would also be in V'' , the semantics of SCX would imply that s could not be successful.) Since I occurs before s , r is not removed before s . When I occurs, since r is in the data structure, all of its children are also in the data structure. Since no mutable field of r changes between I and s , all of r 's children read by I are in the data structure at all times after the last LLX linked to s , and before s .

structure throughout this time. Thus, each node and edge in G is in the data structure at all times after the last LLX linked to s , and before s .

Finally, we prove that no node in N is in the data structure before s is successful. Since O follows the tree update template, s is its only modification to shared memory. Since each $r' \in N$ is newly created by O , it is clear that r' can only be in the data structure after s .

Proof of Claim 2. Suppose s is invocation of $\text{LLX}(r)$. Then, $r \neq \text{NIL}$ (by the discussion in Sec. 5.3). By the code in Figure 5.3, either $r = \text{top}$, or r was obtained from the return value of some invocation L of $\text{LLX}(r')$ previously performed by O . If r was obtained from the return value of L , then Lemma 3.95 implies that r' is in the data structure when L occurs. Hence, r is in the data structure when L occurs, which implies that r is initiated when s occurs. Now, suppose $r = \text{top}$. By the precondition of O , r was reached by following child pointers from root since the last operation by p . By inductive Claim 3, Constraint 3 is satisfied at all times before s . Therefore, we can apply Lemma 3.97, which implies that r was in the data structure at some point before the start of O (and, hence, before s). By Definition 3.92, r is initiated when s begins.

Suppose s is an invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$. By PC2, fld is a mutable (child) field of some node $\text{parent} \in V$. By PC3, R is a subsequence of V . Therefore, the arguments to s are valid. By PC1 and the definition of σ , for each $r \in V$, O performs an invocation I of $\text{LLX}(r)$ before s that returns a value different from FAIL or FINALIZED (and, hence, is linked to s), so s satisfies Precondition 1 of SCX.

We now prove that s satisfies SCX Precondition 2. Let $\text{parent}.c_i$ be the field pointed to by fld . If $\text{parent}.c_i$ initially contains \perp then, by PC8, new is a Node, and we are done. Suppose $\text{parent}.c_i$ initially points to some Node r . We argued in the previous paragraph that O performs an $\text{LLX}(\text{parent})$ linked to s before s . By inductive Claim 3, Constraint 3 is satisfied at all times before s . Therefore, we can apply Lemma 3.97 to show that r was in the data structure at some point before the start of O (and, hence, before s). However, by inductive Claim 1 (which we have proved for s), new cannot be initiated before s , so $\text{new} \neq r$.

Finally, we show s satisfies SCX Precondition 3. Fix any $r' \in V$, and let L be the $\text{LLX}(r')$ linked to s performed by O . To derive a contradiction, suppose a successful invocation S' of $\text{SCX}(V', R', \text{fld}, \text{new})$ occurs before L (which occurs before s). By Lemma 3.96 (which we can apply since Constraint 3 is satisfied at all times before s), new would be in the data structure (and, hence, initiated) before s occurs. However, this contradicts our argument that new cannot be initiated before s occurs.

Proof of Claim 3 and Claim 4. To affect these claims, s must be a successful invocation of $\text{SCX}(V, R, \text{fld}, \text{new})$. The semantics of SCX and the fact that s is successful imply that, for each $r \in V$, no successful $\text{SCX}(V', R', \text{fld}', \text{new}')$ with $r \in V'$ occurs after the invocation I of $\text{LLX}(r)$ linked to s and before s . Thus, O satisfies tree update template PC5, PC7, PC8 and PC6. By inductive Claim 1 (which we have proved for s), all nodes and edges in G are in the data structure just before s , and no node in N is in the data structure before s . Let $\text{parent}.c_i$ be the mutable (child) field changed by s , and old be the value read from $\text{parent}.c_i$ by the $\text{LLX}(\text{parent})$ linked to s .

Suppose $R \neq \emptyset$ (as in Fig. 5.1). Then, by PC7, G_R is a tree rooted at old and $F_N = F_R$. Since G_R is a sub-graph of G , inductive Claim 1 implies that each node and edge of G_R is in the data structure just before s . Further, since O performs an $\text{LLX}(r)$ linked to s for each $r \in R$, and no child pointer changes between this LLX and s , G_R contains every node that was a child of a node in R just before s . Thus, F_R contains every node $r \notin R$ that was a child of a node in R just before s . This implies that, just before s , for each node $r \notin R$ in the sub-tree rooted at old , F_R contains r or an ancestor of r . By inductive Claim 4, just before s , every path from root to a descendent of old passes through old . Therefore, just before s , every path from root to a node in $\{\text{descendants of } \text{old}\} - R$ passes through a node in

F_R . Just before s , by the definition of F_R , and the fact that the nodes in R form a tree, $R \cap F_R$ is empty and no node in R is a descendent of a node in F . By PC8, G_N is a non-empty tree rooted at new with node set $N \cup F_N = N \cup F_R$, where N contains nodes that have not been in the data structure before s . Since $parent.c_i$ is the only field changed by s , s replaces a connected sub-graph with node set $R \cup F_R$ by a connected sub-graph with node set $N \cup F_R$. We prove that $parent$ was in the data structure just before s . Since s modifies $parent.c_i$, just before s , $parent$ must not have been finalized. Thus, no successful $SCX(V', R', fld', new')$ with $parent \in R'$ can occur before s . By inductive Claim 3, Constraint 3 is satisfied at all times before s , so $parent$ cannot be removed from the data structure before s . By inductive Claim 2, the precondition of the $LLX(parent)$ linked to s implies that $parent$ was initiated, so $parent$ was in the data structure just before s . Since no node in N is in the data structure before s , the Data-records added by s are precisely those in N . Since, just before s , no node in R is in F , or a descendent of a node in F , and every $r \in \{\text{descendents of } old\} - R$ is reachable from a node in F_R , the Data-records removed by s are precisely those in R .

Now, to prove Claim 4, we need only show that $parent.c_i$ is the root of a sub-tree just after s . We have argued that old is the root of G_R just before s . Since $parent.c_i$ points to old just before s , the inductive hypothesis implies that old is the root of a subtree, and $parent$ is not a descendent of old . Therefore, just before s , $parent$ is not in any sub-tree rooted at a node in F_R . This implies that no descendent of old is changed by s . By inductive Claim 4, each $r \in F$ is the root of a sub-tree just before s , so each $r \in F$ is the root of a sub-tree just after s . Finally, since PC8 states that G_N is a non-empty down-tree rooted at new , and we have argued that G_N has node set $N \cup F_R$, $parent$ is the root of a sub-tree just after s .

The two other cases, where $R = \emptyset$ and $old = \text{NIL}$ (as in Fig. 5.2a), and where $R = \emptyset$ and $old \neq \text{NIL}$ (as in Fig. 5.2b), are similar (and substantially easier to prove). ■

Corollary 5.2 *A Data-record is finalized when (and only when) it is removed from the tree (Constraint 3).*

Lemma 5.3 *After a Data-record r is removed from the data structure, it cannot be added back into the data structure.*

Proof: Suppose r is removed from the data structure. The only thing that can add r back into the data structure is a successful invocation S of $SCX(V, R, fld, new)$. Such an SCX must occur in an effective update O . By Lemma 5.1.3, every Data-record added by S is in O 's N set. By Lemma 5.1.1, no node in N is in the data structure at any time before S . Thus, r cannot be added to the data structure by S . ■

Observation 5.4 *A template operation O can modify the tree only if it is an effective update (i.e., it performs a successful invocation of SCX).*

Lemma 5.5 *Consider an effective update O . The invocation I of $UPDATEPHASE(m)$ performed by O behaves exactly as it would if it were performed atomically at O 's invocation S of SCX . (That is, I returns the same value, and performs the same invocations of LLX and SCX (with the same arguments), as it would if it were performed atomically at S .)*

Proof: Let I_L be the invocation of $UPDATEPHASE(m)$ performed by the update in the linearized execution that corresponds to O (i.e., that performs S). Note that I_L occurs atomically at S . We prove that the arguments and return values of the LLX s performed by I and I_L are the same. Recall that $s_0, s'_0, \dots, s_i, s'_i$ consist of the return values of the LLX s performed by the operation, and the immutable fields of the nodes accessed by these LLX s. It immediately follows that I and I_L have the same inputs to their local SCX -ARGUMENTS($s_0, s'_0, \dots, s_i, s'_i$) and $RESULT(s_0, s'_0, \dots, s_i, s'_i)$ computations. Consequently, I and I_L perform invocations of SCX with exactly the same arguments, and have the same return value.

We start by proving that each LLX performed by I returns the same value as it would if it were performed atomically at S (which is also when I_L occurs). Since S is successful, by the definition of LLX and SCX, for each $\text{LLX}(r)$ performed by I , no invocation of $\text{SCX}(V', R', \text{fld}', \text{new}')$ with $r \in V'$ occurs after this LLX and before S . Thus, no $r \in V$ changes after I 's $\text{LLX}(r)$ and before S .

We now prove that I and I_L perform the same LLXs, which return the same values. Let I^k and I_L^k be the k th LLXs by I and I_L , respectively, a^k and a_L^k be the respective arguments to I^k and I_L^k , and v^k and v_L^k be the respective return values of I^k and I_L^k . We prove by induction that $a^k = a_L^k$ and $v^k = v_L^k$ for all $k \geq 1$.

Base case: Since I and I_L have the same argument m , they both identify the same starting node n_0 for their LLXs. Hence, $a^1 = a_L^1 = n_0$. Since each LLX performed by I returns the same value as it would if it were performed atomically at S , $v^1 = v_L^1$.

Inductive step: Suppose the inductive hypothesis holds for $k - 1$, where $k > 1$. The NEXTNODE computation from which I obtains a^k depends only on v^1, \dots, v^{k-1} and the immutable fields of nodes a^1, \dots, a^{k-1} . Similarly, the NEXTNODE computation from which I_L obtains a_L^k depends only on v_L^1, \dots, v_L^{k-1} and the immutable fields of nodes a_L^1, \dots, a_L^{k-1} . Thus, by the inductive hypothesis, $a^k = a_L^k$. Since each LLX performed by I returns the same value as it would if it were performed atomically at S , we have $v^k = v_L^k$. Therefore, the inductive hypothesis holds for k , and the claim is proved. ■

Lemma 5.6 *Consider an effective update O that satisfies DTP. The invocation of SEARCHPHASE performed by O returns the same value m that it would if it were performed atomically at O 's SCX.*

Proof: We first argue that, when O performs its SCX, all of the nodes in m are in the tree, and their fields match the values in m . Since O is an effective update, it performs a successful $\text{SCX}(V, R, \text{fld}, \text{new})$. By inspection of the code in Figure 5.3, since O performs an SCX, for each node $n_i \in V$, O must perform an invocation of $\text{LLX}(n_i)$ that returns a value different from FAIL or FINALIZED, and subsequently perform an invocation of $\text{CONFLICT}(n_i, s_i, m)$ that returns FALSE. By PC4, V includes every node in m . Therefore, by the semantics of LLX and SCX, for each node $n_i \in m$, at all times after O 's $\text{LLX}(n_i)$ and before its SCX, n_i is in the tree and its fields match their values stored in m . The claim then follows from the fact that O satisfies DTP. ■

Theorem 5.7 *Every effective update whose SEARCHPHASE satisfies DTP is atomic.*

Proof: Consider an effective update O that satisfies DTP. The argument m to O 's invocation of UPDATEPHASE is returned from O 's invocation of SEARCHPHASE . By Lemma 5.6, O 's invocation of SEARCHPHASE would return the same argument m if it were performed atomically at O 's SCX. By Lemma 5.5, O 's invocation of $\text{UPDATEPHASE}(m)$ would have the same behaviour if it were performed atomically at O 's SCX. Thus, O has the same effect on the tree (and returns the same value) as it would if it were performed atomically at its SCX. ■

5.5 Progress proof

Recall the following definitions and progress property from Chapter 3. We use this progress property to prove progress for template operations. An SCX-UPDATE algorithm performs LLXs on a sequence V of Data-records and invokes $\text{SCX}(V, R, \text{fld}, \text{new})$ if all of these LLXs return snapshots. A successful SCX-UPDATE is one in which the SCX returns TRUE. Similarly, a VLX-QUERY algorithm performs LLXs on a sequence V of Data-records and invokes

$\text{VLX}(V)$ if all of these LLXs return snapshots. A successful VLX-QUERY is one in which the VLX returns TRUE. Note that every execution of a template operation that does not return at line 6 or line 17 is an execution of an SCX-UPDATE algorithm.

P2: Suppose that (a) there is always some non-finalized Data-record reachable by following pointers from an entry point, (b) for each Data-record r , each process performs finitely many invocations of $\text{LLX}(r)$ that return FINALIZED, and (c) processes perform infinitely many executions of SCX-UPDATE and/or VLX-QUERY algorithms. Then, infinitely many SCX or VLX operations succeed.

We briefly describe the proof at a high level. First, to ensure that processes can perform infinitely many template operations, we argue that all template operations are lock-free (resp., wait-free) if their invocations of SEARCHPHASE are lock-free (resp., wait-free). Next, we show that every data structure whose operations are template operations satisfies conditions (a), (b) and (c) of P2. Observe that, if each SCX and VLX is performed by a template operation, then P2 implies the following: if infinitely many template operations are performed, then infinitely many template operations succeed.

Lemma 5.8 *Let $\mathcal{P} \in \{\text{lock-freedom}, \text{wait-freedom}\}$. If the SEARCHPHASE of a template operation satisfies \mathcal{P} , and the implementation of LLX and SCX given in Chapter 3 is used, then the entire template operation satisfies \mathcal{P} .*

Proof: A template operation consists of an invocation of SEARCHPHASE, possibly followed by a loop in which LLX, CONDITION, CONFLICT and NEXTNODE are invoked (see Fig. 5.3), possibly followed by an invocation of SCX. All of the other steps comprise a finite local computation. The implementation of LLX and SCX given in Chapter 3 is wait-free. Similarly, NEXTNODE, CONDITION and CONFLICT perform finite local computations, and CONDITION must eventually return TRUE in every template operation that invokes it, causing the operation to exit the loop. Thus, apart from the SEARCHPHASE, the template operation is wait-free. ■

Lemma 5.9 *No process performs more than one invocation of $\text{LLX}(r)$ that returns FINALIZED, for any r , during template operations.*

Proof: Fix any Data-record r . Suppose, to derive a contradiction, that a process p performs two different invocations L and L' of $\text{LLX}(r)$ that return FINALIZED, during template operations. Then, since each template operation returns FAIL immediately after performing an $\text{LLX}(r)$ that returns FINALIZED, L and L' must occur in different template operations O and O' . Without loss of generality, suppose O occurs before O' .

Since L returns FINALIZED, it occurs after an invocation S of $\text{SCX}(\sigma, R, \text{fld}, \text{new})$ with $r \in R$ (by the semantics of LLX and SCX). By Lemma 5.1.3, r is removed from the data structure by S . By Lemma 5.3, r cannot be added back into the data structure. Thus, r is not in the data structure at any time after S occurs (and, hence, after O occurs). Recall that the first node n_0 on which O' performs LLX is the root of the part m of the tree returned by an invocation of SEARCHPHASE, and that n_0 is in the data structure at some point during this invocation of SEARCHPHASE. Thus, n_0 must be in the tree at some time after O but before O' . Consequently, n_0 cannot be r .

By inspection of the template in Figure 5.3, O' first performs LLX on n_0 , then performs LLX *only* on nodes that it obtained from the results of previous LLXs, and finally performs S . Since $r \neq n_0$, L' (which has r as its argument) must have obtained r from the result of an invocation L'' of $\text{LLX}(r')$ by O' that occurs after O and before L' . By Lemma 3.95, r' must be in the data structure just before L'' . Since L'' returns r , r' also points to r just before L'' . Therefore, r is in the data structure just before L'' , which contradicts the fact that r is not in the data structure at any time after O . ■

Theorem 5.10 *If SEARCHPHASE is lock-free, and template operations are performed infinitely often, then template operations succeed infinitely often.*

Proof: Suppose, to derive a contradiction, that template operations are performed infinitely often but, after some configuration C , no template operation succeeds. We show that the conditions of P2 are satisfied. Recall that the entry point to the data structure is never deleted (see Section 5.2). Consequently, Constraint 3 implies that the entry point is never finalized, so condition (a) of P2 is satisfied. Condition (b) of P2 is satisfied by Lemma 5.9. If a template operation performs a successful SCX, or returns at line 6, 17 or 18, then it is successful. Consequently, after C , no successful SCX occurs, and no template operation returns at line 6, 17 or 18. It follows that, after C , every template operation returns FAIL at line 11, after performing an LLX that returns FAIL or FINALIZED, or at line 19, after performing an unsuccessful SCX. Therefore, every execution of a template operation after C is an execution of an SCX-UPDATE algorithm, and Lemma 5.8 implies that each such operation is lock-free. Hence, there must be infinitely many executions of SCX-UPDATE algorithms, so condition (c) of P2 is satisfied. By P2, there must be infinitely many successful invocations of SCX or VLX. Since there are no invocations of VLX, and SCX is invoked only by template operations, and each template operation that performs a successful SCX succeeds, there must be infinitely many successful template operations, which is a contradiction. ■

Chapter 6

Chromatic tree implemented with the template

6.1 Chromatic trees

Here, we show how the tree update template can be used to implement an ordered dictionary ADT using chromatic trees. The ordered dictionary stores a set of keys, each with an associated value, where the keys are drawn from a totally ordered universe. The dictionary supports five operations. If key is in the dictionary, $\text{GET}(key)$ returns its associated value. Otherwise, $\text{GET}(key)$ returns \perp . $\text{SUCCESSOR}(key)$ returns the smallest key in the dictionary that is larger than key (and its associated value), or \perp if no key in the dictionary is larger than key . $\text{PREDECESSOR}(key)$ is analogous. $\text{INSERT}(key, value)$ replaces the value associated with key by $value$ and returns the previously associated value, or \perp if key was not in the dictionary. If the dictionary contains key , $\text{DELETE}(key)$ removes it and returns the value that was associated immediately beforehand. Otherwise, $\text{DELETE}(key)$ simply returns \perp .

A RBT is a BST in which the root and all leaves are coloured black, and every other node is coloured either red or black, subject to the constraints that no red node has a red parent, and the number of black nodes on a path from the root to a leaf is the same for all leaves. These properties guarantee that the height of a RBT is logarithmic in the number of nodes it contains. We consider search trees that are leaf-oriented, meaning the dictionary keys are stored in the leaves, and internal nodes store keys that are used only to direct searches towards the correct leaf. In this context, the BST property says that, for each node x , all descendants of x 's left child have keys less than x 's key and all descendants of x 's right child have keys that are greater than *or equal to* x 's key.

To decouple rebalancing steps from insertions and deletions, so that each is localized, and rebalancing steps can be interleaved with insertions and deletions, it is necessary to relax the balance properties of RBTs. A *chromatic tree* [104] is a relaxed-balance RBT in which colours are replaced by non-negative integer weights, where weight zero corresponds to red and weight one corresponds to black. As in RBTs, the sum of the weights on each path from the root to a leaf is the same. However, RBT properties can be violated in the following two ways. First, a red child node may have a red parent, in which case we say that a *red-red violation* occurs at this child. Second, a node may have weight $w > 1$, in which case we say that $w - 1$ *overweight violations* occur at this node. The root always has weight one, so no violation can occur there.

Rebalancing steps are localized updates to a chromatic tree that are performed at the location of a violation. Their goal is to eventually eliminate all red-red and overweight violations, while maintaining the invariant that the tree is a chromatic tree. If no rebalancing step can be applied to a chromatic tree (or, equivalently, the chromatic tree contains no violations), then it is a RBT. We use the set of rebalancing steps of Boyar, Fagerberg and Larsen [22] (which appear in Figure 6.5). This set of rebalancing steps has a number of desirable properties: No rebalancing step increases the number of violations in the tree, rebalancing steps can be performed in any order, and, after sufficiently many rebalancing steps, the tree will always become a RBT. Furthermore, in any sequence of insertions, deletions and rebalancing steps starting from an empty chromatic tree, the amortized number of rebalancing steps is at most three per insertion and one per deletion.

Figure 6.1 gives visualizations of three example chromatic trees. The first is the result of a single process sequentially inserting the keys 0, 1, 2, ..., 99,999. The depth of the tree is 34, but the average depth of leaves is approximately 19. This demonstrates the impact of rebalancing in the canonical worst-case example. The second tree was produced by having four processes uniformly randomly perform 50% insertions and 50% deletions of keys drawn uniformly from $[0, 10^3]$ for six seconds. The resulting tree contains approximately 500 keys, and has depth 13, with an average leaf depth of 11. The third tree was produced just like the second, but with the larger key range $[0, 10^6]$. The resulting tree contains approximately 500,000 keys, and has depth 25, with an average leaf depth of 21.

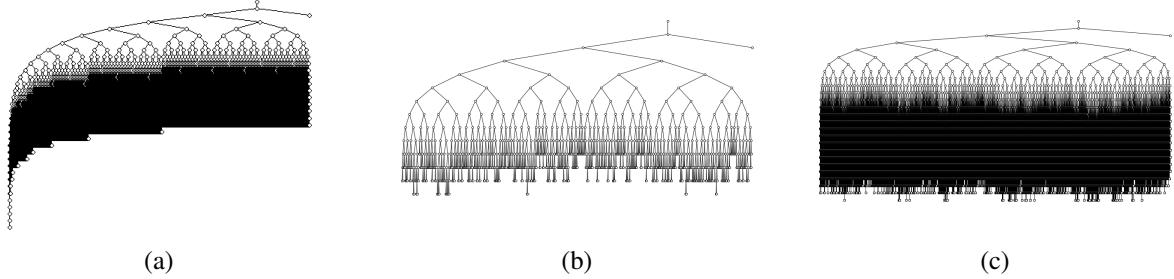


Figure 6.1: Visualizations of chromatic trees: (a) a single process inserting $0, 1, 2, \dots, 99999$. (b) four processes performing millions of random insertions and deletions of keys in $[0, 10^3]$. (c) four processes performing millions of random insertions and deletions of keys in $[0, 10^6]$.

```
type Node
  ▷ User-defined fields
  left, right    ▷ child pointers (mutable)
  k, v, w        ▷ key, value, weight (immutable)
  ▷ Fields used by LLX/SCX algorithm
  info          ▷ pointer to SCX-record
  marked        ▷ Boolean
```

Figure 6.2: Data definition for a node in the chromatic tree.

6.2 Implementation

We represent each node by a Data-record with two mutable child pointers, and immutable fields k , v and w that contain the node's key, associated value, and weight, respectively. (See Figure 6.2.) The child pointers of a leaf are always NIL, and the value field of an internal node is always NIL. To avoid special cases when the chromatic tree is empty, we add sentinel nodes at the top of the tree (see Figure 6.3). The entry point to the data structure, $entry$, is the topmost sentinel node. The chromatic tree is rooted at the leftmost grandchild of $entry$. For convenience we use $root$ to denote the current leftmost grandchild of $entry$. The sentinel nodes have key ∞ to avoid special cases for GET, INSERT and DELETE, and weight one to avoid special cases for rebalancing steps. The node $root$ also always has weight one. The sum of weights is the same for all paths from the root of the chromatic tree to its leaves.

Detailed pseudocode for GET, INSERT and DELETE is given in Figure 6.4, 6.7 and 6.6. Note that an expression of the form $P ? A : B$ evaluates to A if the predicate P evaluates to true, and B otherwise. The expression $x.y$, where x is a Data-record, denotes field y of x , and the expression $\&x.y$ represents a pointer to field y .

GET, INSERT and DELETE each execute an auxiliary procedure, $SEARCH(key)$, which appears in Figure 6.4. $SEARCH(key)$ starts at $entry$ and traverses nodes as in an ordinary BST search, using READS of child pointers until reaching a leaf, which it then returns (along with the leaf's parent, grandparent and great grandparent). Because of the sentinel nodes, the leaf's parent always exists, and the grandparent exists whenever the chromatic tree is non-empty. The grandparent is used whenever a DELETE actually deletes a key (which can happen only if the tree is non-empty) and in some rebalancing steps (which are performed only in a non-empty tree). The great grandparent is used only in some rebalancing steps (which are performed only when the great grandparent exists). If the grandparent (or great

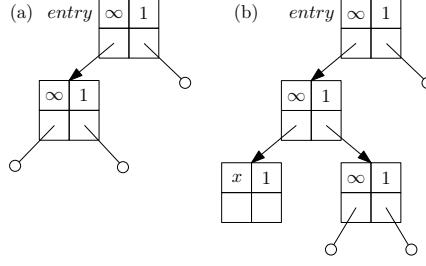


Figure 6.3: (a) empty tree, (b) non-empty tree.

```

1 GET(key)
2   ⟨−, −, −, l⟩ := SEARCH(key)
3   return (key = l.k) ? l.v : NIL
4
5 SEARCH(key)
6   ggp := NIL; gp := NIL; p := entry; l := entry.left
7   while l is internal
8     ggp := gp; gp := p; p := l
9     l := (key < p.k) ? p.left : p.right
return ⟨ggp, gp, p, l⟩

```

Figure 6.4: Pseudocode for GET and SEARCH.

grandparent) does not exist, then SEARCH returns NIL in its place. We define the *search path* for *key* at any time to be the path that SEARCH(*key*) would follow, if it were done instantaneously. The GET(*key*) operation simply executes a SEARCH(*key*) and then returns the value found in the leaf if the leaf's key is *key*, or \perp otherwise.

Figure 6.5 shows the complete set of transformations for the chromatic tree. INSERTNEW, INSERTREPLACE and DELETE change the keys and values in the tree. The other transformations are rebalancing steps. (Each rebalancing step also has a symmetric mirror-image version, denoted by an S after the name, except BLK and W7, which are their own mirror images.) The weight of a node appears to its right. The name of a node appears below it or to its left. We use a simple naming scheme for the nodes in the diagram. Consider the node u_x . We denote its left child by u_{xl} , and its right child by u_{xr} . Similarly, we denote the left child of u_{xl} by u_{xll} , and so on. (The subscript X indicates that we do not care whether a node is a left or right child.) Each transformation shown in Figure 6.5 is achieved by an SCX that swings a child pointer of u and depends on LLXs of all of the shaded nodes. The nodes marked with \times are finalized (and removed from the data structure). The nodes marked by a $+$ are newly created nodes. Nodes drawn as squares are leaves. Circular nodes may be either internal nodes or leaves. Nodes with no marking or shading, and no weight written beside them, do not necessarily have to exist for an operation to be applied.¹ For example, DELETE can be applied even if u_{xr} is a leaf. In this case, the children of u_{xr} are drawn simply to show where they will be reattached to the tree after the transformation, if they *do* exist.

At a high level, INSERT and DELETE are quite similar. INSERT(*key*, *value*) and DELETE(*key*) each perform SEARCH(*key*) and then make the required update at the leaf reached, in accordance with the tree update template. INSERT and DELETE can perform three different updates (shown in Figure 6.5). INSERTNEW inserts a key that was

¹However, sometimes these nodes are guaranteed to exist because, e.g., the tree is a full binary tree, all leaves have positive weight, and internal nodes do not change into leaves (or vice versa).

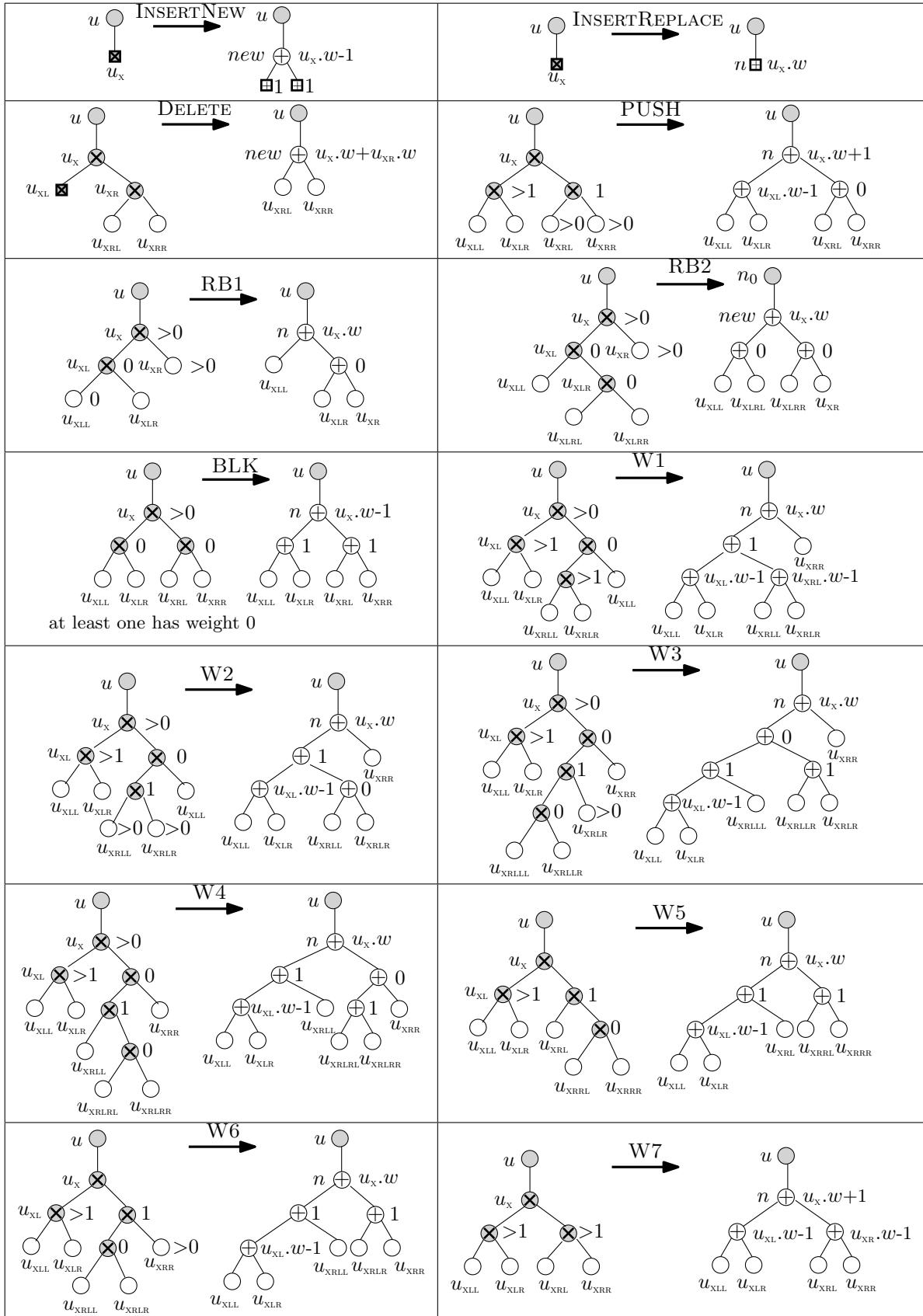


Figure 6.5: Transformations for chromatic search trees. Each transformation also has a mirror image.

not previously in the tree by replacing a leaf with a subtree containing a new internal (routing) node and two new leaves. `INSERTREPLACE` changes the value associated with a key in the tree by replacing a leaf with a new leaf. `DELETE` removes a key and its associated value from the tree by replacing a leaf, its sibling, and its parent with a new (routing) internal node. If the modification performed by an `INSERT` or `DELETE` fails, then the operation restarts from scratch. If it succeeds, it may increase the number of violations in the tree by one, and the new violation occurs on the search path to *key*. If a new violation is created, then an auxiliary procedure `CLEANUP` is invoked to fix it before the `INSERT` or `DELETE` returns.

Observe that an `INSERT(x, val)` in an empty tree will perform `INSERTNEW`, which will yield the tree in Figure 6.3(b). Similarly, deleting the last key in the tree will yield the tree in Figure 6.3(a). Thus, the sentinel nodes are automatically maintained by insertions and deletions without adding any special cases.

6.2.1 Detailed description of insertion

`INSERT(key, value)` invokes `TRYINSERT` to search for a leaf containing *key* and perform the localized update that actually associates *value* with *key*. In Figure 5.3, the template is presented using a loop, however `TRYINSERT` requires only two iterations, so we unroll the loop. To make this transformation as clear as possible, we still conceptually organize the steps of `TRYINSERT` into iterations.

`TRYINSERT` begins by invoking `SEARCH(key)` to find the leaf *l* on the search path to *key* and its parent *p*. In terms of the template, `SEARCH` represents the `SEARCHPHASE` procedure, *p* is *n₀*, and *p* and *l* are in the part *m* of the tree returned by `SEARCHPHASE(key)`. (Specifically, *m* contains an edge from *p* to *l*, and the keys of *p* and *l*.)

In iteration 0 (where *i* = 0), `TRYINSERT` performs `LLX(p)`. Then, it uses the result of the `LLX(p)` to verify that *p* still points to *l*. This verification step is conceptually part of the `CONFLICT` procedure. It is necessary because *p* might be changed between the `SEARCH` and `LLX(p)` so that it no longer points to *l*. If *p* no longer points to *l* (so it no longer matches *m*), then `TRYINSERT` returns `FAIL`, which indicates that the `INSERT` should be retried. In terms of the template, this corresponds to an invocation of `CONFLICT` returning `TRUE`. As in the template, `TRYINSERT` returns `FAIL` if any of its invocations of `LLX` return `FAIL` or `FINALIZED`. In iteration 1, `TRYINSERT` performs `LLX(l)`. This is the last iteration (which corresponds to `CONDITION` returning `TRUE` in the template).

`TRYINSERT` then computes `SCX-ARGUMENTS` over the next few lines. `TRYINSERT` uses locally stored values to construct the sequences *V* and *R* that it will use for its `SCX`, ordering their elements according to a breadth-first traversal, in order to satisfy PC10. Line 36 determines which child pointer *fld* of *p* should be changed by the insertion using the result of its `LLX(p)`. Line 37 creates a new leaf *newL* with weight 1, and the new key and value. Line 38 then checks whether *l* contains *key*. If so, `TRYINSERT` takes *newL* to be the `SCX` argument *new* at line 40 (so that *newL* will be inserted in place of *l*). Otherwise, `TRYINSERT` creates an internal node *new* (at line 45 or 47) with the appropriate weight, routing key, the value `NIL`, and the children *newL* and *l*, ordered according to their keys.

Finally, `TRYINSERT` invokes `SCX(V, R, fld, new)`. If the `SCX` succeeds, then `TRYINSERT` returns the result of the expression (*new.w* = *p.w* = 0), which indicates whether `TRYINSERT` created a red-red violation at *new*, and *oldValue*, which contains the value that was previously associated with *key* (or \perp if *key* was not in the tree). Otherwise, `TRYINSERT` returns `FAIL`.

An invocation of `INSERT` performs at most one successful invocation of `TRYINSERT`, and succeeds only if it performs a successful `TRYINSERT`. Unsuccessful invocations of `TRYINSERT` cannot create violations. A successful invocation of `TRYINSERT` can create at most one red-red violation at *new*, and cannot create any other violations. Thus,

```

10  INSERT(key,value)
11    ▷ Associates value with key in the dictionary and returns the old associated value, or  $\perp$  if none existed
12    do
13      result := TRYINSERT(key,value)
14      while result = FAIL
15       $\langle \text{createdViolation}, \text{value} \rangle := \text{result}$ 
16      if createdViolation then CLEANUP(key)
17      return value

18 TRYINSERT(key,value)
19   ▷ Returns  $\langle \text{TRUE}, \perp \rangle$  if key was not in the dictionary and inserting it caused a violation,
20    $\langle \text{FALSE}, \perp \rangle$  if key was not in the dictionary and inserting it did not cause a violation,
21    $\langle \text{FALSE}, \text{oldValue} \rangle$  if  $\langle \text{key}, \text{oldValue} \rangle$  was in the dictionary, and FAIL if we should try again

22   ▷ Search for key in the tree
23    $\langle -, -, p, l \rangle := \text{SEARCH}(key)$ 

24   ▷ Template iteration 0 (parent of leaf)
25   result := LLX(p)
26   if result  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return FAIL else  $\langle p_L, p_R \rangle := \text{result}$ 
27   if l  $\notin \{p_L, p_R\}$  then return FAIL           ▷ CONFLICT: verify p still points to l

28   ▷ Template iteration 1 (leaf)
29   result := LLX(l)
30   if result  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return FAIL else  $\langle l_L, l_R \rangle := \text{result}$ 

31   ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
32   V :=  $\langle p, l \rangle$ 
33   R :=  $\langle l \rangle$ 
34   fld := (l = pL) ? &p.right : &p.left
35   newL := new leaf with weight 1, key key and value value //
36   if l.k = key then
37     oldValue := l.v
38     new := newL
39   else
40     oldValue :=  $\perp$ 
41     if l is a sentinel node then w := 1 else w := l.w - 1
42     if key < l.k then
43       new := new node with weight w, key l.k, value NIL, and children newL and l
44     else
45       new := new node with weight w, key key, value NIL, and children l and newL

46   if SCX(V,R,fld,new) then return  $\langle (\text{new.w} = p.w = 0), \text{oldValue} \rangle$ 
47   else return FAIL

```

Figure 6.6: Pseudocode for INSERT and TRYINSERT.

```

51  DELETE(key)
52    ▷ Deletes key and returns its associated value, or returns ⊥ if key was not in the dictionary
53    do
54      result := TRYDELETE(key)
55      while result = FAIL
56      ⟨createdViolation,value⟩ := result
57      if createdViolation then CLEANUP(key)
58      return value
59
60 TRYDELETE(key)
61   ▷ Returns ⟨FALSE, ⊥⟩ if key was not in the dictionary,
62   ▷ ⟨FALSE, oldValue⟩ if ⟨key, oldValue⟩ was in the dictionary and deleting it did not create a violation,
63   ▷ ⟨TRUE, oldValue⟩ if ⟨key, oldValue⟩ was in the dictionary and deleting it created a violation, and
64   ▷ FAIL if we should try again
65
66   ▷ Search for key in the tree
67   ⟨−, gp, p, l⟩ := SEARCH(key)
68
69   ▷ UPDATENOTNEEDED: check if tree is empty
70   if gp = NIL then return ⟨FALSE, NIL⟩
71
72   ▷ UPDATENOTNEEDED: check if key is not in the dictionary
73   if l.k ≠ key then return ⟨FALSE, ⊥⟩
74
75   ▷ Template iteration 0 (grandparent of leaf)
76   result := LLX(gp)
77   if result ∈ {FAIL, FINALIZED} then return FAIL else ⟨gpl,gpr⟩ := result
78   if p ∉ {gpl,gpr} then return FAIL           ▷ CONFLICT: verify gp still points to p
79
80   ▷ Template iteration 1 (parent of leaf)
81   result := LLX(p)
82   if result ∈ {FAIL, FINALIZED} then return FAIL else ⟨pl,pr⟩ := result
83   if l ∉ {pl,pr} then return FAIL           ▷ CONFLICT: verify p still points to l
84
85   ▷ Template iteration 2 (leaf)
86   result := LLX(l)
87   if result ∈ {FAIL, FINALIZED} then return FAIL else ⟨lL,lR⟩ := result
88   s := (key < p.k) ? pr : pl
89
90   ▷ Template iteration 3 (sibling of leaf)
91   result := LLX(s)
92   if result ∈ {FAIL, FINALIZED} then return FAIL else ⟨sL,sR⟩ := result
93
94   ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
95   w := (p.k = ∞ or gp.k = ∞) ? 1 : p.w + s.w
96   new := new node with weight w, key s.k, value s.v, and children sL and sR
97   V := (key < p.k) ? ⟨gp, p, l, s⟩ : ⟨gp, p, s, lR := (key < p.k) ? ⟨p, l, s⟩ : ⟨p, s, lfld := (key < gp.k) ? &gp.left : &gp.right
100
101  if SCX(V,R,fld,new) then return ⟨(w > 1),l.v⟩
102  else return FAIL

```

Figure 6.7: Code for DELETE and TRYDELETE.

the expression ($new.w = p.w = 0$) returned by TRYINSERT indicates whether TRYINSERT created any violation. If TRYINSERT created a violation, then INSERT invokes CLEANUP(key) (which is described in more detail, below) to fix it before INSERT returns.

A simple inspection of the pseudocode suffices to prove that TRYINSERT follows the template, and the arguments to SCX satisfy the postconditions of SCX-ARGUMENTS.

6.2.2 Detailed description of deletion

DELETE(key) invokes TRYDELETE to search for a leaf containing key and perform the localized update that actually deletes key and its associated value. As in the description of insertion, we unroll the template loop, and organize the steps of TRYDELETE into iterations.

By inspection of the pseudocode, one can see that TRYDELETE follows the template. TRYDELETE begins by invoking SEARCH(key) to find the leaf l on the search path to key and its parent p and grandparent gp . In terms of the template, SEARCH represents the SEARCHPHASE procedure, gp is n_0 , and gp, p and l are in the part m of the tree returned by SEARCHPHASE(key). (Specifically, m contains an edge from gp to p , an edge from p to l , and the keys of gp, p and l .) If the grandparent does not exist, then the tree is empty (and it looks like Figure 6.3(a)), so TRYDELETE returns successfully at line 66. In terms of the template, this corresponds to an invocation of UPDATENOTNEEDED(m) that returns TRUE. Similarly, if l does not contain key , we can show there is a time during the SEARCH when the tree does not contain key , and TRYDELETE returns successfully at line 69. Note that, if TRYDELETE does *not* return at line 69, then l contains key .

In iteration 0 (where $i = 0$), TRYDELETE performs LLX(gp), and uses the result to verify that gp still points to p (as in m). If gp no longer points to p , then TRYDELETE returns FAIL. In terms of the template, this corresponds to an invocation of CONFLICT returning TRUE. As in the template, TRYDELETE returns FAIL if any of its invocations of LLX return FAIL or FINALIZED. In iteration 1, TRYDELETE performs LLX(p), and uses the result to verify that p still points to l . If p no longer points to l , then TRYDELETE returns FAIL. In iteration 2, TRYDELETE performs LLX(l). At line 84, TRYDELETE uses the result of its previous LLX(p) to obtain a pointer to the sibling, s , of the leaf to be deleted. In iteration 3, TRYDELETE performs LLX(s). This is the last iteration (which corresponds to CONDITION returning TRUE in the template).

TRYDELETE then computes SCX-ARGUMENTS over the next few lines. Line 91 computes the weight of the node new in the depiction of DELETE in Figure 6.3, ensuring that it has weight one if it is taking the place of a sentinel or $root$. Line 92 creates new , reading the key, and value directly from s (since they are immutable) and the child pointers from the result of the LLX(s). Next, TRYDELETE uses locally stored values (and the immutable key of p) to construct the sequences V and R that it will use for its SCX, ordering their elements according to a breadth-first traversal, in order to satisfy PC10. Finally, TRYDELETE uses key and $gp.k$ to decide which child pointer fld of gp should be changed by the deletion, and invokes SCX(V, R, fld, new) to perform the modification. If the SCX succeeds, then TRYDELETE returns the immutable value stored in node l , and the result of the expression $w > 1$, which indicates whether TRYDELETE created an overweight violation at the new node. If the SCX fails, then TRYDELETE returns FAIL. Inspection of the pseudocode suffices to prove that the arguments to SCX satisfy postconditions of SCX-ARGUMENTS.

An invocation of DELETE performs at most one successful invocation of TRYDELETE, and succeeds only if it performs a successful TRYDELETE. Unsuccessful invocations of TRYDELETE cannot create violations. A successful

```

99  CLEANUP(key)
100 > Ensures the violation created by an INSERT or DELETE of key gets eliminated
101 while TRUE
102   ggp := NIL; gp := NIL; p := NIL; l := entry           > Save four last nodes traversed
103   while TRUE
104     if l is a leaf then return                                > Arrived at leaf without finding a violation
105     if key < l.key then {ggp := gp; gp := p; p := l; l := l.left}
106     else {ggp := gp; gp := p; p := l; l := l.right}
107     if l.w > 1 or (p.w = 0 and l.w = 0) then    > Found a violation
108       TRYREBALANCE(ggp,gp,p,l)                  > Try to perform a rebalancing step
109       exit loop                                     > Go back to entry and traverse again

```

Figure 6.8: Pseudocode for CLEANUP.

invocation of TRYDELETE can create overweight violations at *new*, but cannot create any other violations. Thus, the expression $w > 1$ returned by TRYDELETE indicates whether TRYDELETE created any violation. If TRYDELETE creates a new violation, then DELETE invokes CLEANUP(*key*) to fix it.

6.2.3 The rebalancing algorithm

Since rebalancing is decoupled from updating, there must be a scheme that determines when processes should perform rebalancing steps to eliminate violations. In [22], the authors suggest maintaining one or more *problem queues* which contain pointers to nodes that contain violations, and dedicating one or more *rebalancing processes* to simply perform rebalancing steps as quickly as possible. This approach does not yield a bound on the height of the tree, since rebalancing may lag behind insertions and deletions. It is possible to obtain a height bound with a different queue based scheme, but we present a way to bound the tree's height without the (significant) overhead of maintaining any auxiliary data structures. The linchpin of our method is the following claim concerning violations, which is satisfied by each rebalancing step in Figure 6.5.

VIOL: If a violation is on the search path to *key* before a rebalancing step, then the violation is still on the search path to *key* after the rebalancing step, or it has been eliminated.

The rebalancing steps shown in Figure 6.5 are actually a slight modification of those in [22]. Specifically, in the original rebalancing steps, W1, W2, W3 and W4 (and their symmetric versions) require the node labeled u_x to have non-negative weight. In our rebalancing steps, we require u_x to have positive weight. The reason for this restriction is as follows.

While studying the original rebalancing steps, we realized that most of them satisfied VIOL. However, if W1, W2, W3 or W4 (or a symmetric version) were performed when u_x had weight 0, then VIOL could be violated. In each of these cases, we found that another rebalancing step which satisfies VIOL could be applied instead. (We argue, below, that a rebalancing step can always be applied whenever there is a violation in the tree.) Thus, under this restriction, all rebalancing steps satisfy VIOL (a fact that is easily verified). Consequently, each violation created by INSERT(*key*,*value*) or DELETE(*key*) stays on the search path to *key* until it is eliminated.

In our implementation, each INSERT or DELETE that increases the number of violations in the tree cleans up after itself by invoking CLEANUP(*key*). Pseudocode for CLEANUP appears in Figure 6.8. CLEANUP(*key*) behaves like

`SEARCH(key)` until it finds the first node l on the search path where a violation occurs. Then, `CLEANUP(key)` attempts to eliminate or move the violation at l by invoking another procedure `TRYREBALANCE` which applies one localized rebalancing step at l , following the tree update template. (`TRYREBALANCE` is described in more detail, below.) `CLEANUP(key)` repeats these actions, searching for key and invoking `TRYREBALANCE` to perform a rebalancing step, until the search goes all the way to a leaf without finding a violation.

In order to prove that each `INSERT` or `DELETE` cleans up after itself, we must prove that while an invocation of `CLEANUP(key)` searches for key by reading child pointers, it does not somehow miss the violation it is responsible for eliminating, even if a concurrent rebalancing step moves the violation upward in the tree, above where `CLEANUP` is currently searching. To see why this is true, consider any rebalancing step that occurs while `CLEANUP` is searching. The rebalancing step is implemented using the tree update template, and looks like Figure 5.1. It takes effect at the point it changes a child pointer fld of some node $parent$ from a node old to a node new . If `CLEANUP` reads fld while searching, we argue that it does not matter whether fld contains old or new . First, suppose the violation is at a node that is removed from the tree by the rebalancing step, or a child of such a node. If the search passes through old , it will definitely reach the violation, since nodes do not change after they are removed from the tree. If the search passes through new , `VIOL` implies that the rebalancing step either eliminated the violation, or moved it to a new node on the search path through new . Finally, if the violation is further down in the tree, below the section modified by the concurrent rebalancing step, a search through either old or new will reach it.

6.2.4 Deciding which rebalancing step to perform

Whenever `CLEANUP` finds a violation, it invokes `TRYREBALANCE`, and provides it with pointers to the node l where the violation was found, along with the node's parent p , grandparent gp and great grandparent ggp . In terms of the template, the search in `CLEANUP` corresponds to the `SEARCHPHASE` procedure, ggp is n_0 , and ggp, gp, p and l are in the part m of the tree returned by `SEARCHPHASE`. (Specifically, m contains an edge from ggp to gp , an edge from gp to p , an edge from p to l , and the keys of ggp, gp, p and l .) In the first three template iterations, `TRYREBALANCE` performs `LLX` on ggp, gp and p (at lines 113, 117 and 121), and verifies that ggp still points to gp , gp still points to p , and p still points to l . In terms of the template, these verification steps are part of the `CONFLICT` procedure. If any of these verification steps fails, then a node no longer matches m (so `CONFLICT` conceptually returns `TRUE`), and `TRYREBALANCE` returns `FAIL`. As in the template, if an `LLX` returns `FAIL` or `FINALIZED`, then `TRYREBALANCE` returns `FAIL`. If `TRYREBALANCE` returns `FAIL`, then `CLEANUP` will restart its search for violations from $entry$. In the remaining template iterations, `TRYREBALANCE` traverses the decision tree in Figure 6.9 to decide which rebalancing step should be performed. Finally, `TRYREBALANCE` invokes a procedure (e.g., `DOBLK` or `DORB2`) to perform the specific rebalancing step according to Figure 6.5.

We now describe the implementation of the decision tree. Pseudocode for `TRYREBALANCE`, including the implementation of the decision tree, appears in Figure 6.10 and 6.11. The first task in `TRYREBALANCE` is to identify whether the violation found by `CLEANUP` is a left or right overweight violation, or a left or right red-red violation. A left (resp., right) overweight violation is an overweight violation at a node that is a left (right) child. A left (resp., right) red-red violation is a red-red violation at a node whose $parent$ is a left (right) child. `TRYREBALANCE` identifies the type of violation by first determining whether the violation is an overweight violation or a red-red violation at line 125, and then further determining whether it is a left or right violation at line 126 (if it is an overweight violation) or line 133 (if it is a red-red violation). This information is sufficient to determine where `TRYREBALANCE` should

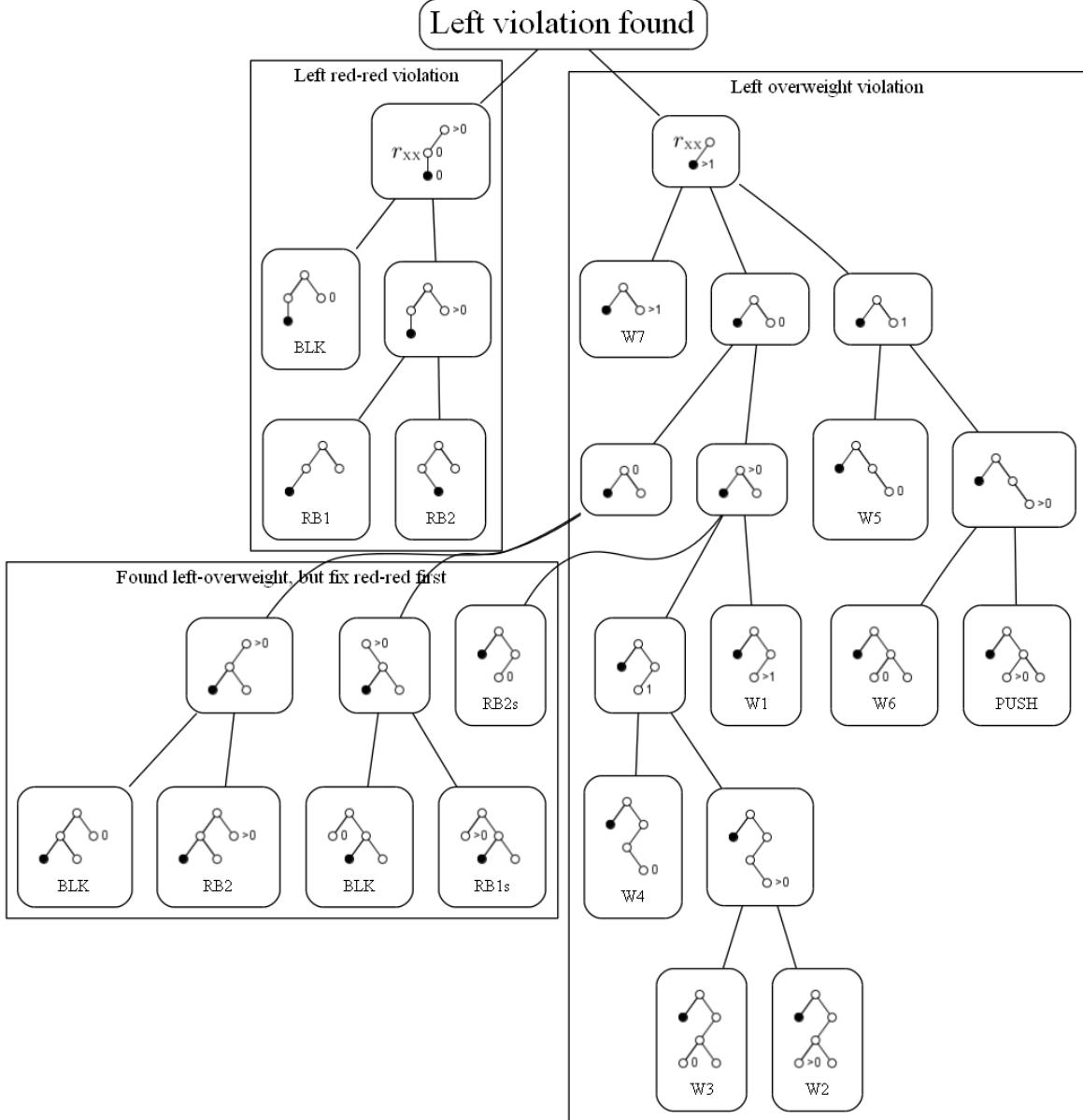


Figure 6.9: Decision tree used by the algorithm to determine which rebalancing operation to apply when a violation is encountered at a node (shaded black). The corresponding diagram to cover right violations can be obtained by: horizontally flipping each small tree diagram, changing each rebalancing step DOX to its symmetric version DOXS, and changing each symmetric version back to its original version.

```

110 TRYREBALANCE( $ggp, gp, p, l$ )
111 ▷ Precondition:  $l.w > 1$  or  $l.w = p.w = 0 \neq gp.w$ 
112    $r := ggp$ 
113   if ( $result := LLX(r)$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{l}}, r_{\text{r}} \rangle := result$ 
114   if  $gp \notin \{r_{\text{l}}, r_{\text{r}}\}$  then return                                ▷ CONFLICT: verify  $ggp$  still points to  $gp$ 

116    $r_x := gp$ 
117   if ( $result := LLX(r_x)$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xl}}, r_{\text{xr}} \rangle := result$ 
118   if  $p \notin \{r_{\text{xl}}, r_{\text{xr}}\}$  then return                                ▷ CONFLICT: verify  $gp$  still points to  $p$ 

120    $r_{\text{xx}} := p$ 
121   if ( $result := LLX(r_{\text{xx}})$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xxl}}, r_{\text{xxr}} \rangle := result$ 
122   if  $l \notin \{r_{\text{xxl}}, r_{\text{xxr}}\}$  then return                                ▷ CONFLICT: verify  $p$  still points to  $l$ 

124    $r_{\text{xxx}} := l$ 
125   if  $r_{\text{xxx}}.w > 1$  then                                              ▷ Overweight violation at  $l$ 
126     if  $r_{\text{xxx}} = r_{\text{xxl}}$  then                                              ▷ Left overweight violation ( $l$  is a left child)
127       if ( $result := LLX(r_{\text{xxl}})$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xxll}}, r_{\text{xxlr}} \rangle := result$ 
128       OVERWEIGHTLEFT(all  $r$  variables)
129     else ▷  $r_{\text{xxx}} = r_{\text{xxr}}$                                               ▷ Right overweight violation ( $l$  is a right child)
130       if ( $result := LLX(r_{\text{xxr}})$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xxrl}}, r_{\text{xxrr}} \rangle := result$ 
131       OVERWEIGHTRIGHT(all  $r$  variables)

132   else                                              ▷ Red-red violation at  $l$ 
133     if  $r_{\text{xx}} = r_{\text{xl}}$  then                                              ▷ Left red-red violation ( $p$  is a left child)
134       if  $r_{\text{xxr}}.w = 0$  then
135         if ( $result := LLX(r_{\text{xxr}})$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xrl}}, r_{\text{xrr}} \rangle := result$ 
136         DOBLK( $\langle r, r_x, r_{\text{xx}}, r_{\text{xxr}} \rangle$ , all  $r$  variables)
137       else if  $r_{\text{xxx}} = r_{\text{xxl}}$  then
138         DORB1( $\langle r, r_x, r_{\text{xx}} \rangle$ , all  $r$  variables)
139       else ▷  $r_{\text{xxx}} = r_{\text{xxr}}$ 
140         if ( $result := LLX(r_{\text{xxr}})$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xxrl}}, r_{\text{xxrr}} \rangle := result$ 
141         DORB2( $\langle r, r_x, r_{\text{xx}}, r_{\text{xxr}} \rangle$ , all  $r$  variables)

142   else ▷  $r_{\text{xx}} = r_{\text{xr}}$                                               ▷ Right red-red violation ( $p$  is a right child)
143     if  $r_{\text{xxl}}.w = 0$  then
144       if ( $result := LLX(r_{\text{xxl}})$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xll}}, r_{\text{xlr}} \rangle := result$ 
145       DOBLK( $\langle r, r_x, r_{\text{xxl}}, r_{\text{xx}} \rangle$ , all  $r$  variables)
146     else if  $r_{\text{xxx}} = r_{\text{xxr}}$  then
147       DORB1S( $\langle r, r_x, r_{\text{xx}} \rangle$ , all  $r$  variables)
148     else ▷  $r_{\text{xxx}} = r_{\text{xxl}}$ 
149       if ( $result := LLX(r_{\text{xxl}})$ )  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then return else  $\langle r_{\text{xxll}}, r_{\text{xxlr}} \rangle := result$ 
150       DORB2S( $\langle r, r_x, r_{\text{xx}}, r_{\text{xxl}} \rangle$ , all  $r$  variables)

```

Figure 6.10: Pseudocode for TRYREBALANCE.

```

151 OVERWEIGHTLEFT( $r, r_s, r_{xx}, r_{xxl}, r_l, r_r, r_{xl}, r_{xr}, r_{xxr}$ )
152   if  $r_{xxr}.w = 0$  then
153     if  $r_{xx}.w = 0$  then
154       if  $r_{xx} = r_{xl}$  then
155         if  $r_{xr}.w = 0$  then
156           if ( $result := LLX(r_{xr}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xrl}, r_{xrr} \rangle := result$ 
157           DOBLK( $\langle r, r_s, r_{xx}, r_{xr} \rangle$ , all  $r$  variables)
158         else  $\triangleright r_{xr}.w > 0$ 
159           if ( $result := LLX(r_{xxr}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrl}, r_{xxrr} \rangle := result$ 
160           DORB2( $\langle r, r_s, r_{xx}, r_{xxr} \rangle$ , all  $r$  variables)
161       else  $\triangleright r_{xx} = r_{xr}$ 
162         if  $r_{xl}.w = 0$  then
163           if ( $result := LLX(r_{xl}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xll}, r_{xlr} \rangle := result$ 
164           DOBLK( $\langle r, r_s, r_{xl}, r_{xx} \rangle$ , all  $r$  variables)
165         else  $\triangleright r_{xx} = r_{xl}$ 
166           DORB1S( $\langle r, r_s, r_{xx} \rangle$ , all  $r$  variables)
167     else  $\triangleright r_{xx}.w > 0$ 
168       if ( $result := LLX(r_{xxr}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrl}, r_{xxrr} \rangle := result$ 
169       if ( $result := LLX(r_{xxrl}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrll}, r_{xxrlr} \rangle := result$ 
170       if  $r_{xxrl}.w > 1$  then
171         DOW1( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr}, r_{xxrl} \rangle$ , result, all  $r$  variables)
172       else if  $r_{xxrl}.w = 0$  then
173         DORB2S( $\langle r_s, r_{xx}, r_{xxr}, r_{xxrl} \rangle$ , all  $r$  variables)
174       else  $\triangleright r_{xxrl}.w = 1$ 
175         if  $r_{xxrlr} = NIL$  then return  $\triangleright$  Special case: a node we performed LLX on was modified
176         if  $r_{xxrlr}.w = 0$  then
177           if ( $res := LLX(r_{xxrlr}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrll}, r_{xxrlr} \rangle := res$ 
178           DOW4( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr}, r_{xxrl}, r_{xxrlr} \rangle$ , all  $r$  variables)
179       else  $\triangleright r_{xxrlr}.w > 0$ 
180         if  $r_{xxrll}.w = 0$  then
181           if ( $res := LLX(r_{xxrl}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrll}, r_{xxrlr} \rangle := res$ 
182           DOW3( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr}, r_{xxrl}, r_{xxrll} \rangle$ , all  $r$  variables)
183         else  $\triangleright r_{xxrll}.w > 0$ 
184           DOW2( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr}, r_{xxrl} \rangle$ , all  $r$  variables)
185     else if  $r_{xxr}.w = 1$  then
186       if ( $result := LLX(r_{xxr}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrl}, r_{xxrr} \rangle := result$ 
187       if  $r_{xxr} = NIL$  then return  $\triangleright$  Special case: a node we performed LLX on was modified
188       if  $r_{xxr}.w = 0$  then
189         if ( $result := LLX(r_{xxr}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrll}, r_{xxrr} \rangle := result$ 
190         DOW5( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr}, r_{xxrr} \rangle$ , all  $r$  variables)
191     else if  $r_{xxr}.w = 0$  then
192       if ( $result := LLX(r_{xxrl}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrll}, r_{xxrlr} \rangle := result$ 
193       DOW6( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr}, r_{xxrl} \rangle$ , all  $r$  variables)
194     else  $\triangleright r_{xxr}.w > 0$  and  $r_{xxrl}.w > 0$ 
195       DOPUSH( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr} \rangle$ , all  $r$  variables)
196   else  $\triangleright r_{xxr}.w > 1$ 
197     if ( $result := LLX(r_{xxr}) \in \{FAIL, FINALIZED\}$ ) then return else  $\langle r_{xxrl}, r_{xxrr} \rangle := result$ 
198     DOW7( $\langle r_s, r_{xx}, r_{xxl}, r_{xxr} \rangle$ , all  $r$  variables)
199 OVERWEIGHTRIGHT( $r, r_s, r_{xx}, r_{xxr}, r_{xxl}, r_l, r_r, r_{xl}, r_{xr}, r_{xxl}$ )
200  $\triangleright$  Obtained from OVERWEIGHTLEFT by flipping each  $R$  in the subscript of an  $r$  variable to an  $L$  (and vice versa), and
    by flipping each rebalancing step DoX to its symmetric version DoXS (and vice versa).

```

Figure 6.11: Pseudocode for OVERWEIGHTLEFT (and effectively also OVERWEIGHTRIGHT).

start in the decision tree shown in Figure 6.9 (or in the symmetric, mirror-image version of this decision tree).

Once the violation type has been identified, TRYREBALANCE performs a sequence of LLXs, and uses their results (and the immutable fields of nodes) to decide which path to traverse in the decision tree. More specifically, at each node of the decision tree, TRYREBALANCE decides which child to proceed to by looking at the weight of one node, as indicated in the child (in Figure 6.9). (For clarity, we factorize the *left overweight* and *right overweight* parts of the decision tree traversal into two other functions, OVERWEIGHTLEFT and OVERWEIGHTRIGHT.) The leaves of the decision tree are labeled by the rebalancing step to apply. Note that this decision tree is a component of the sequential chromatic tree algorithm that was left to the implementer in [22].

Whenever TRYREBALANCE follows a pointer to a node in the chromatic tree, either we must be able to argue that the pointer is not NIL, or TRYREBALANCE must explicitly check whether the pointer is NIL. It is easy to prove that most pointers followed by TRYREBALANCE are non-NIL with the help of three simple invariants (which are easy to prove if rebalancing steps are atomic). First, each node is created with two NIL child pointers or two non-NIL child pointers, and a child pointer does not change from a non-NIL value to NIL, or vice versa. (Thus, if TRYREBALANCE has followed one child pointer of a node, then the other child pointer is non-NIL.) Second, each node with weight zero has two non-NIL child pointers, and node weights never change. Third, every node that has a red-red violation always has a parent, grandparent and great grandparent. These invariants suffice to prove that only two explicit NIL checks are necessary, specifically, at lines 175 and 187.

In each of these cases, TRYREBALANCE returns without performing any rebalancing step, and CLEANUP repeats its search for violations. This creates two potential problems. First, we would like to argue that a rebalancing step can always be applied, whenever the chromatic tree contains a violation, but in this case we return without performing a rebalancing step. Second, we would like to argue that returning without performing a rebalancing step will not cause an infinite loop where CLEANUP repeatedly finds the same violation, and TRYREBALANCE repeatedly returns without fixing it.

We prove both of the above claims for the case where TRYREBALANCE returns because of the NIL check at line 187. (The proof for the other case is similar.) Consider the node labeled $W5$ in Figure 6.9 and its parent. Moving from the parent to the node labeled $W5$ in the decision tree corresponds to following the right child pointer of the node r_{xxr} (where r_{xxr} is the right child of the node labeled r_{xx}). We prove that, in a chromatic tree, the right child pointer of r_{xxr} must be non-NIL. Suppose, to obtain a contradiction, that it is NIL (i.e., r_{xxr} is a leaf). Recall that the sum of node weights is always the same on all paths from the root to a leaf in a chromatic tree, and all node weights are non-negative. Let $SW(u)$ be the sum of weights from *root* to a node u . Since r_{xxr} is a leaf, it follows that $SW(r_{xxl}) \leq SW(r_{xxr})$ (and, if r_{xxl} is also a leaf, then $SW(r_{xxl}) = SW(r_{xxr})$). By definition, $SW(r_{xxl}) = SW(r_{xx}) + r_{xxl}.w$, and $SW(r_{xxr}) = SW(r_{xx}) + r_{xxr}.w$. Furthermore, in this case, $r_{xxr}.w = 1$, so $SW(r_{xxr}) = SW(r_{xx}) + 1$. However, r_{xxl} is overweight (indicated with shading in Figure 6.9), so $r_{xxl}.w > 1$, which implies $SW(r_{xxl}) > SW(r_{xxr})$, which is a contradiction. Consequently, if TRYREBALANCE sees that the right child pointer of r_{xxr} is NIL at line 187, then the view of the chromatic tree that it obtained from its invocations of LLX was not a consistent view of the chromatic tree. In other words, another update must have concurrently modified the tree. Thus, the configuration of nodes observed by TRYREBALANCE is transient, and, when CLEANUP repeats its search for violations, it will eventually see a different configuration (possibly after helping the other update to complete).

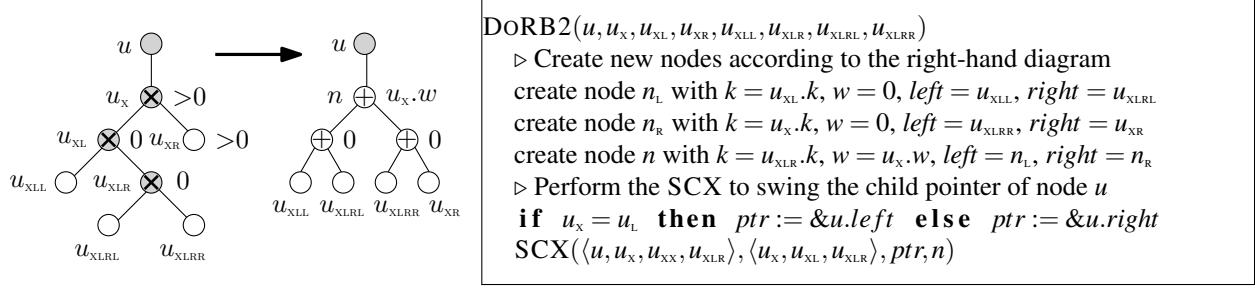


Figure 6.12: Implementing rebalancing step RB2. Other rebalancing steps are handled similarly using the diagrams shown in Figure 6.5.

6.2.5 Implementing a rebalancing step

We now describe how the individual rebalancing steps in Figure 6.5 are implemented. As an example, we consider one of the rebalancing steps, named RB2 (shown in Figure 6.12), which eliminates a red-red violation at node u_{xlr} . The other rebalancing steps are implemented similarly.

Pseudocode for RB2 appears in Figure 6.12. This code is invoked from TRYREBALANCE, after performing a sequence of LLXs and deciding which rebalancing step to perform. We first describe the steps taken in TRYREBALANCE before RB2 is invoked. TRYREBALANCE performs LLXs on each of the shaded nodes in Figure 6.12, starting with u . LLX is performed on u because it will be changed by the rebalancing step. LLX is performed on u_x , u_{xl} and u_{xlr} , because the rebalancing step will remove them from the tree. For RB2 to be applicable, u_x and u_{xr} must have positive weights and u_{xl} and u_{xlr} must both have weight 0. Since weight fields are immutable, TRYREBALANCE can verify constraints on weights at any time. One might wonder why LLX is not performed on u_{xr} , since RB2 should be applied only if u_x has a right child with positive weight. Suppose TRYREBALANCE simply verifies that $u_{xr}.w > 0$ without performing LLX on u_{xr} . Since weight fields are immutable, the only way that $u_{xr}.w$ can change is if the right child of u_x changes. Thus, if $u_{xr}.w$ changes after TRYREBALANCE performs LLX(u_x), then the right child of u_x will change, so the SCX will fail.

Now we describe the behaviour of RB2. First, n and its two children are created. N consists of these three nodes. The keys stored in the newly created nodes are the same as in the removed nodes (so that an in-order traversal encounters them in the same order). Finally, SCX(V, R, fld, new) is invoked, where fld is the child pointer of u that pointed to u_x in the result of LLX(u).

If the SCX modifies the tree, then no node $r \in V$ has changed since the update performed LLX(r). In this case, the SCX replaces the directed graph G_R by the directed graph G_N and the nodes in R are finalized. This ensures that other updates cannot erroneously modify these old nodes after they have been replaced. The nodes in the set $F_R = F_N = \{u_{xll}, u_{xrl}, u_{xlr}, u_{xr}\}$ each have the same keys, weights, and child pointers before and after the rebalancing step, so they can be reused. $V = \langle u, u_x, u_{xx}, u_{xlr} \rangle$ is the sequence of nodes on which TRYREBALANCE performs LLXs, and $R = \langle u_x, u_{xl}, u_{xlr} \rangle$ is a subsequence of V , so PC1, PC2 and PC3 are satisfied. Clearly, we satisfy PC8 and PC9 when we create new and its two children. It is easy to verify that PC4, PC5, PC6 and PC7 are satisfied. If the tree does not change during the update, then the nodes in V are ordered consistently with a breadth-first traversal of the tree, so PC10 is satisfied.

```

209 SUCCESSOR(key)
210   ▷ Returns the successor of key and its associated value (or  $\langle \perp, \perp \rangle$  if there is no such successor)
211   l := entry
212   loop until l is a leaf
213     if LLX(l) ∈ {FAIL, FINALIZED} then retry SUCCESSOR(key) from scratch
214     if key < l.key then
215       lastLeft := l
216       l := l.left
217       V :=  $\langle \text{lastLeft} \rangle$ 
218     else
219       l := l.right
220       add l to end of V
221     if lastLeft = entry then return  $\langle \perp, \perp \rangle$            ▷ Dictionary is empty
222     else if key < l.k then return  $\langle l.k, l.v \rangle$       ▷ Find next leaf after l in in-order traversal
223     else
224       succ := lastLeft.right
225       loop until succ is a leaf
226         if LLX(succ) ∈ {FAIL, FINALIZED} then retry SUCCESSOR(key) from scratch
227         add succ to end of V
228         succ := succ.left
229       if succ.key =  $\infty$  then result :=  $\langle \perp, \perp \rangle$  else result :=  $\langle \text{succ}.k, \text{succ}.v \rangle$ 
230       if VLX(V) then return result
231       else retry SUCCESSOR(key) from scratch

```

Figure 6.13: Code for SUCCESSOR.

6.2.6 Successor queries

SUCCESSOR(*key*) runs an ordinary BST search algorithm, using LLXs to read the child fields of each node visited, until it reaches a leaf. If the key of this leaf is larger than *key*, it is returned and the operation is linearized at any time during the operation when this leaf was on the search path for *key*. Otherwise, SUCCESSOR finds the next leaf. To do this, it remembers the last time it followed a left child pointer and, instead, follows one right child pointer, and then left child pointers until it reaches a leaf, using LLXs to read the child fields of each node visited. If any LLX it performs returns FAIL or FINALIZED, SUCCESSOR restarts. Otherwise, it performs a validate-extended (VLX), which returns TRUE only if all nodes on the path connecting the two leaves have not changed. If the VLX succeeds, the key of the second leaf found is returned and the query is linearized at the linearization point of the VLX. If the VLX fails, SUCCESSOR restarts.

6.3 Correctness proof

We start with a high-level correctness argument for the chromatic tree. As mentioned above, GET(*key*) invokes SEARCH(*key*), which traverses a path from *entry* to a leaf by reading child pointers. Even though this search can pass through nodes that have been removed by concurrent updates, we prove by induction that every node visited was on the search path for *key* at some time during the search. GET can thus be linearized at some time *t* when the leaf it

reaches was on the search path for key . (Moreover, we prove that this leaf is the only one in the tree that could possibly contain key .)

We prove that each insertion, deletion and rebalancing step that performs an SCX follows the template (and, hence, has an atomic update phase). Each `INSERT` that terminates performs a successful SCX, and we linearize the `INSERT` at this SCX. Each `DELETE` that terminates performs a successful SCX, or returns at line 66 or 69. If a `DELETE` performs a successful SCX, we linearize it at this SCX. Otherwise, the `DELETE` behaves like a query, and is linearized in the same way as `GET`. We argue that each invocations of `SEARCH` performed by `INSERT` or `DELETE` satisfies the delayed traversal property (DTP), which was introduced in Section 5.3. This allows us to invoke Theorem 5.7, which proves that `INSERT` and `DELETE` are atomic (including the search phase). Because no rebalancing step modifies the set of keys stored in leaves, the set of leaves always represents the set of dictionary entries.

Detailed proof We first start with two simple definitions, and then prove the major result of this section.

Definition 6.1 *The search path to key starting at a node r is the path that an ordinary BST search starting at r would follow. If $r = \text{entry}$, then we simply call this the search path to key.*

Note that this search is well-defined even if the data structure is not a BST. Moreover, the search path starting at a node is well-defined, even if the node has been removed from the tree. In any case, we simply look at the path that an ordinary BST search would follow, if it were performed on the data structure. Additionally, observe that the `SEARCH` procedure invoked by `TRYINSERT` and `TRYDELETE` is an ordinary BST search, and, hence, an atomic `SEARCH(key)` traverses exactly the search path to key .

Definition 6.2 *The range of a node u in some configuration is the set of keys for which u is on the search path.*

The next lemma is the main result in this section. It establishes several results which are used to prove that searches and updates are linearizable. The first result states that `TRYINSERT`, `TRYDELETE` and `TRYREBALANCE` follow the tree update template, which is a prerequisite to invoking the results in Section 5.4. Our proof that `TRYREBALANCE` follows the tree update template is complicated slightly by the fact that its subroutine, `OVERWEIGHTLEFT`, returns `FAIL` at line 175 and line 187 if it sees a `NIL` child pointer. (The `OVERWEIGHTRIGHT` routine is similar.) These return statements do not explicitly fit into the template, so, to guarantee non-blocking progress, we must prove that they are executed by an update only if it is concurrent with another successful update.

The second result states that the top of the tree is always as shown in Fig. 6.3.

The third and fourth claims are used to linearize searches. Specifically, since a search only reads child pointers, and the tree may change as the search traverses the tree, we must show that it still ends up at the correct leaf. In other words, we must show that the search is linearizable even if it traverses some nodes that are no longer in the tree. Note that these claims are proved in a somewhat similar way to [52], but the proofs here must deal with the additional complication of rebalancing operations occurring while a search traverses the tree.

The fifth claim states that the search phase of each insertion (resp., deletion) satisfies the delayed traversal property. This allows us to argue in the sixth claim that the entire insertion (resp., deletion) is atomic. The sixth claim also states that the update phase of each rebalancing step is atomic. (The search phase of a rebalancing step is not part of the atomic operation, but this is fine, because a chromatic tree allows rebalancing steps to be freely applied anywhere their preconditions are met, and in any order. Thus, it is *not* important for a rebalancing step to be performed specifically at the location that would be found by an atomic search.)

The final claim establishes the chromatic tree structure.

Lemma 6.3 *Our implementation of a chromatic tree satisfies the following claims.*

1. TRYINSERT and TRYDELETE follow the tree update template and satisfy all constraints specified by the template. If an invocation of TRYREBALANCE does not return at line 175 or line 187, then it follows the tree update template and satisfies all constraints specified by the template. Otherwise, it follows the tree update template up until it returns without performing an SCX, and it satisfies all constraints specified by the template.
2. The tree rooted at root always looks like Fig. 6.3(a) if it is empty, and Fig. 6.3(b) otherwise.
3. If a node v is in the data structure in some configuration C and v was on the search path for key k in some earlier configuration C' , then v is on the search path for k in C . (Equivalently, updates to the tree do not shrink the range of any node in the tree.)
4. If an invocation of SEARCH(k) reaches a node v , then there was some earlier configuration during the search when v was on the search path for k .
5. The SEARCH procedure used by TRYINSERT and TRYDELETE satisfies DTP.
6. Invocations of TRYINSERT, TRYDELETE and TRYREBALANCE that perform a successful SCX are atomic.
7. The tree rooted at the left child of entry is always a chromatic tree.

Proof: We prove these claims together by induction on the sequence of steps in an execution. That is, for each claim C , we assume that all of the claims hold before an arbitrary step in the execution, and prove that C holds after the step.

Claim 1: This claim follows almost immediately from inspection of the code. The only subtlety is showing that no process invokes LLX(r) where $r = \text{NIL}$. Suppose the inductive hypothesis holds just before an invocation of LLX(r).

For INSERTNEW, INSERTREPLACE and DELETE, $r \neq \text{NIL}$ follows from inspection of the code, inductive Claim 2, and the fact that every key inserted or deleted from the dictionary is less than ∞ (so every key inserted or deleted minimally has a parent and a grandparent).

For rebalancing steps, $r \neq \text{NIL}$ follows from inspection of the code and the decision tree in Fig. 6.9, using a few facts about the data structure. TRYREBALANCE performs LLXs on its arguments ggp, gp, p, l , and then possibly on a sequence of other nodes in the subtree rooted at gp , as it follows the decision tree. From Fig. 6.3(b), it is easy to see that any node with weight $w \neq 1$ minimally has a parent, grandparent, and great-grandparent. Thus, the arguments to TRYREBALANCE are all non-NIL. By Claim 7, each leaf has weight $w \geq 1$, every node has zero or two children, and the child pointers of a leaf do not change. This is enough to argue that all LLXs performed by TRYREBALANCE, and nearly all LLXs performed by OVERWEIGHTLEFT and OVERWEIGHTRIGHT, are passed non-NIL arguments. Without loss of generality, we restrict our attention to LLXs performed by OVERWEIGHTLEFT. The argument for OVERWEIGHTRIGHT is symmetric. The only LLXs that require different reasoning are performed at lines 177, 181, 189 and 192. For lines 177 and 189, the claim follows immediately from lines 175 and line 187, respectively. Consider line 181. If $r_{xxrl} = \text{NIL}$ then, since every node has zero or two children, and the child pointers of a leaf do not change, r_{xxrl} is a leaf, so $r_{xxrl} = \text{NIL}$. Therefore, OVERWEIGHTLEFT will return before it reaches line 181. By the same argument, $r_{xxrl} \neq \text{NIL}$ when line 192 is performed. Thus, $r \neq \text{NIL}$ no matter where the LLX occurs in the code.

Claim 2: The only step that can modify the tree (and, hence, affect this claim) is an invocation S of SCX performed by an invocation I of TRYINSERT, TRYDELETE or TRYREBALANCE. Suppose the inductive hypothesis holds just before S . By Claim 1, I follows the tree update template up until it performs S . By Lemma 5.5 and Lemma 5.1, the update phase of I is atomic. Thus, S atomically performs one of the transformations in Fig. 6.5. By inspection of

these transformations, when the tree is empty, `INSERTNEW` at the left child of *entry* changes the tree from looking like Fig. 6.3(a) to looking like Fig. 6.3(b) and, otherwise, does not affect the claim. When the tree has only one node with $key \neq \infty$, `DELETE` at the leftmost grandchild of *entry* changes the tree back to looking like Fig. 6.3(a) and, otherwise, does not affect the claim. `INSERTREPLACE` does not affect the claim.

Without loss of generality, suppose S performs a left rebalancing step. (The argument for right rebalancing steps is symmetric.) Each rebalancing step in $\{\text{BLK}, \text{RB1}, \text{RB2}\}$ applies only if $u_{\text{xl}}.w = 0$, and every other rebalancing step applies only if $u_{\text{xl}}.w > 1$. By the inductive hypothesis, just before S , all nodes that are not in the subtree rooted at the leftmost grandchild of *entry* have weight one. Therefore, S must change a child pointer in the subtree rooted at the leftmost grandchild of *entry*. Since the child pointer changed by S was traversed while a process was searching for a key that it inserted or deleted, and every such key must be less than ∞ , S can replace only nodes with $key < \infty$ (and, hence, cannot affect the claim).

Note: using similar reasoning, it is easy to verify that, each time a process accesses a field of a node u , $u \neq \text{NIL}$.

Claim 3: Initially, the claim is trivially true (since the tree only contains sentinel nodes, which are on every search path). In order for v to change from being on the search path for k in configuration C' to no longer being on the search path for k in configuration C , the tree must change between C' and C . Thus, there must be a successful SCX S between C' and C . Moreover, this is the only kind of step that can affect this claim. We show S preserves the property that v is on the search path for k .

By inductive Claim 1, S is performed by a template operation. Thus, by Lemma 5.1, S changes a pointer of a node from *old* to *new*, removing a connected set R of nodes (rooted at *old*) from the tree, and inserting a new connected set N of nodes. If v is not a descendant of *old* immediately before S , then this change cannot remove v from the search path for k . So, suppose v is a descendant of *old* immediately prior to S .

Since v is in the data structure in both C' and C , it must be in the data structure at all times between C' and C by Lemma 5.3. Therefore, v is a descendant of *old*, but S does not remove v from the tree. Recall that the fringe F_R is the set of nodes that are children of nodes in R , but are not themselves in R (see Figure 5.1 and Figure 5.2). By definition, v must be a descendant of a node $f \in F_R$. Moreover, since v is on the search path for k just before S , so is f . We argue, for each possible tree modification in Figure 6.5, that if any node in F_R is on the search path for k prior to S , then it is still on the search path for k after S . We proceed by cases.

Case 1: Suppose S is part of an invocation of `TRYINSERT`, which atomically implements the `INSERTNEW` and `INSERTREPLACE` modifications in Figure 6.5. Since S replaces a leaf with either a new leaf, or a new internal node and two new leaves, the fringe set is empty. Thus, the claim is vacuously true.

Case 2: Suppose S is part of an invocation of `TRYDELETE`, which atomically implements the `DELETE` modification in Figure 6.5. (The argument for its mirror image is symmetric.) S replaces a leaf u_{xl} , its sibling u_{xr} , and their parent u_x , with a new copy new of u_{xr} . If u_{xr} is a leaf, then F_R is empty, and the claim is vacuously true. Otherwise, F_R consists of the children u_{xrl} and u_{xrr} of u_{xr} . Since new has the same key as u_{xr} , the ranges of u_{xrl} and u_{xrr} after S are supersets of what they were before S . Thus, the claim holds.

Case 3: Suppose S is part of an invocation of `TRYREBALANCE`, which atomically implements the remaining modifications in Figure 6.5 (and their mirror images). In each rebalancing step, $|R| = |N|$ and the set of keys in R is the same as the set of keys in N . Furthermore, in-order traversals on $R \cup F_R$ just before S and $N \cup F_R$ just after S yields the same sequence of keys. By Claim 7, this sequence is sorted. Consequently, the nodes in R partition the range of *old* in exactly the same way that the nodes in N partition the range of *new*. Thus, the claim holds.

Claim 4: Consider a read r of a child pointer in an invocation I of $\text{SEARCH}(k)$. This is the only kind of step that can affect this claim. Let v be the node pointed to by the value returned by r . We prove that r preserves the claim. If r is the first read of a child pointer by I , then v is *entry*, which is always on the search path for k , so r preserves the claim.

Now, suppose there is a previous read r' of a child pointer by I . By the inductive hypothesis, the node v' that was returned by r' was on the search path for k in some configuration C' after the beginning of I and before r' . Our goal is to prove that there is a configuration C , after C' and before r , when v is on the search path for k . If v' is in the tree when I performs r , then C is the configuration just before r . Otherwise, C is the last configuration before v' was removed from the tree.

Without loss of generality, assume $k < v'.\text{key}$. (The argument when $k \geq v'.\text{key}$ is symmetric.) By inductive Claim 7, the data structure is a chromatic search tree just before r , so I must reach v by following the left child pointer of v' . We now prove that $v'.\text{left}$ points to v in C . Suppose v' is in the tree when I performs r (so C is just before r). This case follows immediately from our assumption that r reads a pointer to v from $v'.\text{left}$. Now, suppose v' is not in the tree when r occurs, so C is the last configuration before v' was removed. Recall that v' is in the tree in C' . By Lemma 5.3, v' cannot be added back into the data structure after it is removed. Since v' is in the tree in C' , and is not in the tree when I subsequently performs r , v' must be removed after C' and before r occurs. By inductive Claim 1 and template Constraint 3, v' becomes finalized precisely when it is removed. Since v' cannot change after it is finalized, and $v'.\text{left}$ points to v when r occurs (which is after v' is removed), we can see that $v'.\text{left}$ must point to v in C (which is just before v' is removed).

Finally, we prove that v is on the search path for k in C . Since v' was on the search path for k in C' , and it is in the data structure in C , which is after C' but before r , inductive Claim 3 implies that v' is on the search path for k in C . Since $k < v'.\text{key}$ and $v = v'.\text{left}$, v is also on the search path for k in C .

Claim 5: Initially, the claim holds vacuously (since no steps have been taken). Suppose an invocation S of $\text{SEARCH}(k)$ in TRYDELETE terminates and returns m . (The proof for TRYINSERT is similar.) Consider any configuration C , after S returns m , in which all of the nodes in m are in the tree and their fields agree with the values in m . Suppose the inductive hypothesis up until C . We prove that an invocation S' of $\text{SEARCH}(k)$ in TRYDELETE would return m if S' were performed atomically just after configuration C .

The value $m = \langle -, gp, p, l \rangle$ returned by S contains a leaf l , its parent p and its grandparent gp . By inductive Claim 4, gp , p and l were each on the search path at some point during S (which is before C). Since gp , p and l are in the tree in C , inductive Claim 3 implies that they are all on the search path for k in C . Therefore, S' will visit each of them. Conceptually, m also encodes the facts that gp points to p and p points to l . These facts are checked at line 74 and line 79 (as part of the CONFLICT procedure). By our assumption (that the fields of the nodes in m in configuration C agree with their values in m), these facts also hold when S' is performed. Consequently, S' will also return $m = \langle -, gp, p, l \rangle$.

Claim 6: By Claim 5 and Theorem 5.7, all invocations of TRYINSERT and TRYDELETE that perform a successful SCX are atomic. By Lemma 5.5, all invocations of TRYREBALANCE that perform a successful SCX are atomic.

Claim 7: Only successful invocations of SCX can affect this claim. These are performed only in TRYINSERT , TRYDELETE and TRYREBALANCE . The claim holds in the initial state of the tree shown in Figure 6.3(a). We show that every successful invocation S of SCX preserves the claim. We proceed by cases.

Case 1: S is performed in an invocation I of $\text{TRYINSERT}(key, value)$ or $\text{TRYDELETE}(key)$. By Claim 6, the entirety of I is atomic, including its search phase. Thus, when I occurs, its search procedure returns the unique leaf on the

search path for key . Consequently, I atomically performs one of the transformations INSERTNEW, INSERTREPLACE or DELETE in Figure 6.5 to replace l (and possibly some of its neighbouring nodes). Since I is entirely atomic (including its search phase), and it simply performs one of the chromatic tree updates, it is easy to verify that it preserves the claim.

Case 2: S is performed in an invocation I of TRYREBALANCE, which performs one of the rebalancing transformations PUSH, RB1, RB2, BLK, W1, W2, W3, W4, W5, W6, or W7. By Claim 6, the update phase of I is atomic (but the search phase is not necessarily atomic). Therefore, I atomically performs one of the rebalancing transformations at some location in the tree (but not necessarily the same rebalancing transformation, at the same location, that it would perform if I 's search were also part of the atomic update). All of the rebalancing transformations preserve the claim (regardless of where in the tree they are performed). ■

We define the linearization points for chromatic tree operations as follows.

- $\text{GET}(key)$ is linearized at a time during the operation when the leaf reached was on the search path for key . (This time exists, by Lemma 6.3.4.)
- An INSERT is linearized at its successful SCX inside TRYINSERT (if such an SCX exists). (Note: every INSERT that terminates performs a successful SCX.)
- A DELETE that returns \perp is linearized at a time during the operation when the leaf returned by its last invocation of SEARCH was on the search path for key . (This time exists, by Lemma 6.3.4.)
- A DELETE that does not return \perp is linearized at its successful SCX inside TRYDELETE (if such an SCX exists). (Note: every DELETE that terminates, but does not return \perp , performs a successful SCX.)
- A SUCCESSOR query that returns at line 221 is linearized when it performs $\text{LLX}(entry)$.
- A SUCCESSOR query that returns at line 222 at the time during the operation that l was on the search path for key . (This time exists, by Lemma 6.3.4.)
- A SUCCESSOR query that returns at line 230 is linearized when it performs its successful VLX .

It is easy to verify that every operation that terminates is assigned a linearization point during the operation.

Theorem 6.4 *The chromatic search tree is a linearizable implementation of an ordered dictionary with the operations GET , INSERT , DELETE , SUCCESSOR .*

Proof: Lemma 6.3.6 proves that the SCXs implement atomic changes to the tree as shown in Figure 6.5. By inspection of these transformations, the set of keys and associated values stored in leaves are not altered by any rebalancing steps. Moreover, the transformations performed by each linearized INSERT and DELETE maintain the invariant that the set of keys and associated values stored in leaves of the tree is exactly the set that should be in the dictionary.

When a $\text{GET}(key)$ is linearized, the search path for key ends at the leaf returned by its invocation of SEARCH . If that leaf contains key , GET returns the associated value, which is correct. If that leaf does not contain key , then, by Lemma 6.3.7, it is nowhere else in the tree, so GET is correct to return \perp .

If $\text{SUCCESSOR}(key)$ returns $\langle \perp, \perp \rangle$ at line 221, then at its linearization point, the left child of the $entry$ is a leaf. By Lemma 6.3.2, the dictionary is empty.

If $\text{SUCCESSOR}(key)$ returns $\langle l.k, l.v \rangle$ at line 221, then at its linearization point, l is the leaf on the search path for key . So, l contains either key or its predecessor or successor at the linearization point. Since $key < l.k$, l is key 's successor.

Finally, suppose $\text{SUCCESSOR}(key)$ returns $\langle succ.k, succ.v \rangle$ at line 230. Then l was on the search path for key at some time during the search. Since l is among the nodes validated by the VLX , it is not finalized, so it is still on the

search path for key at the linearization point, by Lemma 6.3.3. Since $key \geq l.k$, the successor of key is the next leaf after l in an in-order traversal of the tree. Leaf l is the rightmost leaf in the subtree rooted at the left child of $lastLeft$ and the key returned is the leftmost leaf in the subtree rooted at the right child of $lastLeft$. The paths from $lastLeft$ to these two leaves are not finalized and therefore are in the tree. Thus, the correct result is returned. ■

6.4 Progress proof

Our goal is to prove that, if processes take steps infinitely often, then chromatic tree operations succeed infinitely often. At a high level, this follows from Theorem 5.10 (the final progress result for template operations), and the fact that at most $3i + d$ rebalancing steps can be performed after i insertions and d deletions have occurred (proved in [22]). Theorem 5.10 applies only if processes perform infinitely many template operations, so we must prove that processes will perform infinitely many template operations if they take steps infinitely often. The subtlety is that some invocations of TRYREBALANCE might not follow the template. One can imagine a pathology in which non-blocking progress is violated, because processes perform only finitely many invocations of TRYREBALANCE that follow the template, but perform infinitely many invocations that do *not* follow the template. We first prove that this does not happen. Then, we prove the main result.

Lemma 6.5 *If infinitely many invocations of TRYREBALANCE are performed, then infinitely many invocations follow the tree update template.*

Proof: As we explained near the beginning of this section, each invocation either follows the template or returns at line 175 or line 187 without invoking SCX. We prove that each invocation I of TRYREBALANCE that returns at line 175 or line 187 is concurrent with a template operation that performs a successful SCX during I (and, hence, follows the template). Suppose not, to derive a contradiction. Suppose I returns at line 175. Consider the configuration immediately before I returns. By inspection of OVERWEIGHTLEFT, we have $r_{xxrl}.w = 1$ and $r_{xxrlr} = \text{NIL}$, so r_{xxrl} is a leaf. Moreover, since the tree does not change during I , we also have the following facts: r_{xxr} is the parent of r_{xxrl} , $r_{xxr}.w = 0$, r_{xxl} is the sibling of r_{xxr} , and $r_{xxl}.w > 1$. Therefore, the sum of weights on a path from $root$ to a leaf in the sub-tree rooted at r_{xxr} is different from the sum of weights on a path from $root$ to a leaf in the sub-tree rooted at r_{xxl} . So, the tree is not a chromatic tree, which is a contradiction. The proof when I returns at line 187 is similar, and is left as an exercise. ■

Theorem 6.6 *The chromatic tree operations are non-blocking.*

Proof: To derive a contradiction, suppose there is some configuration C after which some processes continue to take steps but no successful chromatic tree operations occur. We first argue that eventually the tree stops changing. Since no successful chromatic tree operations occur after C , the only steps that can change the tree after C are successful invocations of SCX performed by TRYREBALANCE. Boyar, Fagerberg and Larsen [22] proved that after a bounded number of rebalancing steps, the tree becomes a red-black tree, and then no further rebalancing steps can be applied. Thus, eventually the tree must stop changing.

This implies that every invocation of SEARCH (and, hence, GET) terminates after a finite number of steps, unless the process executing it crashes. Thus, SEARCH and GET are non-blocking. It follows that no invocation of SEARCH or GET occurs after C . To prove progress for the other operations, we consider two cases.

Suppose processes take infinitely many steps in INSERT or DELETE operations. Then, processes perform infinitely many invocations of TRYINSERT, TRYDELETE or TRYREBALANCE. By Lemma 6.3.1 and Lemma 6.5, infinitely many of these invocations follow the tree update template. By Theorem 5.10, infinitely many of them will succeed, so infinitely many must succeed after C , which is a contradiction.

Now, suppose there is a configuration C' (after C) after which no process takes a step in an INSERT or DELETE operation. Then, every operation after C' must be an invocation of PREDECESSOR or SUCCESSOR. Observe that every invocation of PREDECESSOR or SUCCESSOR is an execution of a VLX-QUERY algorithm (see the definition in Section 5.5). By progress property P2 of LLX, SCX and VLX (see Section 3.2.2), infinitely many invocations of SCX or VLX must succeed. Since there are only finitely many invocations of SCX, infinitely many invocations of VLX must succeed, and infinitely many of these must occur after C' . Observe that VLX is performed only by invocations of PREDECESSOR and SUCCESSOR, and an invocation of PREDECESSOR or SUCCESSOR is successful precisely if it performs a successful VLX. Therefore, infinitely many successful invocations of PREDECESSOR and/or SUCCESSOR must occur after C' , which is a contradiction. ■

6.5 Bounding the height of the tree

We now show that the height of the chromatic search tree at any time is $O(c + \log n)$ where n is the number of keys stored in the tree and c is the number of INSERT and DELETE operations currently in progress. Since we always perform rebalancing steps that satisfy VIOL, if we reach a leaf without finding the violation that an INSERT or DELETE created, then the violation has been eliminated. This allows us to prove that the number of violations in the tree at any time is bounded above by c . Further, since removing all violations would yield a red-black tree with height $O(\log n)$, and eliminating each violation reduces the height by at most one, the height of the chromatic tree is $O(c + \log n)$.

Definition 6.7 Let x be a node that is in the data structure. We say that $x.w - 1$ **overweight violations occur at x** if $x.w > 1$. We say that a **red-red violation occurs at x** if x and its parent in the data structure both have weight 0.

The following lemma says that red-red violations can never be created at a node, except when the node is first added to the data structure.

Lemma 6.8 Let v be a node with weight 0. Suppose that when v is added to the data structure, its (unique) parent has non-zero weight. Then v is never the child of a node with weight 0.

Proof: Node weights are immutable. It is easy to check by inspection of each transformation in Figure 6.5 that if v is not a newly created node and it acquires a new parent in the transformation with weight 0, then v had a parent of weight 0 prior to the transformation. ■

Definition 6.9 A process P is **in a cleanup phase for k** if it is executing an $\text{INSERT}(k)$ or a $\text{DELETE}(k)$ and it has performed a successful SCX inside a TRYINSERT or TRYDELETE that returns $\text{createdViolation} = \text{TRUE}$. If P is between line 102 and 109, $\text{location}(P)$ and $\text{parent}(P)$ are the values of P 's local variables l and p ; otherwise $\text{location}(P)$ is entry and $\text{parent}(P)$ is NIL.

We use the following invariant to show that each violation in the data structure has a pending update operation that is responsible for removing it before terminating: either that process is on the way towards the violation, or it will find another violation and restart from the top of the tree, heading towards the violation.

Lemma 6.10 In every configuration, there exists an injective mapping ρ from violations to processes such that, for every violation x ,

- (A) process $\rho(x)$ is in a cleanup phase for some key k_x and
- (B) x is on the search path from entry for k_x and
- (C) either
 - (C1) the search path for k_x from $location(\rho(x))$ contains the violation x , or
 - (C2) $location(\rho(x)).w = 0$ and $parent(\rho(x)).w = 0$, or
 - (C3) in the prefix of the search path for k_x from $location(\rho(x))$ up to and including the first non-finalized node (or the entire search path if all nodes are finalized), there is a node with weight greater than 1 or two nodes in a row with weight 0.

Proof: In the initial configuration, there are no violations, so the invariant is trivially satisfied. We show that any step S by any process P preserves the invariant. We assume there is a function ρ satisfying the claim for the configuration C immediately before S and show that there is a function ρ' satisfying the claim for the configuration C' immediately after S . The only step that can cause a process to leave its cleanup phase is the termination of an INSERT or DELETE that is in its cleanup phase. The only steps that can change $location(P)$ and $parent(P)$ are P 's execution of line 109 or the read of the child pointer on line 105 or 106. (We think of all of the updates to local variables in the braces on those lines as happening atomically with the read of the child pointer.) The only steps that can change child pointers or finalize nodes are successful SCXs. No other steps S can cause the invariant to become false.

Case 1 S is the termination of an INSERT or DELETE that is in its cleanup phase: We choose $\rho' = \rho$. S happens when the test in line 104 is true, meaning that $location(P)$ is a leaf. Leaves always have weight greater than 0. The weight of the leaf cannot be greater than 1, because then the process would have exited the loop in the previous iteration after the test at line 107 returned true (since weights of nodes never change). Thus, $location(P)$ is a leaf with weight 1. So, P cannot be $\rho(x)$ for any violation x , so S cannot make the invariant become false.

Case 2 S is an execution of line 109: We choose $\rho' = \rho$. Step S changes $location(P)$ to $entry$. If $P \neq \rho(x)$ for any violation x , then this step cannot affect the truth of the invariant. Now suppose $P = \rho(x_0)$ for some violation x_0 . The truth of properties (A) and (B) are not affected by a change in $location(P)$ and property (C) is not affected for any violation $x \neq x_0$. Since ρ satisfies property (B) for violation x_0 before S , it will satisfy property (C1) for x_0 after S .

Case 3 S is a read of the left child pointer on line 105: We choose $\rho' = \rho$. Step S changes $location(P)$ from some node v to node v_L , which is v 's left child when S is performed. If $P \neq \rho(x)$ for any violation x , then this step cannot affect the truth of the invariant. So, suppose $P = \rho(x_0)$ for some violation x_0 . By (A), P is in a cleanup phase for k_{x_0} . The truth of (A) and (B) are not affected by a change in $location(P)$ and property (C) is not affected for any violation $x \neq x_0$. So it remains to prove that (C) is true for violation x_0 in C' .

First, we prove $v.w \leq 1$, and hence there is never an overweight violation at v . If v is $entry$, then $v.w = 1$. Otherwise, S does not occur during the first iteration of CLEANUP's inner loop. In the previous iteration, $v.w \leq 1$ at line 107 (otherwise, the loop would have terminated).

Next, we prove that there is no red-red violation at v when S occurs. If v is $entry$ or is not in the data structure when S occurs, then there cannot be a red-red violation at v when S occurs, by definition. Otherwise, node v was read as the child of some other node u in the previous iteration of CLEANUP's inner loop and line 107 found that $u.w \neq 0$ or $v.w \neq 0$ (otherwise the loop would have terminated). So, at some time before S (and when v was in the data structure), there was no red-red violation at v . By Lemma 6.8, there is no red-red violation at v when S is performed.

Next, we prove that (C2) cannot be true for x_0 in configuration C . If S is in the first iteration of CLEANUP's inner loop, then $\text{location}(P) = \text{entry}$, which has weight 1. If S is not in the first iteration of CLEANUP's inner loop, then the previous iteration found $\text{parent}(P).w \neq 0$ or $\text{location}(P).w \neq 0$ (otherwise the loop would have terminated).

So we consider two cases, depending on whether (C1) or (C3) is true in configuration C .

Case 3a (C1) is true in configuration C : Thus, when S is performed, the violation x_0 is on the search path for k_{x_0} from v , but it is not at v (as argued above). S reads the *left* child of v , so $k_{x_0} < v.k$ (since the key of node v never changes). So, x_0 must be on the search path for k_{x_0} from v_L . This means (C1) is satisfied for x_0 in configuration C' .

Case 3b (C3) is true in configuration C : We argued above that $v.w \leq 1$, so the prefix must contain two nodes in a row with weight 0. If they are the first two nodes, v and v_L , then (C2) is true after S . Otherwise, (C3) is still true after S .

Case 4 S is a read of the right child pointer on line 106: The argument is symmetric to Case 3.

Case 5 S is a successful SCX: We must define the mapping ρ' for each violation x in configuration C' . By Lemma 6.8 and the fact that node weights are immutable, no transformation in Figure 6.5 can create a new violation at a node that was already in the data structure in configuration C . So, if x is at a node that was in the data structure in configuration C , x was a violation in configuration C , and $\rho(x)$ is well-defined. In this case, we let $\rho'(x) = \rho(x)$.

If x is at a node that was added to the data structure by S , then we must define $\rho(x)$ on a case-by-case basis for all transformations described in Figure 6.5. (The symmetric operations are handled symmetrically.)

Transformation	Red-red violations x created by S	$\rho'(x)$
RB1	none created	–
RB2	none created	–
BLK	at n (if $u_x.w = 1$ and $u.w = 0$)	$\rho(\text{red-red violation at one of } u_{\text{xLL}}, u_{\text{xLR}}, u_{\text{xRL}}, u_{\text{xRR}})$ †
PUSH	none created	–
W1,W2,W3,W4	none created	–
W5	at n (if $u_x.w = u.w = 0$)	$\rho(\text{red-red violation at } u_x)$
W6	at n (if $u_x.w = u.w = 0$)	$\rho(\text{red-red violation at } u_x)$
W7	none created	–
INSERT1	at n (if $u_x.w = 1$ and $u.w = 0$)	process performing the INSERT
INSERT2	none created	–
DELETE	at n (if $u_x.w = u_{\text{xR}}.w = u.w = 0$)	$\rho(\text{red-red violation at } u_x)$

Figure 6.14: Description of how ρ' maps red-red violations at newly added nodes to processes responsible for them.

†By inspection of the decision tree in Figure 6.9, BLK is only applied if one of u_{xLL} , u_{xLR} , u_{xRL} or u_{xRR} has weight 0, and therefore a red-red violation, in configuration C , and this red-red violation is eliminated by the transformation.

If x is a red-red violation at a newly added node, we define $\rho'(x)$ according to the table in Figure 6.14. For each newly added node that has k overweight violations after S , ρ' maps them to the k distinct processes $\{\rho(q) : q \in Q\}$, where Q is given by the table in Figure 6.15.

The function ρ' is injective, since ρ' maps each violation created by S to a distinct process that ρ assigned to a violation that has been removed by S , with only two exceptions: for red-red violations caused by INSERTNEW and one overweight violation caused by DELETE, ρ' maps the red-red violation to the process that has just begun its cleanup phase (and therefore was not assigned any violation by ρ).

Let x be any violation in the tree in configuration C' . We show that ρ' satisfies properties (A), (B) and (C) for x in configuration C' .

Transformation	Overweight violations created by S	Set Q of overweight violations before S
RB1	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
RB2	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
BLK	$u_x.w - 2$ at n (if $u_x.w > 2$)	$u_x.w - 2$ of the $u_x.w - 1$ at u_x
PUSH	$u_x.w$ at n (if $u_x.w > 0$)	$u_x.w - 1$ at u_x , and 1 at u_{xl}
PUSH	$u_{xl}.w - 2$ at n_L (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W1	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
W1	$u_{xl}.w - 2$ at n_{LL} (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W1	$u_{xrl}.w - 2$ at n_{LR} (if $u_{xrl}.w > 2$)	$u_{xrl}.w - 2$ of the $u_{xrl}.w - 1$ at u_{xl}
W2	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
W2	$u_{xl}.w - 2$ at n_{LL} (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W3	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
W3	$u_{xl}.w - 2$ at n_{LLL} (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W4	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
W4	$u_{xl}.w - 2$ at n_{LL} (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W5	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
W5	$u_{xl}.w - 2$ at n_{LL} (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W6	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
W6	$u_{xl}.w - 2$ at n_{LL} (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W7	$u_x.w$ at n (if $u_x.w > 0$)	$u_x.w - 1$ at u_x , and 1 at u_{xl}
W7	$u_{xl}.w - 2$ at n_L (if $u_{xl}.w > 2$)	$u_{xl}.w - 2$ of the $u_{xl}.w - 1$ at u_{xl}
W7	$u_{xrl}.w - 2$ at n_R (if $u_{xrl}.w > 2$)	$u_{xrl}.w - 2$ of the $u_{xrl}.w - 1$ at u_{xrl}
INSERT1	$u_x.w - 2$ at n (if $u_x.w > 2$)	$u_x.w - 2$ of the $u_x.w - 1$ at u_x
INSERT2	$u_x.w - 1$ at n (if $u_x.w > 1$)	$u_x.w - 1$ at u_x
DELETE	$u_x.w + u_{xrl}.w - 1$ at n (if $u_x.w + u_{xrl}.w > 1$)	$\max(0, u_x.w - 1)$ at u_x and $\max(0, u_{xrl}.w - 1)$ at $u_{xrl} \dagger$

Figure 6.15: Description of how ρ' maps overweight violations at newly added nodes to processes responsible for them. (Note that Q is a set, since a node can have many overweight violations.) \dagger In this case, the number of violations in Q is one too small if both $u_x.w$ and $u_{xrl}.w$ are greater than 0, so the remaining violation is assigned to the process that performed the DELETE's SCX.

Property (A): Every process in the image of ρ' was either in the image of ρ or a process that just entered its cleanup phase at step S , so every process in the image of ρ' is in its cleanup phase.

Property (B) and (C): We consider several subcases.

Subcase 5a Suppose S is an INSERTNEW's SCX, and x is the red-red violation created by S . Then, P is in its cleanup phase for the inserted key, which is one of the children of the node containing the red-red violation x . Since the tree is a BST, x is on the search path for this key, so (B) holds.

In this subcase, $location(\rho'(x)) = entry$ since $P = \rho'(x)$ has just entered its cleanup phase. So property (B) implies property (C1).

Subcase 5b Suppose S is a DELETE's SCX, and x is the overweight violation assigned to P by ρ' . Then, P is in a cleanup phase for the deleted key, which was in one of the children of u_x before S . Therefore, x (at the root of the newly inserted subtree) is on the search path for this key, so (B) holds.

As in the previous subcase, $location(\rho'(x)) = entry$ since $P = \rho'(x)$ has just entered its cleanup phase. So property (B) implies property (C1).

Subcase 5c If x is at a node that was added to the data structure by S (and is not covered by the above two cases),

then $\rho'(x)$ is $\rho(y)$ for some violation y that has been removed from the tree by S , as described in the above two tables. Let k be the key such that process $\rho(y) = \rho'(x)$ is in the cleanup phase for k . By property (B), y was on the search path for k before S . It is easy to check by inspection of the tables and Figure 6.5 that any search path that went through y 's node in configuration C goes through x 's node in configuration C' . (We designed the tables to have this property.) Thus, since y was on the search path for k in configuration C , x is on the search path for k in configuration C' , satisfying property (B).

If (C2) is true for violation y in configuration C , then (C2) is true for x in configuration C' (since S does not affect $location()$ or $parent()$ and $\rho(y) = \rho'(x)$). If (C3) is true for violation y in configuration C , then (C3) is true for x in configuration C' (since any node that is finalized remains finalized forever, and its child pointers do not change).

So, for the remainder of the proof of subcase 5c, suppose (C1) is true for y in configuration C . Let $l = location(\rho(y))$ in configuration C . Then y is on the search path for k from l in configuration C .

First, suppose S removes l from the data structure.

- If y is a red-red violation at node l in configuration C , then the red-red violation was already there when process $\rho(y)$ read l as the child of some other node (by Lemma 6.8) and (C2) is true for x in configuration C' .
- If y is an overweight violation at node l in configuration C , then it makes (C3) true for x in configuration C' .
- Otherwise, since both l and its descendant, the parent of the node that contains y , are removed by S , the entire path between these two nodes is removed from the data structure by S . So, all nodes along this path are finalized by S because Constraint 3 is satisfied. Thus, the violation y makes (C3) true for x in configuration C' .

Now, suppose S does not remove l from the data structure. In configuration C , the search path from l for k contains y . It is easy to check by inspection of the tables defining ρ' and Figure 6.5 that any search path from l that went through y 's node in configuration C goes through x 's node in configuration C' . So, (C1) is true in configuration C' .

Subcase 5d If x is at a node that was in the data structure in configuration C , then $\rho'(x) = \rho(x)$. Let k be the key such that this process is in the cleanup phase for k . Since x was on the search path for k in configuration C and S did not remove x from the data structure, x is still on the search path for k in configuration C' (by inspection of Figure 6.5). This establishes property (B).

If (C2) or (C3) is true for x in configuration C , then it is also true for x in configuration C' , for the same reason as in Subcase 5c.

So, suppose (C1) is true for x in configuration C . Let $l = location(\rho(x))$ in configuration C . Then, (C1) says that x is on the search path for k from l in configuration C . If S does not change any of the child pointers on this path between l and x , then x is still on the search path from $location(\rho'(x)) = l$ in configuration C' , so property (C1) holds for x in C' . So, suppose S does change the child pointer of some node on this path from old to new . Then the search path from l for k in configuration C goes through old to some node f in the Fringe set F of S and then onward to the node containing violation x . By inspection of the transformations in Figure 6.5, the search path for k from l in configuration C' goes through new to the same node f , and then onward to the node containing the violation x . Thus, property (C1) is true for x in configuration C' . ■

Corollary 6.11 *The number of violations in the data structure is bounded by the number of incomplete INSERT and DELETE operations.*

In the following discussion, we are discussing “pure” chromatic trees, without the dummy nodes with key ∞ that appear at the top of our tree. The sum of weights on a path from the root to a leaf of a chromatic tree is called the *path weight* of that leaf. The *height* of a node v , denoted $h(v)$ is the maximum number of nodes on a path from v to a leaf

descendant of v . We also define the *weighted height* of a node v as follows.

$$wh(v) = \begin{cases} v.w & \text{if } v \text{ is a leaf} \\ \max(wh(v.left), wh(v.right)) + v.w & \text{otherwise} \end{cases}$$

Lemma 6.12 Consider a chromatic tree rooted at root that contains n nodes and c violations. Suppose T is any red black tree rooted at root_T that results from performing a sequence of rebalancing steps on the tree rooted at root to eliminate all violations. Then, the following claims hold.

1. $h(\text{root}) \leq 2wh(\text{root}) + c$

2. $wh(\text{root}) \leq wh(\text{root}_T) + c$

3. $wh(\text{root}_T) \leq h(\text{root}_T)$

Proof: **Claim 1:** Consider any path from root to a leaf. It has at most $wh(\text{root})$ non-red nodes. So, there can be at most $wh(\text{root})$ red nodes that do not have red parents on the path (since root has weight 1). There are at most c red nodes on the path that have red parents. So the total number of nodes on the path is at most $2wh(\text{root}) + c$.

Claim 2: Consider any rebalancing step that is performed by replacing some node u_x by n (using the notation of Figure 6.5). If u_x is not the root of the chromatic tree, then $wh(u_x) = wh(n)$, since the path weights of all leaves in a chromatic tree must be equal. (Otherwise, the path weight to a leaf in the subtree rooted at n would become different from the path weight to a leaf outside this subtree.)

Thus, the only rebalancing steps that can change the weighted height of the root are those where u_x is the root of the tree. Recall that the weight of the root is always one. If u_x and n are supposed to have different weights according to Figure 6.5, then blindly setting the weight of the n to one will have the effect of changing the weighted height of the root. By inspection of Figure 6.5, the only transformation that increases the weighted height of the root is BLK, because it is the only transformation where the weight of n is supposed to be less than the weight of u_x . Thus, each application of BLK at the root increases the weighted height of the root by one, but also eliminates at least one red-red violation at a grandchild of the root (without introducing any new violations). Since none of the rebalancing steps increases the number of violations in the tree, performing any sequence of steps that eliminates c violations will change the weighted height of the root by at most c . The claim then follows from the fact that T is produced by eliminating c violations from the chromatic tree rooted at root .

Claim 3: Since T is a RBT, it contains no overweight violations. Thus, the weighted height of the tree is a sum of zeros and ones. It follows that $wh(\text{root}_T) \leq h(\text{root}_T)$. ■

Corollary 6.13 If there are c incomplete INSERT and DELETE operations and the data structure contains n keys, then its height is $O(\log n + c)$.

Proof: Let root , T , and root_T be defined as in Lemma 6.12. We immediately obtain $h(\text{root}) \leq 2h(\text{root}_T) + 3c$ from Corollary 6.11 and Lemma 6.12. Since the height of a RBT is $O(\log n)$, it follows that the height of our data structure is $O(\log n + c)$ (including the two dummy nodes at the top of the tree with key ∞). ■

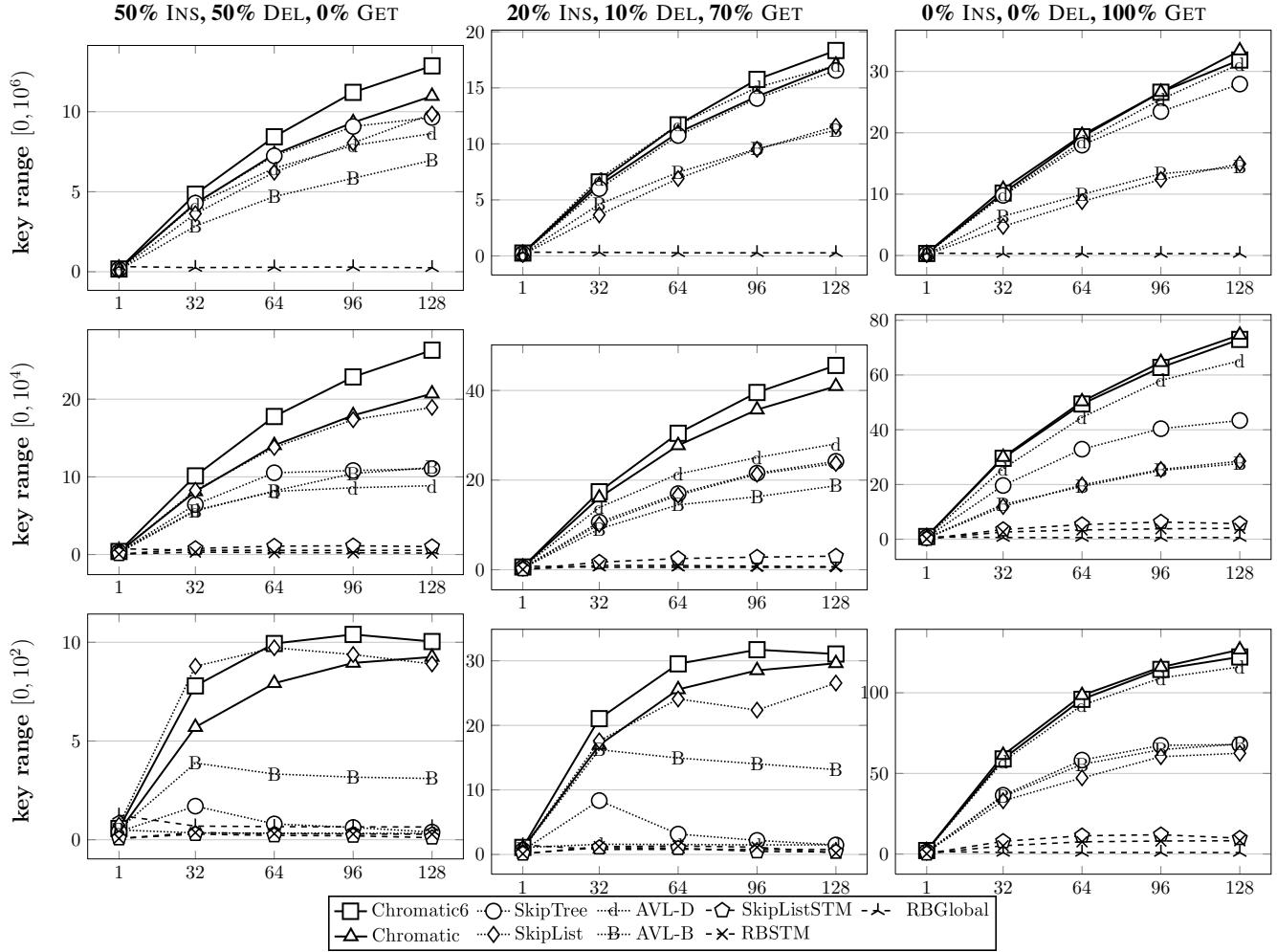


Figure 6.16: *Multithreaded* throughput (millions of operations/second) for 2-socket SPARC T2+ (128 hardware threads) on y-axis versus number of threads on x-axis.

6.6 Allowing more violations

Forcing insertions and deletions to rebalance the chromatic tree after creating only a single violation can cause unnecessary rebalancing steps to be performed, for example, because an overweight violation created by a deletion might be eliminated by a subsequent insertion. In practice, we can reduce the total number of rebalancing steps that occur by modifying our `INSERT` and `DELETE` procedures so that `CLEANUP` is invoked only once the number of violations on a path from *entry* to a leaf exceeds some constant k . The resulting data structure has height $O(k + c + \log n)$. Since searches are significantly faster than updates, slightly increasing search costs to reduce update costs yields performance benefits for many workloads.

6.7 Experimental results

We compared the performance of our chromatic tree (Chromatic), and the variant of our chromatic tree that invokes CLEANUP only when the number of violations on a path exceeds six (Chromatic6), against several leading data structures that implement ordered dictionaries: the non-blocking skip-list (SkipList) of the Java Class Library, the non-blocking multiway search tree (SkipTree) of Spiegel and Reynolds [115], the lock-based relaxed-balance AVL tree with non-blocking searches (AVL-D) of Drachsler et al. [49], and the lock-based relaxed-balance AVL tree (AVL-B) of Bronson et al. [27]. Our comparison also includes an STM-based red-black tree optimized by Oracle engineers (RBSTM) [67], an STM-based skip-list (SkipListSTM), and the highly optimized Java red-black tree, `java.util.TreeMap`, with operations protected by a global lock (RBGlobal). The STM data structures are implemented using DeuceSTM 1.3.0, which is one of the fastest STM implementations that does not require modifications to the Java virtual machine. We used the offline instrumentation capability of DeuceSTM to minimize the overhead of the STM instrumentation. All of the implementations used were made publicly available by their respective authors. For a fair comparison between data structures, we made slight modifications to RBSTM and SkipListSTM to use generics, instead of hardcoding the type of keys as `int`, and to store values in addition to keys.

We tested the data structures for three different operation mixes, 0i-0d, 20i-10d and 50i-50d, where x % INSERTS, y % DELETES, and $(100 - x - y)\%$ GETS. These operation mixes represent workloads where: all operations are queries, a moderate proportion operations are INSERTS and DELETES, and all operations are INSERTS and DELETES. We used three key ranges, $[0, 10^2]$, $[0, 10^4]$ and $[0, 10^6]$, to test different contention levels. For example, for key range $[0, 10^2]$, data structures will be small, so updates are likely to affect overlapping parts of the data structure.

For each data structure, each operation mix, each key range, and each thread count in $\{1, 32, 64, 96, 128\}$, we ran five trials which each measured the total throughput (operations per second) of all threads for five seconds. Each trial began with an untimed prefilling phase, which continued until the data structure was within 5% of its expected size in the steady state. For operation mix 50i-50d, the expected size is half of the key range. This is because, eventually, each key in the key range has been inserted or deleted at least once, and the last operation on any key is equally likely to be an insertion (in which case it is in the data structure) or a deletion (in which case it is not in the data structure). Similarly, 20i-10d yields an expected size of two thirds of the key range since, eventually, each key has been inserted or deleted and the last operation on a particular key is twice as likely to be an insertion as a deletion. For 0i-0d, we prefilled the data structures so that they contain half of the key range.

We used a Sun Enterprise T5240 with 32GB of RAM and two UltraSPARC T2+ processors, for a total of $16 \times 1.2\text{GHz}$ cores supporting a total of 128 hardware threads. The Sun 64-bit JVM version 1.7.0_03 was run in server mode, with 3GB minimum and maximum heap sizes. Different experiments run within a single instance of a Java virtual machine (JVM) are not statistically independent, so each batch of five trials was run in its own JVM instance. Prior to running each batch, a fixed set of three trials was run to cause the Java HotSpot compiler to optimize the running code. Garbage collection was manually triggered before each trial. The heap size of 3GB was small enough that garbage collection was performed regularly (approximately ten times) in each trial. We did not pin threads to cores, since this is unlikely to occur in practice.

Figure 6.16 shows our experimental results. Our algorithms are drawn with solid lines. Competing handcrafted implementations are drawn with dotted lines. Implementations with coarse-grained synchronization are drawn with dashed lines. Error bars are not drawn because they are mostly too small to see: The standard deviation is less than 2% of the mean for half of the data points, and less than 10% of the mean for 95% of the data points. The STM data

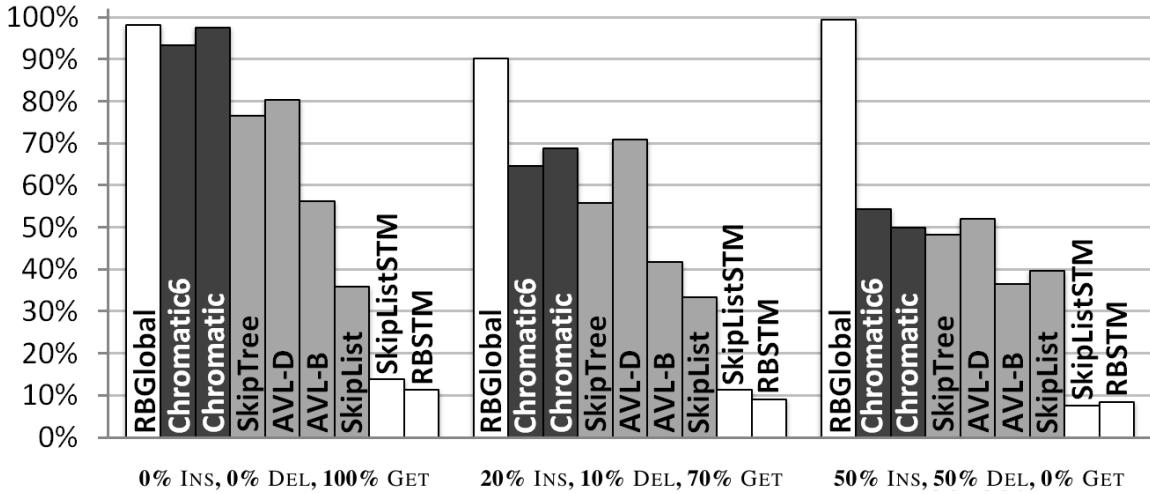


Figure 6.17: *Single threaded* throughput of the data structures relative to Java’s sequential RBT for key range $[0, 10^6]$.

structures are not included in the graphs for key range $[0, 10^6]$, because of the enormous length of time needed just to perform prefilling (more than 120 seconds per five second trial).

Chromatic6 nearly always outperforms Chromatic. The only exception is for an all query workload, where Chromatic performs slightly better. Chromatic6 is prefilled with the Chromatic6 insertion and deletion algorithms, so it has a slightly larger average leaf depth than Chromatic; this accounts for the performance difference. In every graph, Chromatic6 rivals or outperforms the other data structures, even the highly optimized implementations of SkipList and SkipTree which were crafted with the help of Doug Lea and the Java Community Process JSR-166 Expert Group. Under high contention (key range $[0, 10^2]$), Chromatic6 outperforms every competing data structure except for SkipList in case 50i-50d and AVL-D in case 0i-0d. In the former case, SkipList approaches the performance of Chromatic6 when there are many INSERTS and DELETES, due to the simplicity of its updates. In the latter case, the non-blocking searches of AVL-D allow it to perform nearly as well as Chromatic6; this is also evident for the other two key ranges. SkipTree, AVL-D and AVL-B all experience negative scaling beyond 32 threads when there are updates. For SkipTree, this is because its nodes contain many child pointers, and processes modify a node by replacing it (severely limiting concurrency when the tree is small). For AVL-D and AVL-B, this is likely because processes waste time waiting for locks to be released when they perform updates. Under moderate contention (key range $[0, 10^4]$), in cases 50i-50d and 20i-10d, Chromatic6 significantly outperforms the other data structures. Under low contention, the advantages of a non-blocking approach are less pronounced, but Chromatic6 is still at the top of each graph (likely because of low overhead and searches that ignore updates).

Figure 6.17 compares the single-threaded performance of the data structures, relative to that of the sequential RBT, `java.util.TreeMap`. This demonstrates that the overhead introduced by our technique is relatively small.

Although balanced BSTs are designed to give performance guarantees for worst-case sequences of operations, the experiments are performed using random sequences. For such sequences, BSTs *without* rebalancing operations are balanced with high probability and, hence, will have better performance because of their lower overhead. Thus, better experiments are needed to evaluate balanced BSTs.

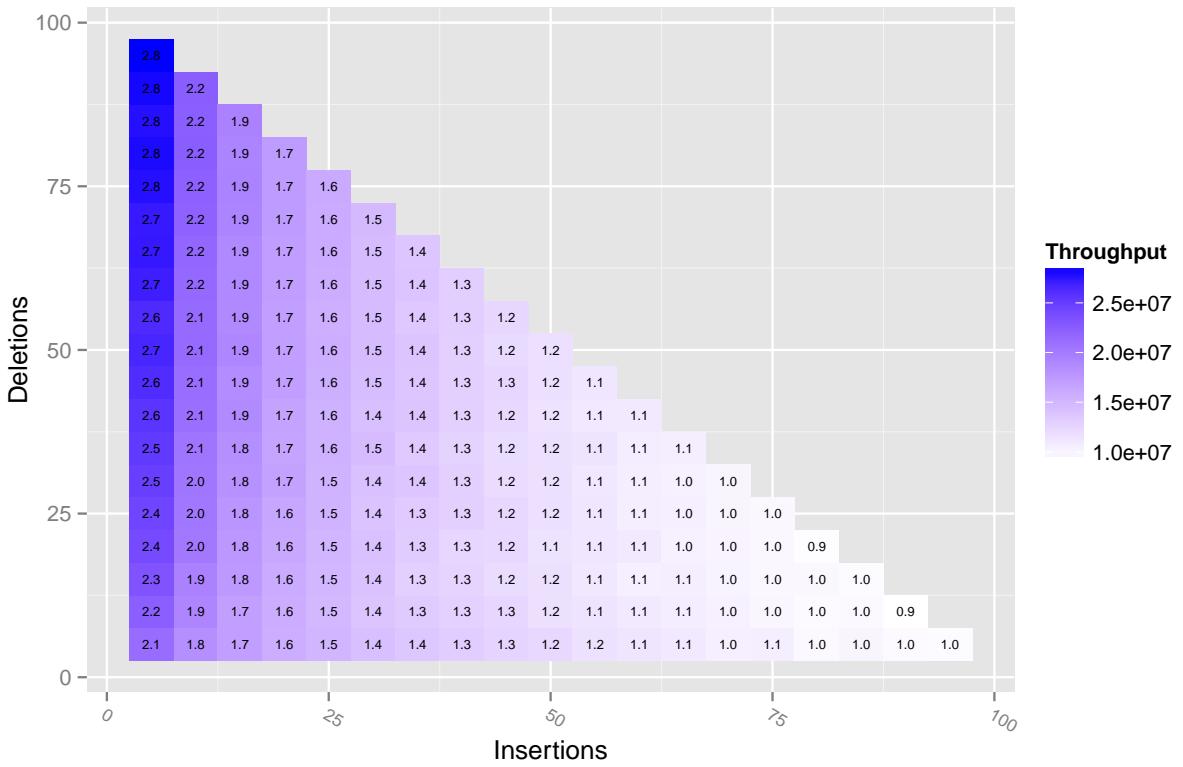


Figure 6.18: Heatmap showing throughput (operations per second) for **Chromatic6** over a wide variety of operation mixes, with key range $[0, 10^6]$ and 128 threads.

Exploring additional operation mixes The results above explored a fairly limited set of operation mixes. In the interest of obtaining a more complete picture of the performance of a subset of these algorithms, we also performed some supplementary experiments for a wider variety of operation mixes. Specifically, for each of the data structures in $\{\text{Chromatic6}, \text{SkipTree}, \text{AVL-B}\}$, we ran trials of the same sort that we performed above, with key range $[0, 10^6]$ and 128 threads, for *all* operation mixes with $x\%$ insertions and $y\%$ deletions, where $x, y \in \{5, 10, 15, \dots, 95\}$ and $x + y \leq 100$. The results for each data structure appear in Figure 6.18, Figure 6.19 and Figure 6.20.

Broadly, the results show that throughput increases as the fraction of operations that are searches increases. The results additionally show that throughput of Chromatic6 increases as the ratio of deletions to insertions increases (causing the size of the tree to be smaller in the steady state). This demonstrates that Chromatic6 scales well, even as contention increases. However, the results for SkipTree and AVL-B do not show this effect. Instead, performance *decreases* as the ratio of deletions to insertions increases. For SkipTree, this is because nodes contain many keys, so updates in a small tree can contend on a large fraction of the data structure. For AVL-B, we believe this is due to increased contention as a result of hand-over-hand locking, which must always start at the root.

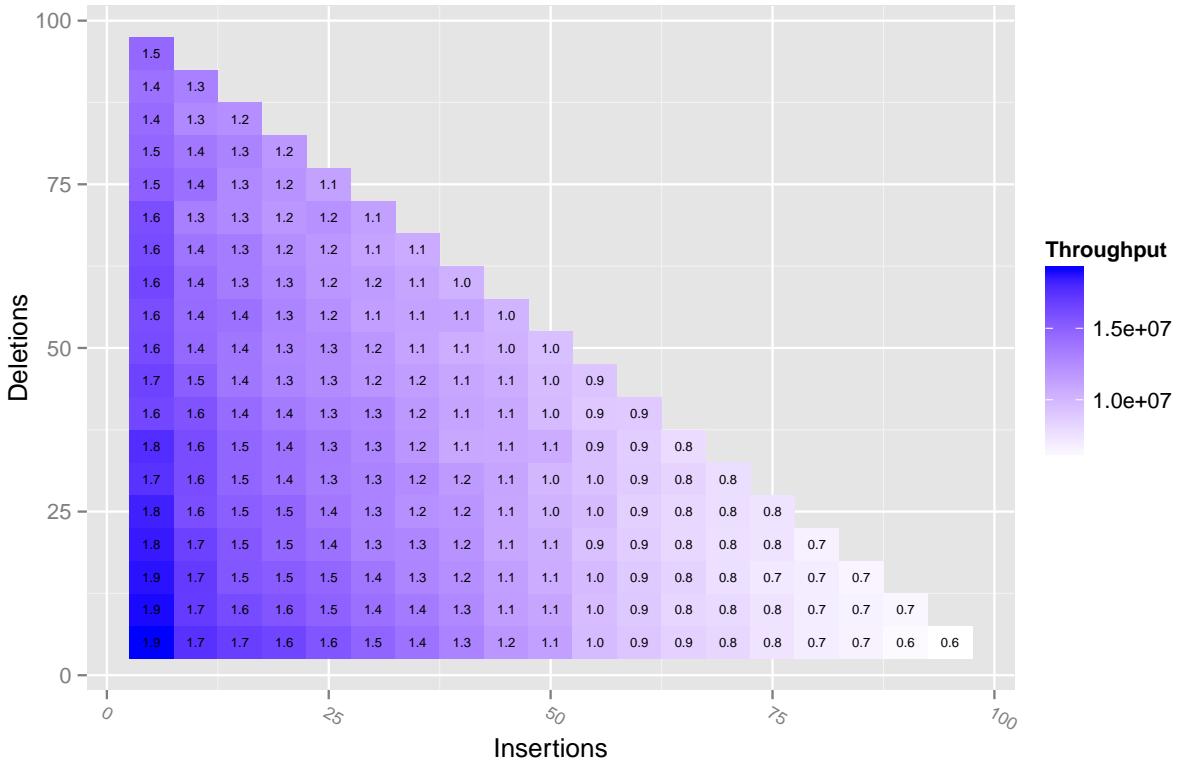


Figure 6.19: Heatmap showing throughput (operations per second) for **SkipTree** over a wide variety of operation mixes, with key range $[0, 10^6]$ and 128 threads.

6.8 Summary

In this chapter, we demonstrated the use of our tree update template by implementing a non-blocking chromatic tree. To the authors' knowledge, this is the first provably correct, non-blocking balanced BST with fine-grained synchronization. Proving the correctness of a direct implementation of a chromatic tree from hardware primitives such as CAS would have been completely intractable. By developing our template abstraction and the chromatic tree in tandem, we were able to minimize overhead, so the chromatic tree is very efficient.

We hope that this work sparks interest in developing more relaxed-balance sequential versions of data structures, since it is now easy to obtain efficient concurrent implementations of them using our template. We also hope this work leads to a greater diversity of advanced lock-free balanced trees. Towards this end, we present four different lock-free balanced trees in this thesis. Additionally, since the paper that presented our tree update template and lock-free chromatic tree was first published at PPoPP 2014, the *weak AVL trees* of Haeupler, Sen and Tarjan [60] have been implemented using the tree update template by He and Li [64].

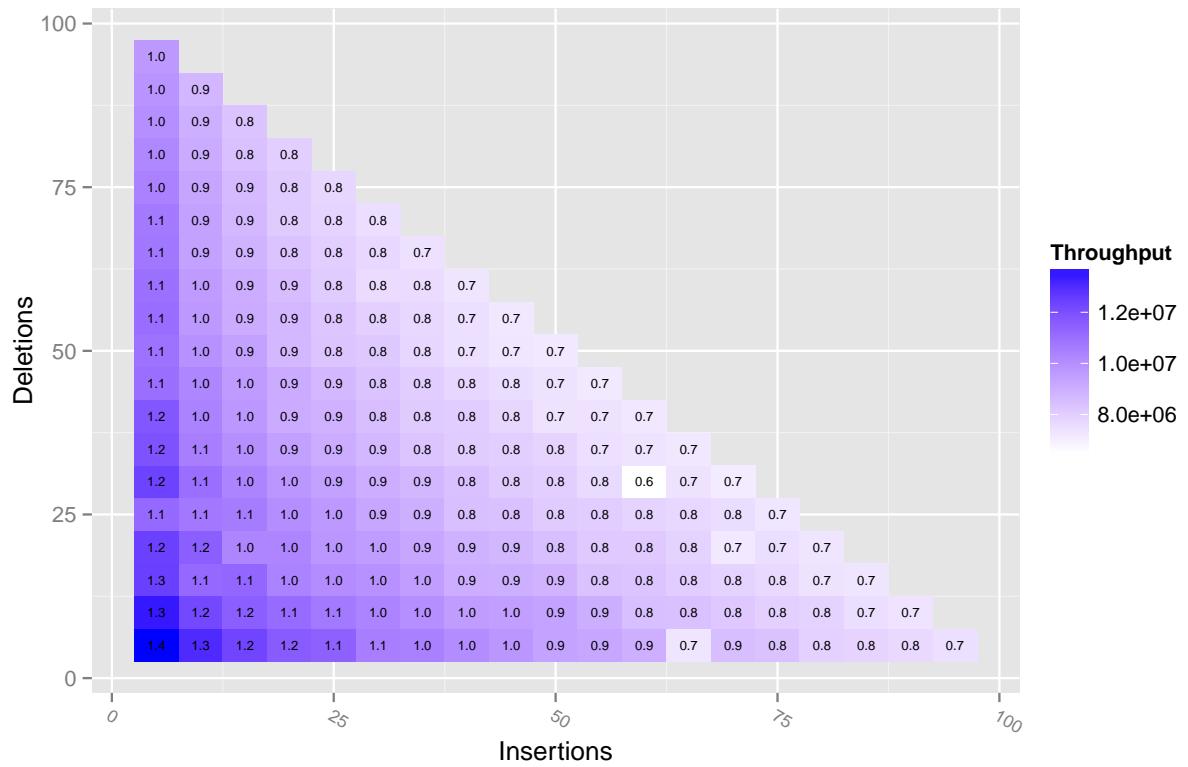


Figure 6.20: Heatmap showing throughput (operations per second) for **AVL-B** over a wide variety of operation mixes, with key range $[0, 10^6]$ and 128 threads.

Chapter 7

Relaxed AVL tree implemented with the template

Several papers have proposed more concurrency friendly versions of AVL trees that allow rebalancing steps to be decoupled from insertions and deletions [21, 81, 83, 105]. This decoupling allows the tree to temporarily become unbalanced while insertions and deletions are in progress, which allows a greater degree of concurrency. We consider the relaxation proposed by Larsen [81]. However, the other proposals could also be implemented using our template. Given a copy of [81] and a description of the tree update template, a first year undergraduate student produced a Java implementation of a relaxed-balance AVL tree in less than a week. We start by defining AVL trees, and then describe Larsen’s relaxation.

7.1 AVL trees

An AVL tree [1] is a balanced search tree with stricter balance than a red-black tree (or chromatic tree). Each *internal* node r in an AVL tree has a *balance factor* $bf(r) = h(r.left) - h(r.right)$, where $h(r)$ is the height of r , defined as follows.

$$h(r) = \begin{cases} 0 & : r \text{ is a leaf} \\ \max\{h(r.left), h(r.right)\} + 1 & : \text{otherwise} \end{cases}$$

Balance is maintained in an AVL tree with the following invariant.

AVL balance invariant: $bf(r) \in \{-1, 0, 1\}$ for all nodes r in the tree.

In other words, in an AVL tree, the heights of the left and right subtrees of each node can differ by at most one. This invariant yields a tree with height at most $\log_{\phi}(\sqrt{5}(n+2)) - 3$ where ϕ is the golden ratio, and n is the number of keys in the tree (see pp.460 of [77]). This is approximately $1.44\log_2(n+2)$, which is somewhat better than the upper bound of $2\log_n(n+1)$ on the height of a red-black tree.

7.2 Relaxed AVL trees

In a relaxed AVL (RAVL) tree, each node r has an integer *tag* that is zero if r is the root and, otherwise, satisfies $r.tag \geq -1$ for internal nodes and $r.tag \geq 1$ for leaves. Similar to balance factors in an AVL tree, each *internal* node r in a RAVL tree has a *relaxed balance factor* $rbf(r) = rh(r.left) - rh(r.right)$, where $rh(r)$ is the *relaxed height* of r , defined as follows.

$$rh(r) = \begin{cases} r.tag & : r \text{ is a leaf} \\ \max\{rh(r.left), rh(r.right)\} + 1 + r.tag & : \text{otherwise} \end{cases}$$

Intuitively, positive (resp., negative) tags let a node pretend that it is the root of a taller (resp., shorter) subtree. Note that a node can pretend to be the root of a *much* taller subtree, but it can only pretend to be the root of a slightly shorter subtree, since $tag \geq -1$. In analogy to AVL trees, we have the following balance invariant.

RAVL balance invariant: $rbf(r) \in \{-1, 0, 1\}$ for all nodes r in the tree.

We now describe the updates to the RAVL tree, which appear in Figure 7.1. Relaxed balance factors appear to the left of nodes, and tag values appear to the right. To simplify the presentation, we define two functions $l(u)$ and $r(u)$, such that $l(u) = 1$ if $rbf(u) = -1$, and $l(u) = 0$ otherwise, and $r(u) = 1$ if $rbf(u) = 1$, and $r(u) = 0$ otherwise.

We consider a dictionary implemented using a RAVL tree. Recall that a dictionary represents a set of key-value pairs, and offers a GET operation to search for the value associated with a key, an INSERT operation to add a new key-value pair (or replace the value in an existing key-value pair), and a DELETE operation to remove any existing

key-value pair for a given key. The tree is leaf-oriented, which means the key-value pairs in the dictionary are contained entirely in the leaves. Internal nodes contain routing keys which direct searches to the appropriate leaf.

To insert a key-value pair $\langle k, val \rangle$, a process begins searching for the leaf l where it should appear. If the leaf contains k , then the process performs **INSERTREPLACE**, which replaces the leaf with a new copy that contains val instead of the value previously associated with k . Otherwise, the process performs **INSERTNEW**, which replaces the leaf by a subtree consisting of an internal node with two leaves as its children. The left leaf contains the smaller of k and the key in l , and the right leaf contains the larger of k and the key in l . Both leaves have tag value zero and relaxed balance factor zero. The internal node contains the same key as the right leaf and has relaxed balanced factor zero. The tag value of the internal node is $l.tag - 1$.

This choice of tag value for the internal node maintains the RAVL balance invariant. To see why, consider the following. Let l be the leaf found by the search, n be the internal node after the update, and n_l and n_r be its left and right children. Let rbf (resp., rbf') be the relaxed balance factor before (resp., after) the update, and rh (resp., rh') be the relaxed height before (resp., after) the update. Suppose the RAVL balance invariant holds before the update. We prove it holds after the update. We first argue that the new internal node n has relaxed balance factor zero. By definition, $rh'(n_l) = n_l.tag = 0$ and $rh'(n_r) = n_r.tag = 0$, so $rbf'(n) = rh'(n_l) - rh'(n_r) = 0$. Now, we argue that no other internal node has its relaxed balance factor changed by the update. Observe that the update can only affect the relaxed balance factors of the ancestors of l . To argue that the update does not change the relaxed balance factor of any ancestor of l , it suffices to prove $rh'(n) = rh(l)$. By definition, $rh'(n) = \max\{rh'(n_l), rh'(n_r)\} + 1 + (l.tag - 1)$. Thus, $rh'(n) = \max\{0, 0\} + 1 + (l.tag - 1) = l.tag$. Since l is a leaf, $rh(l) = l.tag$, so $rh'(n) = rh(l)$.

To delete any key-value pair with key k , a process searches for k , ending at a leaf l . If the leaf does not contain k , then the deletion terminates. Otherwise, the process performs **DELETE**, which removes l and its parent, leaving only the sibling s of l . After the update, s has tag value $l.tag + s.tag + 1 + r(u)$. It is straightforward to verify that this choice of tag value maintains the RAVL balance invariant.

We now describe how rebalancing works in a RAVL tree. If a node r in a RAVL tree has a negative tag value, then we say a *negative violation* occurs at r . If r has a positive tag value, then *r.tag positive violations* occur at r . Observe that, if a RAVL tree T contains no violations, then for all $r \in T$, $r.tag = 0$, so $rh(r) = h(r)$, so $rbf(r) = bf(r)$, so T is an AVL tree. Thus, the goal of rebalancing is to perform rebalancing steps to eliminate all violations, while maintaining the RAVL balance invariant, to produce an AVL tree.

The updates starting with R in Figure 7.1 are rebalancing steps. For each rebalancing step, in the left side of the diagram, there is a violation that is either eliminated or moved upwards in the tree by the update. For example, R3 applies when there is a negative violation at v , and R3 either eliminates this violation or moves it to the parent. Each of the rebalancing steps also has a mirror-image that is obtained by flipping the left and right child pointers, and flipping any relaxed balance factor inequalities. We append **SYM** to a rebalancing step's name to indicate that it is the horizontal symmetry of the original rebalancing step. For example, R3SYM applies if $b_u \geq 0$ and the right child of u has $tag = -1$.

The set of rebalancing steps presented here is slightly different from the set presented in [81]. There, the rebalancing steps were introduced simply as *operation 3*, *operation 4*, and so on, through *operation 13*. Operations 3 and 4 correspond to R3 and R4, respectively. However, unlike R3 and R4, operations 3 and 4 do not require $t_u \geq 0$. Performing these operations when $t_u < 0$ causes the relaxed balance factor of a node to become -2 or +2, violating the RAVL balance invariant. Conceptually, the authors of [81] treated these relaxed balance factors of -2 and +2 as a new type of violation, and provided operations 5 through 13 to eliminate this new type of violation. Specifically,

operations 5 through 8 were designed to be performed after operation 3, and operations 9 through 13 were designed to be performed after operation 4. Thus, the elimination of a positive or negative violation required performing two rebalancing steps: either operation 3 or 4, followed by one of the operations 5 through 13. Here, rather than performing two separate rebalancing steps, we simply combine the two steps into one. Thus, the rebalancing steps presented in Figure 7.1 consist of R3 and R4, as well as *composite* rebalancing steps R3:5, R3:6, R3:7, R3:8, R4:9, R4:10, R4:11, R4:12 and R4:13. Each composite rebalancing step Rx:y has the same effect as first performing operation x , then performing operation y on the result.

7.3 Implementation overview

Here, we give only a high level overview of how the RAVL tree could be implemented, with the understanding that the implementation details would be similar to those of the chromatic tree in Chapter 6 and the relaxed (a, b) -tree that we describe in Chapter 8 (both of which are described and proved in full detail).

Broadly, the implementation is very similar to our chromatic tree. `INSERT` and `DELETE` respectively invoke `TRYINSERT` and `TRYDELETE` procedures, which perform a BST search, and return a leaf and its parent (and possibly grandparent). The updates in Figure 7.1 are implemented (using the template) just like the updates in the chromatic tree. Whenever an `INSERTNEW` or `DELETE` update is performed, a tag violation might be created. If an `INSERT` or `DELETE` creates a violation, it invokes a `CLEANUP` procedure. `CLEANUP` repeatedly searches for the key that was inserted or deleted, fixing any violations it sees, until it performs a search and sees no violations.

We now explain how `CLEANUP` decides which rebalancing step to perform when it finds a violation. If a violation occurs at the left (resp., right) child of a node, we call it a *left violation* (resp., *right violation*). The rebalancing steps in Figure 7.1 are applied when a left violation is found, and the symmetric updates are applied when a right violation is found. Without loss of generality, suppose `CLEANUP` encounters a left violation. In the following, we refer to nodes by their names in Figure 7.1. The violation found by `CLEANUP` always occurs at node v . Observe that $t_u \geq 0$, or else `CLEANUP` would have stopped at node u . If the violation is a negative violation, then one of R3, R3:5, R3:6, R3:7 or R3:8 applies. Otherwise, the violation is a positive violation, and there is either a nearby negative violation that can be fixed, or one of R4, R4:9, R4:10, R4:11, R4:12 or R4:13 applies.

We give the complete decision tree for determining which rebalancing step to perform. Suppose a negative violation occurs at v . If $rbf(u) \leq 0$, then R3 applies. So, suppose $rbf(u) > 0$. By the RAVL balance invariant, $rbf(u) = 1$. If $rbf(v) \geq 0$, then R3:5 applies. So, suppose $rbf(v) < 0$. If $w.tag > 0$, then R3:6 applies. Otherwise, if $w.tag = 0$, then R3:7 applies. Otherwise, R3:8 applies.

Now, suppose a positive violation occurs at v . If $rbf(u) \geq 0$, then R4 applies. So, suppose $rbf(u) < 0$. If $w.tag > 0$, then R4:9 applies. So, suppose $w.tag \leq 0$. In this case, before attempting to fix the positive violation, we first check if a negative violation occurs at v 's sibling w . If so, the decision tree for negative violations is used, and that violation is fixed first. So, we can assume that $w.tag \geq 0$, which implies that $w.tag = 0$. If $rbf(w) \leq 0$ then R4:10 applies. So, suppose $rbf(w) > 0$. By the RAVL balance invariant, $rbf(w) = 1$. If the left child x of w satisfies $x.tag > 0$, then R4:11 applies. Otherwise, if $x.tag = 0$, then R4:12 applies. Otherwise, R4:13 applies.

Correctness and progress Since we demonstrate with the chromatic tree and relaxed (a, b) -tree how such a proof would proceed, we leave this as an exercise. A proof of correctness and progress for the RAVL tree would follow the

exact same structure as the proof for the chromatic tree. One can simply copy the proof for the chromatic tree, and work through it, making minor changes.

Java implementation Given copies of [81] and [30], a first year undergraduate student produced a Java implementation of the RAVL tree in less than a week. Its performance was slightly lower than that of Chromatic. Similar to Chromatic6, we modified CLEANUP so that it only performs rebalancing steps if at least k violations appear in a path. Doing this improved performance for the RAVL tree, bringing it in line with Chromatic6. This suggests that the performance bottleneck in each data structure is the search phase of operations.

7.4 Towards a height bound

We claim that an RAVL tree containing n keys has height at most $O(\log n + c)$, where c is the number of unfinished insertions and deletions. A full proof of this would be similar to the proof of the height bound for the chromatic tree. Such a proof has two components: a proof that each RAVL tree contains at most c violations, and a proof that an RAVL tree containing c violations has height $O(\log n + c)$. The first component would follow the same approach as the proof for the chromatic tree, so we leave it as an exercise. However, the second component is somewhat subtle, so we prove it.

Lemma 7.1 *A relaxed AVL tree containing n keys and c violations has height $O(\log n + c)$.*

Proof: Suppose a relaxed AVL tree rooted at root contains n keys and c violations. By Corollary 2 of [81], if the topmost node in a rebalancing step is not root , then the rebalancing step does not change the relaxed height of the topmost node. In fact, each rebalancing step would leave the relaxed height of the root unchanged if it they were able to change its tag value. However, the tag value of the root is always zero, so, a rebalancing step that would normally set the tag value of the topmost node to a non-zero tag value (and, in doing so, preserve the relaxed height of the topmost node) will have the effect of changing the root's relaxed height. Thus, the root is the only node that can have its relaxed height changed by a rebalancing step (and it is clearly the topmost node in any rebalancing step that modifies it). In order to understand how the relaxed height of the root can be changed by a rebalancing step, we consider how the relaxed height of the topmost node in a rebalancing step changes depending on whether its tag value is forced to zero.

Consider any rebalancing step S . Before S , let u be its topmost node (shown in Figure 7.1), u_L and u_R be its children, and t be the value of $u.\text{tag}$. After S , let u' be the topmost node, u'_L and u'_R be its children, and t' be the value of $u'.\text{tag}$. By definition, $\text{rh}(u) = \max\{\text{rh}(u_L), \text{rh}(u_R)\} + 1 + t$ and $\text{rh}(u') = \max\{\text{rh}(u'_L), \text{rh}(u'_R)\} + 1 + t'$. By Corollary 2 of [81], $\text{rh}(u) = \text{rh}(u')$.

Now, suppose that S sets $u.\text{tag}$ to 0 instead of t' (which is what happens if $u = \text{root}$). After S , let u'' be the topmost node. Observe that the children of u'' are simply u'_L and u'_R . We compute $\text{rh}(u'') - \text{rh}(u)$. By definition $\text{rh}(u'') = \max\{\text{rh}(u'_L), \text{rh}(u'_R)\} + 1 + 0$. Thus, $\text{rh}(u'') - \text{rh}(u) = \max\{\text{rh}(u'_L), \text{rh}(u'_R)\} + 1 - (\max\{\text{rh}(u_L), \text{rh}(u_R)\} + 1 + t) = \max\{\text{rh}(u'_L), \text{rh}(u'_R)\} - \max\{\text{rh}(u_L), \text{rh}(u_R)\} - t$. Since $\text{rh}(u) = \text{rh}(u')$, we have $\max\{\text{rh}(u_L), \text{rh}(u_R)\} + 1 + t = \max\{\text{rh}(u'_L), \text{rh}(u'_R)\} + 1 + t'$, so $\max\{\text{rh}(u_L), \text{rh}(u_R)\} = \max\{\text{rh}(u'_L), \text{rh}(u'_R)\} + t' - t$. Substituting for $\max\{\text{rh}(u_L), \text{rh}(u_R)\}$ in the equation for $\text{rh}(u'') - \text{rh}(u)$, we have $\text{rh}(u'') - \text{rh}(u) = \max\{\text{rh}(u'_L), \text{rh}(u'_R)\} - (\max\{\text{rh}(u'_L), \text{rh}(u'_R)\} + t' - t) - t$. The right side of this equation simplifies to $-t'$. Therefore, S changes the relaxed height of u by $-t'$. By inspection of the rebalancing steps, t' is at least $t - 1$ and at most $t + 1$, where t is the value of $u.\text{tag}$ before S . Thus, when a

rebalancing step is applied where $u = \text{root}$ (so $t = 0$), t' is at least -1 and at most 1 . It follows that S changes the relaxed height of the root by $-1, 0$ or 1 .

By inspection of the rebalancing steps, each rebalancing step where $u = \text{root}$ reduces the number of violations in the tree by at least one. Therefore, removing c violations from the tree changes $|rh(\text{root})|$ by at most c . Suppose T is any relaxed AVL tree rooted at root_T that results from removing all c violations from the tree rooted at root by performing rebalancing steps. Then, $|rh(\text{root}) - rh(\text{root}_T)| \leq c$. Since T contains no violations, it is an AVL tree. Hence, $rh(\text{root}_T) = h(\text{root}_T)$, which implies $|rh(\text{root}) - h(\text{root}_T)| \leq c$. Since the tree rooted at root contains only c violations, we have $|rh(\text{root}) - h(\text{root})| \leq c$. Since $|rh(\text{root}) - h(\text{root})| \leq c$ and $|rh(\text{root}) - h(\text{root}_T)| \leq c$, we have $|h(\text{root}) - h(\text{root}_T)| \leq 2c$. Finally, since $h(\text{root}_T) \in O(\log n)$, $h(\text{root}) \in O(\log n + c)$. ■

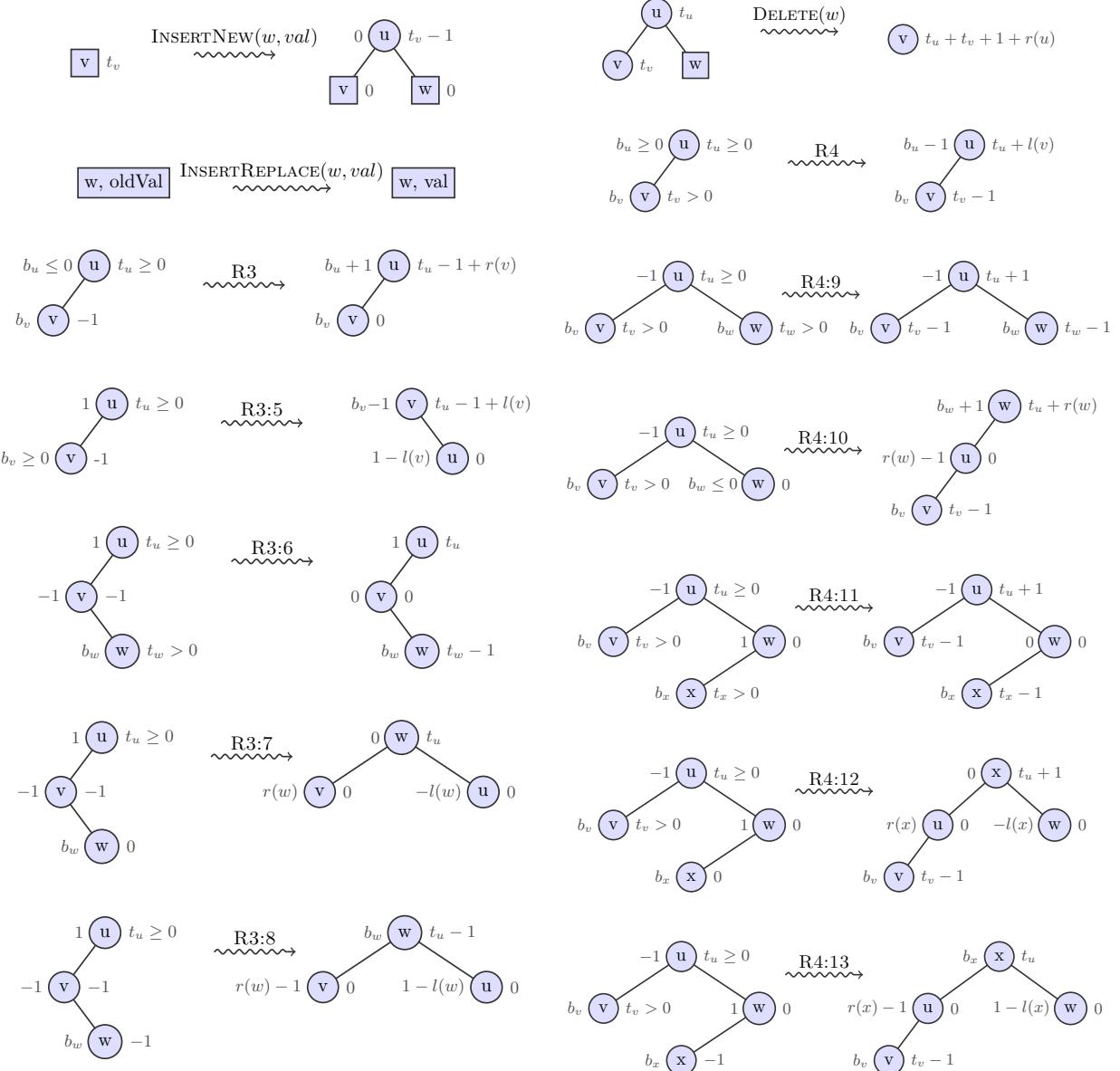


Figure 7.1: Updates for the RAVL tree. Relaxed balance factors appear on the left of nodes, and tag values on the right.

Chapter 8

Relaxed (a, b) -tree implemented with the template

8.1 (a,b) -trees and the relaxation

We first define (a,b) -trees, then explain how they are relaxed. An (a,b) -tree is a balanced, leaf-oriented search tree. Leaf-oriented B-trees, 2-3 trees and 2-3-4 trees are all special cases of (a,b) -trees.¹ Excluding the root, each leaf in an (a,b) -tree has between a and b keys, and each internal node has between a and b child pointers, where $b \geq 2a - 1$. If the root is a leaf, then it has between 1 and b keys. Otherwise, it has between 2 and b child pointers. Each internal node has one more child pointer than it has keys. The *degree* of a node u , denoted $|u|$, is the number of pointers it contains. We denote the parent of a node r by $\pi(r)$. The *level* of r is defined as follows.

$$l(r) = \begin{cases} 0 & : r \text{ is a leaf} \\ l(\pi(r)) + 1 & : \text{otherwise} \end{cases}$$

As in B-trees, the balance invariant in (a,b) -trees is quite strict: All leaves in an (a,b) -tree have the same level. Consequently, insertions and deletions of keys can involve a significant amount of work rebalancing the tree. In particular, rebalancing is necessary whenever a process tries to insert a key into a leaf that is already full (i.e., contains b keys), or delete a key from a leaf that contains only a keys. Sometimes rebalancing can be achieved with a small localized update, but other times it affects many nodes (even an entire path from the root to a leaf). Synchronizing on many nodes adds algorithmic complexity and limits concurrency.

In order to avoid these issues, relaxed (a,b) -trees: allow leaves to have different levels, allow nodes to contain fewer keys/pointers, split up monolithic rebalancing operations into small localized updates, and decouple rebalancing from insertion and deletion [84]. Each leaf in a relaxed (a,b) -tree has between zero and b keys, and each internal node has between one and b child pointers, where $b \geq 2a - 1$. Each node is augmented with a tag bit, which is zero for leaves and the root. We say a node is *tagged* if its tag bit is set. The *relaxed level* of a node r is defined as follows.

$$rl(r) = \begin{cases} r.tag & : r \text{ is a leaf} \\ rl(\pi(r)) + (1 - r.tag) & : \text{otherwise} \end{cases}$$

The relaxed level of r is just like the level of r , except any tagged nodes on the path from r to the root are not counted. The balance invariant in relaxed (a,b) -trees is: All leaves in a relaxed (a,b) -tree have the same *relaxed level*.

We now describe the updates to a relaxed (a,b) -tree, which appear in Figure 8.1. There, tag bits appear to the right of nodes. We consider a dictionary implemented using a relaxed (a,b) -tree. Recall that a dictionary represents a set of key-value pairs, and offers a GET operation to search for the value associated with a key, an INSERT operation to add a new key-value pair (or replace the value in an existing key-value pair), and a DELETE operation to remove any existing key-value pair for a given key. The tree is leaf-oriented, which means the key-value pairs in the dictionary are contained entirely in the leaves. Internal nodes contain routing keys which direct searches to the appropriate leaf. Leaves contain keys and pointers to values. Let $kv(u)$ be the set of key-value pairs in a node u .

To insert a key-value pair $\langle k, v \rangle$, a process begins searching for the leaf l where it should appear. If the leaf already contains some key-value pair $\langle k, x \rangle$, then the process performs REPLACEPAIR, which replaces that key-value pair with $\langle k, v \rangle$. Now, suppose the dictionary does not contain any key value pair with key k . If l contains fewer than b keys, then the process performs INSERTPAIR, which inserts $\langle k, v \rangle$ into l . Otherwise, the process performs OVERFLOW, which

¹Using (a,b) -tree notation, a 2-3 tree is a $(2,3)$ -tree, a 2-3-4 tree is a $(2,4)$ -tree, and a B-tree is either an $(a,2a)$ -tree or an $(a,2a-1)$ -tree (since there are two common definitions of a B-tree).

DELETEPAIR	$u \boxed{\alpha \boxed{k} v \gamma} 0 \xrightarrow{\text{DELETE}(k)} n \boxed{\alpha \gamma} 0$
REPLACEPAIR	$u \boxed{\alpha \boxed{k} x \gamma} 0 \xrightarrow{\text{INSERT}(k, v)} n \boxed{\alpha \boxed{k} v \gamma} 0$ leaf contains k
INSERTPAIR	$u \boxed{\alpha \gamma} 0 \xrightarrow{\text{INSERT}(k, v)} n \boxed{\alpha \boxed{k} v \gamma} 0$ leaf does not contain k and $ \alpha < b$
OVERFLOW	$u \boxed{\alpha} 0 \xrightarrow{\text{INSERT}(k, v)} n \boxed{} 1$ leaf does not contain k and $ \alpha = b$ $n_1 \boxed{\alpha_1 \boxed{k} \alpha_2} 0 \quad n_2 \boxed{\alpha_3} 0$ leaves containing $b + 1$ pairs evenly distributed
ROOTUNTAG	$\text{root } \boxed{\alpha} 1 \longrightarrow n \boxed{\alpha} 0$
ROOTABSORB	$\text{root } \boxed{} 0 \longrightarrow n \boxed{\alpha} 0$ $\text{root}_1 \boxed{\alpha} 1$ if the root has only one child, make that child the new root
ABSORBCHILD	$u \boxed{\alpha \beta \gamma} 0 \longrightarrow n \boxed{\alpha \beta \gamma} 0$ $u_i \boxed{\beta} 1$ initially, u contains less than b pointers
PROPAGATETAG	$u \boxed{\alpha \beta \gamma} 0 \longrightarrow n \boxed{} 1$ $u_i \boxed{\beta} 1 \quad n_1 \boxed{\alpha \beta \gamma_1} 0 \quad n_2 \boxed{\gamma_2} 0$ initially, u contains b pointers; afterwards, all pointers in α, β, γ are evenly distributed between the two children
ABSORBSIBLING	$u \boxed{\alpha \beta \gamma} 0 \longrightarrow n \boxed{\alpha \beta \gamma} 0$ $u_i \boxed{\alpha} 0 \quad u_{i+1} \boxed{\beta} 0 \quad n_i \boxed{\alpha \beta} 0$ $ \beta < a$ and $ \alpha + \beta < 2a$
DISTRIBUTE	$u \boxed{\alpha \beta \gamma} 0 \longrightarrow n \boxed{\alpha \beta \gamma} 0$ $u_i \boxed{\alpha \beta} 0 \quad u_{i+1} \boxed{\gamma} 0 \quad n_i \boxed{\alpha} 0 \quad n_{i+1} \boxed{\beta \gamma} 0$ $ \gamma < a$ and $ \alpha + \beta + \gamma \geq 2a$ evenly distribute the keys

Figure 8.1: Updates to relaxed (a, b) -trees. In leaves, a single key-value pair is drawn with the key on top, and the value on the bottom. Tag bits appear to the right of nodes. Note that tagged nodes have exactly two pointers. $|\alpha|$ denotes the degree of α (i.e., the number of pointers it contains).

conceptually replaces l with a small subtree consisting of an internal node and two leaves. The key-value pairs in $kv(l) \cup \{\langle k, v \rangle\}$ are distributed evenly between the leaves.

To delete any key-value pair with key k , a process searches for k , ending at a leaf l . If the leaf does not contain any key-value pair with key k , then the deletion terminates. Otherwise, the process performs **DELETEPAIR**, which removes any key-value pair with key k from l .

We now describe how rebalancing works in a relaxed (a, b) -tree. If a node r is tagged, then we say a *tag violation* occurs at r . If r is an untagged node that is not the root, and r has degree less than a , then we say a *degree violation* occurs at r . We say a *degree violation* occurs at the root only if it is an internal node with degree *one*. (Note that tag violations effectively supersede degree violations. That is, a degree violation does occur at a node if a tag violations occurs at that node.) Observe that a relaxed (a, b) -tree that contains no violations is an (a, b) -tree. Therefore, the objective of rebalancing is to eliminate violations, while maintaining the invariant that the data structure is a relaxed (a, b) -tree.

At a high level, violations are always created at leaves. Tag violations are created by **OVERFLOW** and degree violations are created by **DELETEPAIR**. Other updates either remove violations or propagate them up the tree towards the root, where they can always be removed.

We now describe precisely how violations are created, moved and removed by the (a, b) -tree updates in Figure 8.1. **OVERFLOW** creates a tag violation at the new internal node, and **DELETEPAIR** creates a degree violation at a leaf precisely if the leaf contains a keys just before the **DELETEPAIR** removes one of them. **INSERTPAIR** removes a degree violation at a leaf precisely if the leaf contains a keys just after the **INSERTPAIR** adds a key to it. **ROOTUNTAG** removes a tag violation at the root, and will create a degree violation in the process precisely if the root is internal and has degree one. **ROOTABSORB** removes a degree violation at the root. If there is a tag violation at the child of the root before the **ROOTABSORB**, then it removes that tag violation, as well, unless the child of the root had degree one, in which case the degree violation is conceptually *moved* from the child to the root. **ABSORBCCHILD** removes a tag violation at u_i . If its parent u has degree $a - 1$ before the update, and $|\alpha| + |\gamma| + |\beta| \geq a$, then the update also removes a degree violation at u . **PROPAGATETAG** moves a tag violation from u_i to u . **DISTRIBUTE** removes a degree violation at u_i or u_{i+1} . (In Figure 8.1, a degree violation is depicted at u_{i+1} , but the update is also applicable if the degree violation appears at u_i , instead. Note that **DISTRIBUTE** does not apply if degree violations simultaneously appear at both u_i and u_{i+1} , since it requires $|\alpha| + |\beta| + |\gamma| \geq 2a$.) **ABSORBSIBLING** is more complex.

Like **DISTRIBUTE**, **ABSORBSIBLING** attempts to fix a degree violation at u_i or u_{i+1} . If a single degree violation appears at either u_i or u_{i+1} (depicted at u_{i+1} in Figure 8.1), then **ABSORBSIBLING** removes it. However, unlike **DISTRIBUTE**, **ABSORBSIBLING** can be applied when degree violations occur at both u_i and u_{i+1} (which occurs precisely when $|\alpha| < a$ and $|\alpha| + |\beta| \geq a$). In this case, **ABSORBSIBLING** removes both degree violations. However, since **ABSORBSIBLING** removes the node u_{i+1} , if u contains a pointers before the update, then the update will create a degree violation at u . Therefore, depending on the degrees of u , u_i and u_{i+1} , **ABSORBSIBLING** will either: (1) remove degree violations at u_i and u_{i+1} , (2) remove degree violations at u_i and u_{i+1} and create one at u , (3) remove a degree violation at either u_i or u_{i+1} , or (4) remove a degree violation at either u_i or u_{i+1} and create one at u . In case (2) and (4), we think of the violation created at u as having been *moved* there from u_i or u_{i+1} (so, conceptually, no new violation is created). Thus, **ABSORBSIBLING** removes up to two degree violations, and possibly moves a degree violation from a node to its parent.

```

type Node
  ▷ User-defined fields
  tag           ▷ tag bit (immutable)
   $k_1, k_2, \dots, k_d$  ▷ keys (immutable)
   $p_1, p_2, \dots, p_d$  ▷ pointers (mutable)
  d             ▷ degree of the node (immutable)
  ▷ Fields used by LLX/SCX algorithm
  info          ▷ pointer to SCX-record
  marked        ▷ marked bit

```

Figure 8.2: Data definition for a node in the (a, b) -tree.

```

1 GET(key)
2    $\langle -, -, l \rangle := \text{SEARCH}(key)$ 
3   if l contains key then return the value associated with key
4   else return NIL
5
6 SEARCH(key)
7   gp := NIL; p := entry; l := entry.p1
8   while l is internal
9     gp := p; p := l
10    i := 1
11    while i < l.d and key  $\geq l.k_i$  do i := i + 1      ▷ Locate appropriate child pointer to follow
12    l := l.pi                                         ▷ Follow the child pointer
13   return  $\langle gp, p, l \rangle$ 

```

Figure 8.3: GET and SEARCH.

8.2 Implementation

Let a be the minimum degree of nodes, and b be the maximum degree. We represent each node by a Data-record with b mutable pointers, and b immutable keys, as well as immutable fields d and tag which contain the node's degree and tag bit, respectively. (See Figure 8.2.) The degree d represents the number of pointers that are used, and the tag bit indicates whether the node is tagged. Internal nodes have one fewer key than pointers, so k_d is unused in an internal node. Leaves have exactly as many pointers as keys.

To avoid special cases when the (a, b) -tree is empty, we add a sentinel node at the top of the tree. The sentinel node *entry* always has one child and no keys. (Every search that passes through it will simply follow that one child pointer.) The sole child of this sentinel node is initially an empty leaf (with $d = 0$ and no keys or pointers). The actual (a, b) -tree is rooted at the child of the sentinel node. For convenience, we use *root* to refer to the current child of the sentinel node.

Detailed pseudocode for GET, INSERT and DELETE is given in Figure 8.3, 8.4 and 8.5. GET, INSERT and DELETE each execute an auxiliary procedure, $\text{SEARCH}(key)$, which appears in Figure 8.3. $\text{SEARCH}(key)$ starts at *entry* and traverses nodes as in an ordinary B-tree search, reading child pointers until reaching a leaf, which it then returns (along with the leaf's parent and grandparent). Sometimes, the grandparent returned by an invocation of SEARCH is

not accessed. It is easy to argue, by inspection of the code, and because of the sentinel node, that the leaf's parent always exists (i.e., is not NIL), and its grandparent exists whenever it is accessed. We define the *search path* for *key* at any time to be the path that `SEARCH(key)` would follow, if it were done instantaneously. The `GET(key)` operation simply executes a `SEARCH(key)` and then returns the value found in the leaf if it contains *key*, and \perp otherwise.

8.2.1 Detailed description of insertion

Pseudocode for `INSERT` and its helper function `TRYINSERT` appear in Figure 8.4. `INSERT` is identical to the `INSERT` procedure for the chromatic tree. It simply repeatedly invokes `TRYINSERT` until it successfully performs the insertion, and then invokes `CLEANUP` if it created a violation. `TRYINSERT` takes *key* and *value* as its arguments, and returns either `FAIL`, or a pair $\langle \text{createdViolation}, \text{oldValue} \rangle$. In this pair, *createdViolation* is a bit that indicates whether the invocation of `TRYINSERT` created a violation, and *oldValue* is the value that was previously associated with *key*. If `TRYINSERT` returns `FAIL`, this signals to `INSERT` that the update was not successful, and `INSERT` should perform another invocation of `TRYINSERT` to try again.

`TRYINSERT` first invokes `SEARCH(key)`, which returns a leaf *l* and its parent *p*. Then, it performs LLX on *p* and *l*. If either LLX returns `FAIL` or `FINALIZED`, then `TRYINSERT` returns `FAIL`. So, we assume both LLXs succeed. After the `LLX(p)`, `TRYINSERT` verifies that *p* still points to *l* (since *p* may have changed since it was visited by `SEARCH`). This is conceptually part of the `CONFLICT` procedure in the template. After the `LLX(l)`, `TRYINSERT` computes `SCX-ARGUMENTS`.

The creation of the arguments to `SCX` is straightforward. If *l* already contains *key*, then `TRYINSERT` performs `REPLACEPAIR` by setting *oldValue* to the value that was associated with *key* in *l*, and replacing *l* with a new node in which the previous association $\langle \text{key}, \text{oldValue} \rangle$ is replaced with $\langle \text{key}, \text{value} \rangle$. This does not create any violation, so *createdViolation* is set to false. Otherwise, if *l* contains fewer than *b* keys, then `TRYINSERT` performs `INSERTPAIR` by replacing *l* with a new copy that has $\langle \text{key}, \text{value} \rangle$ inserted. Since there was no value previously associated with *key*, *oldValue* is set to `NIL`. As in the previous case, this does not create any violation, so *createdViolation* is set to false. Otherwise, *l* already contains *b* keys, so `TRYINSERT` performs `OVERFLOW` by replacing *l* with a subtree of three newly created nodes (configured as shown in Figure 8.1). Since there was no value previously associated with *key*, *oldValue* is set to `NIL`. Since `OVERFLOW` creates a violation, *createdViolation* is set to true.

8.2.2 Detailed description of deletion

Pseudocode for `DELETE` and its helper function `TRYDELETE` appear in Figure 8.5. `DELETE` is identical to the `DELETE` procedure for the chromatic tree. It simply repeatedly invokes `TRYDELETE` until it successfully performs the deletion, and then invokes `CLEANUP` if it created a violation. `TRYDELETE` takes *key* as its argument, and returns either `FAIL`, or a pair $\langle \text{createdViolation}, \text{oldValue} \rangle$. If `TRYDELETE` returns `FAIL`, this signals that the update was not successful, and `DELETE` should perform another invocation of `TRYDELETE` to try again.

`TRYDELETE` is identical to `TRYINSERT` up until it begins computing `SCX-ARGUMENTS`, so we jump straight to the description of how it computes `SCX-ARGUMENTS`. If *l* does not contain *key*, then *key* is not in the tree, so `TRYDELETE` simply returns $\langle \text{FALSE}, \text{oldValue} \rangle$. Otherwise, `TRYDELETE` performs `DELETEPAIR` by setting *oldValue* to the value that was associated with *key* in *l*, and replacing *l* with a new copy that does not contain *key*. The replacement of *l* creates a (degree) violation if and only if *l* contains exactly *a* key-value pairs. So, *createdViolation* is set

```

13  INSERT(key,value)
14    ▷ Identical to INSERT procedure for the chromatic tree (see Figure 6.6).

15  TRYINSERT(key,value)
16    ▷ Returns  $\langle \text{TRUE}, \perp \rangle$  if key was not in the dictionary and inserting it caused a violation,
       $\langle \text{FALSE}, \perp \rangle$  if key was not in the dictionary and inserting it did not cause a violation,
       $\langle \text{FALSE}, \text{oldValue} \rangle$  if  $\langle \text{key}, \text{oldValue} \rangle$  was in the dictionary, and FAIL if we should try again

18    ▷ Search for key in the tree
19     $\langle -, p, l \rangle := \text{SEARCH}(key)$ 

21    ▷ Template iteration 0 (parent of leaf)
22     $\text{result}_p := \text{LLX}(p)$ 
23    if  $\text{result}_p \in \{\text{FAIL}, \text{FINALIZED}\}$  then return FAIL
24    if  $l \notin \text{result}$  then return FAIL           ▷ CONFLICT: verify p still points to l
25    Let  $p.p_i$  be the child pointer of p that pointed to l at the previous line

27    ▷ Template iteration 1 (leaf)
28     $\text{result}_l := \text{LLX}(l)$ 
29    if  $\text{result}_l \in \{\text{FAIL}, \text{FINALIZED}\}$  then return FAIL

31    ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
32     $V := \langle p, l \rangle$ 
33     $R := \langle l \rangle$ 
34     $fld :=$  a pointer to  $p.p_i$ 
35    if l contains key then                                ▷ Replace the value associated with an existing key
36      Let oldValue be the value associated with key in l
37       $\text{newNode} :=$  new copy of l that contains  $\langle \text{key}, \text{value} \rangle$  instead of  $\langle \text{key}, \text{oldValue} \rangle$ 
38       $\text{createdViolation} := \text{false}$ 
39    else if  $l.d < b$  then                                     ▷ Insert a new key-value pair into a full leaf
40       $\text{newNode} :=$  new copy of l that has the new key-value pair inserted
41       $\text{oldValue} := \text{NIL}$ 
42       $\text{createdViolation} := \text{false}$ 
43    else ▷  $l.d = b$                                          ▷ Insert a new key-value pair into a non-full leaf
44       $\text{newNode} :=$  pointer to a subtree of three newly created nodes: one internal node and two leaves, configured as
        in OVERFLOW in Figure 8.1 (so that the key-value pairs in  $\text{kv}(l) \cup \{\langle \text{key}, \text{value} \rangle\}$  are evenly distributed
        between the leaves, and the internal node is tagged)
45       $\text{oldValue} := \text{NIL}$ 
46       $\text{createdViolation} := \text{true}$ 

48    if  $\text{SCX}(V, R, fld, \text{newNode})$  then return  $\langle \text{createdViolation}, \text{oldValue} \rangle$ 
49    else return FAIL

```

Figure 8.4: Pseudocode for INSERT and TRYINSERT. Here, *a* is the minimum degree of nodes, and *b* is the maximum degree of nodes.

```

50  DELETE(key)
51    ▷ Identical to DELETE procedure for the chromatic tree (see Figure 6.7).
52
52 TRYDELETE(key)
53  ▷ Returns  $\langle \text{FALSE}, \perp \rangle$  if key was not in the dictionary,
54     $\langle \text{FALSE}, \text{oldValue} \rangle$  if  $\langle \text{key}, \text{oldValue} \rangle$  was in the dictionary and deleting it did not create a violation,
55     $\langle \text{TRUE}, \text{oldValue} \rangle$  if  $\langle \text{key}, \text{oldValue} \rangle$  was in the dictionary and deleting it created a violation, and
56    FAIL if we should try again
57
57  ▷ Search for key in the tree
58   $\langle -, p, l \rangle := \text{SEARCH}(key)$ 
59
60  ▷ Template iteration 0 (parent of leaf)
61   $\text{result}_p := \text{LLX}(p)$ 
62  if  $\text{result}_p \in \{\text{FAIL}, \text{FINALIZED}\}$  then return FAIL
63  if  $l \notin \text{result}$  then return FAIL           ▷ CONFLICT: verify p still points to l
64  Let  $p.p_i$  be the child pointer of p that pointed to l at the previous line
65
65  ▷ Template iteration 1 (leaf)
66   $\text{result}_l := \text{LLX}(l)$ 
67  if  $\text{result}_l \in \{\text{FAIL}, \text{FINALIZED}\}$  then return FAIL
68
68  ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
69   $V := \langle p, l \rangle$ 
70   $R := \langle l \rangle$ 
71   $fld :=$  a pointer to  $p.p_i$ 
72  if l does not contain key then           ▷ The tree does not contain key
73     $\text{oldValue} := \text{NIL}$ 
74     $\text{createdViolation} := \text{false}$ 
75    return  $\langle \text{createdViolation}, \text{oldValue} \rangle$ 
76  else
77    Let oldValue be the value associated with key in l
78     $\text{newNode} :=$  new copy of l that does not contain  $\langle \text{key}, \text{oldValue} \rangle$ 
79     $\text{createdViolation} := (l.d = a)$ 
80    if  $\text{SCX}(V, R, fld, \text{newNode})$  then return  $\langle \text{createdViolation}, \text{oldValue} \rangle$ 
81    else return FAIL

```

Figure 8.5: Pseudocode for DELETE and TRYDELETE. Here, *a* is the minimum degree of nodes, and *b* is the maximum degree of nodes.

to the result of the expression $l.d = a$. (Observe that, if $l.d < a$, then the degree violation was actually created by a *previous* deletion at l .)

8.2.3 The rebalancing algorithm

As in the chromatic tree, we assign responsibility for a violation to the process that created it. Whenever a process creates a violation, it executes a CLEANUP procedure that repeatedly searches for violations and performs rebalancing steps to eliminate them, terminating when it no longer finds any.

At a high level, in each iteration of the outer loop, CLEANUP(key) searches for key , stopping at the first violation it encounters. If it does not find a violation, then it terminates. Otherwise, it determines which rebalancing step it should perform, and then invokes the appropriate procedure in {TRYROOTUNTAG, TRYROOTABSORB, TRYABSORBCCHILD, TRYPROPAGATETAG, TRYABSORBSIBLING, TRYDISTRIBUTE}. Conceptually, each of these procedures implements the update phase of an update in Figure 8.1, taking the return value m of the search phase as its argument. (Note that we discuss precisely how these procedures are implemented, below.) After invoking one of these procedures, CLEANUP moves to the next iteration of the loop.

More specifically, a process executing CLEANUP first checks if there is a tag violation at $root$ (line 90). If so, it invokes TRYROOTUNTAG. Otherwise, the process checks if there is a degree violation at $root$ (line 91). If so, it invokes TRYROOTABSORB. Otherwise, it enters the loop at line 93, where it searches for a violation on the search path for key . The process exits this loop when it reaches a leaf without finding a violation and returns from CLEANUP (at line 94), or finds a violation and exits the loop (at line 99). Consequently, at line 103, we know there is a violation at l .

If l is tagged, then CLEANUP invokes TRYABSORBCCHILD or TRYPROPAGATETAG, as appropriate. Suppose l is not tagged. Then there is a degree violation at l . Degree violations are fixed using ABSORB SIBLING and DISTRIBUTE, which manipulate l and its sibling s . Since these rebalancing steps require that l and s are both untagged, CLEANUP must first check whether s is tagged. If so, CLEANUP invokes TRYABSORBCCHILD or TRYPROPAGATETAG to fix the tag violation at s . (After fixing a tag violation at s , CLEANUP will move to the next iteration of the loop, and search again.) Otherwise, CLEANUP invokes TRYABSORBSIBLING or TRYDISTRIBUTE, as appropriate, to fix the degree violation at l .

The argument for why this CLEANUP algorithm yields an upper bound on the height of the tree is very similar to the argument for the chromatic tree. It relies on the fact that an INSERT($key, value$) or DELETE(key) always creates its violation at the terminal leaf on the search path for key , and a violation can only be moved from a node to its parent. Thus, intuitively, a process can always find any violation it created while performing an INSERT($key, value$) or DELETE(key) by searching for key . (And, similarly, if a process finds no violation while searching for key , then any violation it created has been eliminated.) We will discuss this further in Section 8.5.

8.2.4 Implementing a rebalancing step

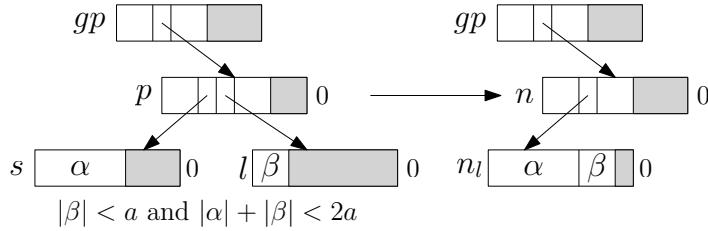
We now explain how TRYABSORBSIBLING is implemented. The procedures for the other rebalancing steps are similar. Pseudocode appears in Figure 8.7. TRYABSORBSIBLING takes four nodes gp, p, l and s , as well as three integers ix_p, ix_l and ix_s , as its arguments. When TRYABSORBSIBLING is invoked by CLEANUP, there is a degree violation at the node l , and no other violations at p, s or l . Furthermore, during its most recent search phase, CLEANUP saw $gp.p_{ix_p} = p, p.p_{ix_l} = l$ and $p.p_{ix_s} = s$.

```

82 CLEANUP(key)
83   ▷ Ensures the violation created by an OVERFLOW or DELETEPAIR (which is, in turn, caused by an
     ▷ INSERT(key, value) or DELETE(key)) gets eliminated
84   while TRUE                                ▷ Repeatedly search until no violation is found
85     ▷ Conceptually, SEARCHPHASE starts here
86     gp := NIL; p := entry; l := entry.p1          ▷ Save the three last nodes traversed
87     ixp = 0; ixl = 1                      ▷ Also save the indices of the pointers to p and l
88
89   ▷ Base case: check for violations at the root of the (a, b)-tree
90   if l.tag = 1 then TRYROOTUNTAG(p, ixl, l)
91   else if l is internal and l.d = 1 then TRYROOTABSORB(p, ixl, l)
92   else ▷ Continue to search for key, looking for a violation
93     loop
94       if l is a leaf then return                  ▷ Arrived at leaf without finding any violation
95       ixp := ixl; ixl := 1
96       while ixl < l.d and key ≥ l.kixl      ▷ Locate appropriate child pointer to follow
97         ixl := ixl + 1
98         gp := p; p := l; l := l.pixl           ▷ Follow the child pointer
99         if l.tag = 1 or l.d < a then exit loop    ▷ Exit the loop if there is a violation at l
100
101  ▷ Note: if we got here, we found a violation at l and gp ≠ NIL
102  ▷ Try to fix any tag violation at l (recall: tag violations supersede degree violations)
103  if l.tag = 1 then
104    if p.d + l.d ≤ b + 1 then TRYABSORBCCHILD(gp, ixp, p, ixl, l)
105    else TRYPROPAGATETAG(gp, ixp, p, ixl, l)
106  else ▷ Try to fix the degree violation at l (assert: l.d < a and gp ≠ NIL)
107    ixs := (ixp > 0 ? ixp - 1 : ixp + 1)          ▷ Compute the index of a sibling of l
108    s := p.pixs                                     ▷ Get a pointer to this sibling
109    ▷ We can only fix the degree violation at l if s is not tagged, so we first check if s is tagged
110    if s.tag = 1 then
111      ▷ Fix the tag violation at s
112      if p.d + s.d ≤ b + 1 then TRYABSORBCCHILD(gp, ixp, p, ixs, s)
113      else TRYPROPAGATETAG(gp, ixp, p, ixs, s)
114    else
115      ▷ Both l and s are untagged, so we try to fix the degree violation at l
116      if l.d + s.d < 2a then TRYABSORBSIBLING(gp, ixp, p, ixl, l, ixs, s)
117      else TRYDISTRIBUTE(gp, ixp, p, ixl, l, ixs, s)

```

Figure 8.6: Pseudocode for CLEANUP. Here, *a* is the minimum degree of nodes, and *b* is the maximum degree of nodes.



TRYABSORBSIBLING($gp, ix_p, p, ix_l, l, ix_s, s$)

▷ Implements the UPDATEPHASE for the ABSORB SIBLING update. Precondition: there is a degree violation at l , and there are no other violations at p, s or l .

▷ Template iteration 0 (grandparent of l)

```
resultgp := LLX(gp)
if resultgp ∈ {FAIL, FINALIZED} then return FAIL
Let ptrp be the value in resultgp that was read from gp.pix_p
if ptrp ≠ p then return FAIL
```

▷ CONFLICT: verify gp still points to p

▷ Template iteration 1 (parent of l)

```
resultp := LLX(p)
if resultp ∈ {FAIL, FINALIZED} then return FAIL
Let ptrl be the value in resultp that was read from p.pix_l
Let ptrs be the value in resultp that was read from p.pix_s
if ptrl ≠ l then return FAIL
```

▷ CONFLICT: verify p still points to l

```
if ptrs ≠ s then return FAIL
```

▷ CONFLICT: verify p still points to s

▷ Determine whether s is the left or right sibling of l (used to order the nodes in V top-to-bottom, then left-to-right)

```
if ixs < ixl then left := s; right := l
```

▷ s is the left sibling of l

```
else left := l; right := s
```

▷ s is the right sibling of l

▷ Template iteration 2 (the leftmost of l and s)

```
resultleft := LLX(left)
if resultleft ∈ {FAIL, FINALIZED} then return FAIL
```

▷ Template iteration 3 (the rightmost of l and s)

```
resultright := LLX(right)
if resultright ∈ {FAIL, FINALIZED} then return FAIL
```

▷ Compute SCX-ARGUMENTS: Start by creating new nodes according to the right-hand of the diagram above

(using the immutable fields of gp, p, l and s , and the results of the LLXs)

create node n_l with key-value pairs $kv(l) \cup kv(s)$

create node new by copying p , removing the child pointer to s (and the corresponding key), and changing the pointer to l to point to n_l

$V := \langle gp, p, left, right \rangle$

$R := \langle p, left, right \rangle$

$fld := \&gp.p_{ix_p}$

▷ Perform the SCX to change child pointer ix_p of node gp

$SCX(V, R, fld, new)$

Figure 8.7: Implementing rebalancing step ABSORB SIBLING. Other rebalancing steps are handled similarly using the diagrams shown in Figure 8.1. Here, a is the minimum degree of nodes, and b is the maximum degree of nodes.

`TRYABSORBSIBLING` begins by invoking `LLX(gp)`. Conceptually, this marks the beginning of iteration zero of the loop in the template. If any `LLX` performed by `TRYABSORBSIBLING` returns `FAIL` or `FINALIZED`, then `TRYABSORBSIBLING` immediately returns `FAIL`, which prompts `CLEANUP` to search again for violations to fix. If the `LLX(gp)` returns a snapshot, then `TRYABSORBSIBLING` verifies that $gp.p_{ix_p}$ still points to p , according to the result of the `LLX(gp)`. If $gp.p_{ix_p}$ no longer points to p , then `TRYABSORBSIBLING` immediately returns `FAIL`. In terms of the template, this verification is considered to be part of the `CONFLICT` procedure.

In iteration one of the loop in the template, `TRYABSORBSIBLING` invokes `LLX(p)`. If this `LLX` returns a snapshot, then `TRYABSORBSIBLING` verifies that $p.p_{ix_l}$ still points to l and $p.p_{ix_s}$ still points to s , according to the result of the `LLX(p)`. If $p.p_{ix_l}$ no longer points to l or $p.p_{ix_s}$ no longer points to s , then `TRYABSORBSIBLING` immediately returns `FAIL`. (As above, these verification steps are considered to be part of `CONFLICT`.)

Recall that the template places an ordering constraint on the nodes in the V sequence that will be passed to `SCX`. Specifically, the sequences V constructed by all updates that take place entirely during a period of time when no `SCX`s change the tree structure must be ordered consistently according to a fixed tree traversal algorithm (for example, an in-order traversal or a breadth-first traversal). We use a breadth-first traversal (i.e., top-to-bottom, then left-to-right). So, `TRYABSORBSIBLING` next determines whether s is the left or right sibling of l , and gives the alias *left* to the leftmost of the two, and the alias *right* to the rightmost of the two.

In iterations two and three of the loop in the template, `TRYABSORBSIBLING` simply invokes `LLX(left)` and `LLX(right)`. Finally, `TRYABSORBSIBLING` computes `SCX-ARGUMENTS`, and invokes `SCX` to perform the update. The return value of `SCX` is ignored, because it does not matter whether the `SCX` succeeds or fails. In either case, `CLEANUP` will continue to perform rebalancing steps until it performs a search without encountering any violation. It is straightforward to verify that each iteration of the outer loop in `CLEANUP` follows the template.

8.3 Correctness proof

In the following, the terms **search path** and **range** are defined as they were in the proof of the chromatic tree (in Section 6.3).

Lemma 8.1 *Our implementation of a relaxed (a,b) -tree satisfies the following claims.*

1. *TRYINSERT and TRYDELETE follow the tree update template and satisfy all constraints specified by the template. Each iteration of the outer loop in CLEANUP follows the template.*
2. *The node entry always has no keys and one child pointer. The child pointer points to a node $\text{root} \neq \text{entry}$.*
3. *If a node v is in the data structure in some configuration C and v was on the search path for key k in some earlier configuration C' , then v is on the search path for k in C . (Equivalently, updates to the tree do not shrink the range of any node in the tree.)*
4. *If an invocation of `SEARCH(k)` reaches a node v , then there was some earlier configuration during the search when v was on the search path for k .*
5. *The `SEARCH` procedure used by `TRYINSERT` and `TRYDELETE` satisfies DTP.*
6. *All invocations of `TRYINSERT` and `TRYDELETE` that perform a successful `SCX` are atomic (including their search phases). All invocations of `TRYROOTUNTAG`, `TRYROOTABSORB`, `TRYABSORBCCHILD`, `TRYPROPAGATETAG`, `TRYABSORBSIBLING` and `TRYDISTRIBUTE` that perform a successful `SCX` are atomic.*
7. *The tree rooted at the child of entry is always a relaxed (a,b) -tree.*

Proof: We prove these claims together by induction on the sequence of steps (invocations and responses of procedures, atomic invocations of SEARCH (for Claim 5) reads from shared memory, and LLXs and SCXs) in an execution. That is, we assume that all of the claims hold before an arbitrary step in the execution, and prove they hold after the step.

Claim 1: This claim follows almost immediately from inspection of the code. The only subtlety is showing that no process invokes $\text{LLX}(r)$ where $r = \text{NIL}$. Suppose the inductive hypothesis holds just before an invocation I of $\text{LLX}(r)$.

Suppose I occurs in TRYINSERT or TRYDELETE . Then, r must be one of the nodes p or l returned by SEARCH . By inspection of the code, l is not NIL . By inductive Claim 2, l has a parent, so p is not NIL . Now, suppose I occurs in TRYABSORBSIBLING . Then, r must be one of the nodes gp , p or l passed to TRYABSORBSIBLING , or a sibling s of l obtained from an $\text{LLX}(p)$. Before CLEANUP invokes TRYABSORBSIBLING it must execute one iteration of the loop at line 93, where it sets $gp := p$. It is straightforward to argue that l , p and gp are all non- NIL when the loop terminates. Thus, if $r \in \{gp, p, l\}$, then $r \neq \text{NIL}$. It remains to prove $s \neq \text{NIL}$. By inductive Claim 7, when I obtains s from the result of its $\text{LLX}(p)$, the tree is a relaxed (a, b) -tree, and p has at least two children. Thus, s exists (and is non- NIL). The proof for TRYABSORBCCHILD , TRYPROPAGATETAG and TRYDISTRIBUTE is similar. The proof when I occurs in TRYROOTUNTAG or TRYROOTABSORB is even simpler, since gp is *not* passed as an argument. Thus, we need only argue that l and p are non- NIL , which follows immediately from line 86.

Claim 2: The only step that can modify the tree (and, hence, affect this claim) is an invocation S of SCX performed by an invocation I of TRYINSERT , TRYDELETE , TRYROOTUNTAG , TRYROOTABSORB , TRYABSORBCCHILD , TRYPROPAGATETAG , TRYABSORBSIBLING or TRYDISTRIBUTE . Proving this claim entails arguing that I cannot replace the entry point, or modify it in any way except by changing its single child pointer.

Suppose the inductive hypothesis holds just before S . By inductive Claim 1, I follows the tree update template up until it performs S . By Lemma 5.5 and Lemma 5.1, the update phase of I atomically performs one of the transformations in Figure 8.1. All of these transformations simply change a single child pointer to replace one or more nodes. However, each node that is replaced has a parent, which entry does not. Thus, entry cannot be replaced.

Claim 3: Initially, the claim is trivially true (since the tree only contains entry , which is on every search path). In order for v to change from being on the search path for k in configuration C' to no longer being on the search path for k in configuration C , the tree must change between C' and C . Thus, there must be a successful $\text{SCX } S$ between C' and C . Moreover, this is the only kind of step that can affect this claim. We show S preserves the property that v is on the search path for k .

By inductive Claim 1, S is performed by a template operation. Thus, by Lemma 5.1, S changes a pointer of a node from old to new , removing a connected set R of nodes (rooted at old) from the tree, and inserting a new connected set N of nodes. If v is not a descendant of old immediately before S , then this change cannot remove v from the search path for k . So, suppose v is a descendant of old immediately prior to S .

Since v is in the data structure in both C' and C , it must be in the data structure at all times between C' and C by Lemma 5.3. Therefore, v is a descendant of old , but S does not remove v from the tree. Recall that the fringe F_R is the set of nodes that are children of nodes in R , but are not themselves in R (see Figure 5.1 and Figure 5.2). By definition, v must be a descendant of a node $f \in F_R$. Moreover, since v is on the search path for k just before S , so is f . We argue, for each possible tree modification in Figure 8.1, that if any node in F_R is on the search path for k prior to S , then it is still on the search path for k after S . We proceed by cases.

Case 1: Suppose S performs a INSERTPAIR, OVERFLOW or DELETEPAIR update. Since S replaces a leaf with either a new leaf, or a new internal node and two new leaves, the fringe set is empty. Thus, the claim is vacuously true.

Case 2: Suppose S performs a ROOTUNTAG or ROOTABSORB update. Then, S does not change the range of any node in the fringe set, so the claim holds.

Case 3: Suppose S performs a TRYABSORBCCHILD update. By inductive Claim 7, the tree is a relaxed (a, b) -tree before S . Since leaves are never tagged in a relaxed (a, b) -tree, the node u_i in the depiction of ABSORBCHILD in Figure 8.1 must be internal. Thus, one can think of each of the nodes u_i and u as a sequence of alternating pointers and keys, starting and ending with a pointer. Consequently, α , β and γ (in Figure 8.1) can be thought of as sequences of alternating pointers and keys, where α starts with a pointer and ends with a key, β starts and ends with pointers, and γ starts with a key and ends with a pointer. The fringe F_R is the set of nodes pointed to by α , β and γ . Observe that the keys in u_i and u partition the range of u , and this partition defines the range of each node in F_R . Specifically, the partition begins with the left endpoint of the range of u , then continues with the alternating pointers and keys of α , β and γ , and finally ends with the right endpoint of the range of u . The update does not change this partition, so the range of each node in F_R is the same before and after the update. Therefore, if a node in F_R is on the search path to key before the update, it is still on the search path after the update. The cases for PROPAGATETAG, ABSORBSIBLING and DISTRIBUTE follow the exact same reasoning.

Claim 4: Consider a read r of a child pointer in an invocation I of $\text{SEARCH}(k)$. This is the only kind of step that can affect this claim. Let v be the node pointed to by the value returned by r . We prove that r preserves the claim. If r is the first read of a child pointer by I , then v is *entry*, which is always on the search path for k , so r preserves the claim.

Now, suppose there is a previous read r' of a child pointer by I . By the inductive hypothesis, the node v' that was returned by r' was on the search path for k in some configuration C' after the beginning of I and before r' . Our goal is to prove that there is a configuration C , after C' and before r , when v is on the search path for k . If v' is in the tree when I performs r , then C is the configuration just before r . Otherwise, C is the last configuration before v' was removed from the tree.

Without loss of generality, suppose I reaches v by following the i th child pointer of v' . By inductive Claim 7, the data structure is a relaxed (a, b) -tree just before r , so $k \geq v'.k_j$ for $j < i$ and $k < v'.k_j$ for $j \geq i$. We now prove that $v'.p_i$ points to v in C . Suppose v' is in the tree when I performs r (so C is just before r). This case follows immediately from our assumption that r reads a pointer to v from $v'.p_i$. Now, suppose v' is not in the tree when r occurs, so C is the last configuration before v' was removed. Recall that v' is in the tree in C' . By Lemma 5.3, v' cannot be added back into the data structure after it is removed. Since v' is in the tree in C' , and is not in the tree when I subsequently performs r , v' must be removed after C' and before r occurs. By inductive Claim 1 and template Constraint 3, v' becomes finalized precisely when it is removed. Since v' cannot change after it is finalized, and $v'.p_i$ points to v when r occurs (which is after v' is removed), we can see that $v'.p_i$ must point to v in C (which is just before v' is removed).

Finally, we prove that v is on the search path for k in C . Since v' was on the search path for k in C' , and it is in the data structure in C , which is after C' but before r , inductive Claim 3 implies that v' is on the search path for k in C . Since $v = v'.p_i$, $k \geq v'.k_j$ for $j < i$ and $k < v'.k_j$ for $j \geq i$, we can see that v must also be on the search path for k in C .

Claim 5: Initially, the claim holds vacuously (since no steps have been taken). Suppose an invocation S of $\text{SEARCH}(k)$ in TRYINSERT terminates and returns m . (The proof for TRYDELETE is similar.) Consider any configuration C , after S returns m , in which all of the nodes in m are in the tree and their fields agree with the values in m . Suppose the inductive hypothesis up until C . We prove that an invocation S' of $\text{SEARCH}(k)$ in TRYINSERT would return m if S'

were performed atomically just after configuration C .

The value $m = \langle -, p, l \rangle$ returned by S contains a leaf l and its parent p . By inductive Claim 4, p and l were each on the search path at some point during S (which is before C). Since p and l are in the tree in C , inductive Claim 3 implies that they are on the search path for k in C . Therefore, S' will visit each of them. Conceptually, m also encodes the fact that p points to l . This fact is checked at line 24 as part of the CONFLICT procedure. By our assumption (that the fields of the nodes in m in configuration C agree with their values in m), p is also the parent of l when S' is performed. Consequently, S' will also return $m = \langle -, p, l \rangle$.

Claim 6: By inductive Claim 5 and Theorem 5.7, all invocations of TRYINSERT and TRYDELETE that perform a successful SCX are atomic. By Lemma 5.5, all invocations of TRYROOTUNTAG, TRYROOTABSORB, TRYABSORBCCHILD, TRYPROPAGATETAG, TRYABSORBSIBLING and TRYDISTRIBUTE that perform a successful SCX are atomic.

Claim 7: Only successful invocations of SCX can affect this claim. Successful invocations of SCX are performed only in TRYINSERT, TRYDELETE and the rebalancing procedures: TRYROOTUNTAG, TRYROOTABSORB, TRYABSORBCCHILD, TRYPROPAGATETAG, TRYABSORBSIBLING and TRYDISTRIBUTE. The claim holds in the initial state of the tree (which is described in Claim 2). We show that every successful invocation S of SCX preserves the claim. We proceed by cases.

Case 1: S is performed in an invocation I of TRYINSERT($key, value$) or TRYDELETE(key). By Claim 6, the entirety of I is atomic, including its search phase. Thus, when I occurs, its search procedure returns the unique leaf l on the search path for key . Consequently, I atomically performs one of the transformations DELETEPAIR, REPLACEPAIR, INSERTPAIR, or OVERFLOW in Figure 8.1 to replace l (and possibly some of its neighbouring nodes). Since I is entirely atomic (including its search phase), and it simply performs one of the (a, b) -tree updates, it is easy to verify that it preserves the claim.

Case 2: S is performed in an invocation I of one of the rebalancing procedures. By Claim 6, the update phase of I is atomic (but the search phase is not necessarily atomic). Therefore, I atomically performs one of the rebalancing transformations at some location in the tree (but not necessarily the same rebalancing transformation, at the same location, that it would perform if I 's search were also part of the atomic update). All of the rebalancing transformations preserve the claim (regardless of where in the tree they are performed). ■

We define the linearization points for relaxed (a, b) -tree operations as follows.

- GET(key) is linearized at a time during the operation when the leaf reached was on the search path for key . (This time exists, by Lemma 8.1.4.)
- An INSERT is linearized at its successful SCX inside TRYINSERT (if such an SCX exists). (Note: every INSERT that terminates performs a successful SCX.)
- A DELETE that returns \perp is linearized at a time during the operation when the leaf returned by its last invocation of SEARCH was on the search path for key . (This time exists, by Lemma 8.1.4.)
- A DELETE that does not return \perp is linearized at its successful SCX inside TRYDELETE (if such an SCX exists). (Note: every DELETE that terminates, but does not return \perp , performs a successful SCX.)

It is easy to verify that every operation that terminates is assigned a linearization point during the operation.

Theorem 8.2 *The relaxed (a, b) -tree is a linearizable implementation of a dictionary with the operations GET, INSERT and DELETE.*

Proof: Lemma 8.1.6 proves that the SCXs implement atomic changes to the tree as shown in Figure 8.1. By inspection of these transformations, the set of keys and associated values stored in leaves are not altered by any rebalancing steps. Moreover, the transformations performed by each linearized INSERT and DELETE maintain the invariant that the set of keys and associated values stored in leaves of the tree is exactly the set that should be in the dictionary. When an invocation of $\text{GET}(key)$ is linearized, the search path for key ends at the leaf returned by its invocation of SEARCH . If that leaf contains key , GET returns the associated value, which is correct. If that leaf does not contain key , then, by Lemma 8.1.7, it is nowhere else in the tree, so GET is correct to return \perp . ■

8.4 Progress proof

Our goal is to prove that, if processes take steps infinitely often, then relaxed (a,b) -tree operations succeed infinitely often. At a high level, this follows from Theorem 5.10 (the final progress result for template operations), and the fact that at most $(i+d)\lfloor \log_a(|T|+i)/2 \rfloor + 1$ rebalancing steps can be performed after i insertions and d deletions have been performed on a standard (a,b) -tree T (proved in [85]). Additionally, note that for $b \geq 2a$, only amortized $O(1)$ rebalancing steps are needed per insertion or deletion to maintain balance.

Theorem 8.3 *The relaxed (a,b) -tree operations are non-blocking.*

Proof: To derive a contradiction, suppose there is some configuration C after which some processes continue to take steps but no successful relaxed (a,b) -tree operations occur. We first argue that eventually the tree stops changing. Since no successful relaxed (a,b) -tree operations occur after C , the only steps that can change the tree after C are successful invocations of SCX performed by invocations of TRYROOTUNTAG, TRYROOTABSORB, TRYABSORBCCHILD, TRYPROPAGATETAG, TRYABSORBSIBLING or TRYDISTRIBUTE. Larsen and Fagerberg [85] proved that after a bounded number of rebalancing steps, a relaxed (a,b) -tree becomes a standard (a,b) -tree, and then no further rebalancing steps can be applied. Thus, eventually the tree must stop changing.

This implies that every invocation of SEARCH (and, hence, GET) succeeds after a finite number of steps, unless the process executing it crashes. It follows that no invocation of SEARCH or GET occurs after C . Therefore, eventually processes only take steps in INSERT and/or DELETE operations. Thus, processes perform infinitely many invocations of TRYINSERT and/or TRYDELETE, and/or infinitely many iterations of the outer loop in CLEANUP. By Lemma 8.1.1, TRYINSERT, TRYDELETE, and iterations of the outer loop in CLEANUP, all follow the template. Thus, Theorem 5.10 implies that infinitely many of these template updates will succeed. Since the number of rebalancing steps that can be performed is finite if the number of successful insertions and deletions is finite, there must be infinitely many successful insertions and/or deletions. Consequently, infinitely many must succeed after C , which is a contradiction. ■

8.5 Bounding the height of the tree

We now show that the height of the relaxed (a,b) -tree at any time is $O(c + \log_a n)$ where n is the number of keys stored in the tree and c is the number of INSERT and DELETE operations *currently* in progress. (When no process is performing INSERT or DELETE , the c term disappears.) At a high level, the upper bound holds for the following reason. Since we always perform rebalancing steps that satisfy VIOL, if we reach a leaf without finding the violation

that an INSERT or DELETE created, then the violation has been eliminated. Since each INSERT or DELETE creates at most one violation, and eliminates it before terminating, we can prove that the number of violations in the tree at any time is bounded above by c . Further, since removing all violations would yield an (a,b) -tree tree with height $O(\log_a n)$, and eliminating each violation reduces the height by *at most* one, the height of the relaxed (a,b) -tree is $O(c + \log_a n)$.

Note that this is a very pessimistic upper bound, because many violations do not increase the height of the tree. In our experiments, the height is typically approximately $1 + \log_a n$. To the authors' knowledge, a tree of height $c + \log_a n$ is only achieved in a specific pathology where all processes cooperate on the same path to cause n consecutive tag violations (by having one process perform OVERFLOW at a node, then a second process perform OVERFLOW at one of the leaves created by the first OVERFLOW, then a third process perform OVERFLOW at one of the leaves created by the second OVERFLOW, and so on). Observe that, to cause an OVERFLOW at a leaf that was created by another OVERFLOW, processes must insert $b - a + 1$ keys at the leaf. Additionally, no operation that performs an OVERFLOW update can terminate until all processes have finished constructing this pathology, or else the operation will perform rebalancing steps to eliminate the tag violations.

Definition 8.4 Let x be a node that is in the data structure. We say that a tag violation occurs at x if x has its tag bit set. Suppose x is not entry, and no tag violation occurs at x . We say that a **degree violation occurs at x** if (1) x is root (i.e., $\text{entry}.p_1$) and x is an internal node with one pointer, or (2) x is not root and x has fewer than a pointers.

Definition 8.5 A process P is **in a cleanup phase for k** if it is executing an $\text{INSERT}(k, \text{value})$ or a $\text{DELETE}(k)$ and it has performed a successful SCX inside a TRYINSERT or TRYDELETE that returns $\text{createdViolation} = \text{TRUE}$. If P is between line 86 and 117, then $\text{location}(P)$ is the value of P 's local variable l ; otherwise, $\text{location}(P)$ is the entry node.

We use the following invariant to show that each violation in the data structure has a pending update operation that is responsible for removing it before terminating: either that process is on the way towards the violation, or it will find another violation and restart from the top of the tree, heading towards the violation.

Lemma 8.6 In every configuration, there exists an injective mapping ρ from violations to processes such that, for every violation x ,

- (A) process $\rho(x)$ is in a cleanup phase for some key k_x and
- (B) x is on the search path for k_x from root (i.e., $\text{entry}.p_1$) and
- (C) either
 - (C1) the search path for k_x from $\text{location}(\rho(x))$ contains the violation x , or
 - (C2) in the prefix of the search path for k_x from $\text{location}(\rho(x))$ up to and including the first non-finalized node (or the entire search path if all nodes are finalized), there is a node where a violation occurs.

Proof: In the initial configuration, there are no violations, so the invariant is trivially satisfied. We show that any step S by any process P preserves the invariant. We assume there is a function ρ satisfying the claim for the configuration C immediately before S and show that there is a function ρ' satisfying the claim for the configuration C' immediately after S . The only step that can cause a process to leave its cleanup phase is the termination of an INSERT or DELETE that is in its cleanup phase. The only steps that can change $\text{location}(P)$ and $\text{parent}(P)$ are P 's execution of line 86 or the read of the child pointer on line 98. (We think of all of the updates to local variables on those lines as happening

atomically with the read of the child pointer.) The only steps that can change child pointers or finalize nodes are successful SCXs. No other steps S can cause the invariant to become false.

Case 1 S is the termination of an INSERT or DELETE that is in its cleanup phase: We choose $\rho' = \rho$. S happens when the test in line 94 is true, meaning that $location(P)$ is a leaf. Leaves are never tagged. There cannot be a degree violation at the leaf, because then the process would have exited the loop in the previous iteration after the test at line 99 returned true (since set of pointers/values in a leaf never changes). Thus, no violation occurs at $location(P)$. So, P cannot be $\rho(x)$ for any violation x , so S cannot make the invariant become false.

Case 2 S is an execution of line 86: We choose $\rho' = \rho$. Step S changes $location(P)$ to $entry.p_1$ and $parent(P)$ to $entry$. If $P \neq \rho(x)$ for any violation x , then this step cannot affect the truth of the invariant. Now suppose $P = \rho(x_0)$ for some violation x_0 . The truth of properties (A) and (B) are not affected by a change in $location(P)$ and property (C) is not affected for any violation $x \neq x_0$. Since ρ satisfies property (B) for violation x_0 before S , it will satisfy property (C1) for x_0 after S .

Case 3 S is a read of the child pointer on line 98: We choose $\rho' = \rho$. Step S changes $location(P)$ from some node l to node l_i , which is l 's i th child when S is performed. If $P \neq \rho(x)$ for any violation x , then this step cannot affect the truth of the invariant. So, suppose $P = \rho(x_0)$ for some violation x_0 . By (A), P is in a cleanup phase for k_{x_0} . The truth of (A) and (B) are not affected by a change in $location(P)$ and property (C) is not affected for any violation $x \neq x_0$. So it remains to prove that (C) is true for violation x_0 in C' .

First, we prove there is no violation at l . Suppose S occurs in the first iteration of CLEANUP's inner loop. Then, l was *root* when it was read from $entry.p_1$ at line 86. Before entering the inner loop, CLEANUP saw that l was untagged (so there is no tag violation at l), and was not an internal node with one pointer (so there is no degree violation at l). Thus, there is no violation at l in this case. Now, suppose S does not occur in the first iteration of CLEANUP's inner loop. In the previous iteration, CLEANUP saw $l.tag = 0$ and $l.d \geq a$ at line 99, otherwise the loop would have terminated. So, there is no violation at l .

We consider two cases, depending on whether (C1) or (C2) is true in configuration C .

Case 3a (C1) is true in configuration C : Thus, when S is performed, the violation x_0 is on the search path for k_{x_0} from l , but it is not at l (as argued above). S reads the i th child of l , so $k \geq l.k_j$ for $j < i$ and $k < l.k_j$ for $j \geq i$ (since the keys of node l never change). So, x_0 must be on the search path for k_{x_0} from l_i . This means (C1) is satisfied for x_0 in configuration C' .

Case 3b (C2) is true in configuration C : This proof is similar to the previous case.

Case 4 S is a successful SCX: We must define the mapping ρ' for each violation x in configuration C' . Since tag bits are immutable and the number of pointers in nodes do not change, no transformation in Figure 8.1 can create a new violation at a node that was already in the data structure in configuration C . So, if x is at a node that was in the data structure in configuration C , x was a violation in configuration C , and $\rho(x)$ is well-defined. In this case, we let $\rho'(x) = \rho(x)$.

If x is at a node that was added to the data structure by S , then we must define $\rho(x)$ on a case-by-case basis for all transformations described in Figure 8.1. If x is a tag violation at a newly added node, we define $\rho'(x)$ according to the table in Figure 8.8. If x is a degree violation at a newly added node, we define $\rho'(x)$ according to the table in Figure 8.9.

The function ρ' is injective, since ρ' maps each violation created by S to a distinct process that ρ assigned to a violation that has been removed by S , with only two exceptions: for tag violations caused by OVERFLOW and degree violation caused by DELETEPAIR. In these exceptional cases, ρ' maps these violations to the process that has just

Transformation	Tag violations x created by S	$\rho'(x)$
DELETEPAIR	none created	—
REPLACEPAIR	none created	—
INSERTPAIR	none created	—
OVERFLOW	at n	process performing the INSERT
ROOTUNTAG	none created	—
ROOTABSORB	none created	—
ABSORBCCHILD	none created	—
PROPAGATETAG	at n	$\rho(\text{tag violation at } u_i)$
ABSORBSIBLING	none created	—
DISTRIBUTE	none created	—

Figure 8.8: Description of how ρ' maps tag violations at newly added nodes to processes responsible for them.

Transformation	Degree violations x created by S	$\rho'(x)$
DELETEPAIR	at n (if $u.d = a$)	process performing the DELETE
DELETEPAIR	at n (if $u.d < a$)	$\rho(\text{degree violation at } u)$
REPLACEPAIR	none created	—
INSERTPAIR	none created	—
OVERFLOW	none created	—
ROOTUNTAG	at n (if root is internal and $\text{root}.d = 1$)	$\rho(\text{tag violation at } u)$
ROOTABSORB	at n (if root_1 is internal and $\text{root}_1.d = 1$)	$\rho(\text{degree violation at } \text{root}_1)$
ABSORBCCHILD	at n (if $u.d \leq a - 2$)	$\rho(\text{degree violation at } u)$
PROPAGATETAG	none created	—
ABSORBSIBLING	at n_i (if $u_i.d + u_{i+1}.d < a$)	$\rho(\text{degree violation at } u_i)$
ABSORBSIBLING	at n (if $u.d = a$ and $u_i.d + u_{i+1}.d < a$)	$\rho(\text{degree violation at } u_{i+1})$
ABSORBSIBLING	at n (if $u.d = a$ and $u_i.d + u_{i+1}.d \geq a$)	$\rho(\text{degree violation at } u_i \text{ or } u_{i+1})$
ABSORBSIBLING	at n (if $u.d < a$)	$\rho(\text{degree violation at } u)$
DISTRIBUTE	at n (if $u.d < a$)	$\rho(\text{degree violation at } u)$

Figure 8.9: Description of how ρ' maps degree violations at newly added nodes to processes responsible for them. Here, $u.d$ denotes the degree of a node u .

begun its cleanup phase (and therefore was not assigned any violation by ρ).

Let x be any violation in the tree in configuration C' . We show that ρ' satisfies properties (A), (B) and (C) for x in configuration C' .

Property (A): Every process in the image of ρ' was either in the image of ρ or a process that just entered its cleanup phase at step S , so every process in the image of ρ' is in its cleanup phase.

Property (B) and (C): We consider several subcases.

Subcase 4a Suppose S is an OVERFLOW's SCX, and x is the tag violation created by S . Then, P is in its cleanup phase for the inserted key, which is one of the children of the node containing the tag violation x . Since the tree is a search tree, x is on the search path for this key, so (B) holds. In this subcase, $\text{location}(\rho'(x)) = \text{entry}$ since $P = \rho'(x)$ has just entered its cleanup phase. So property (B) implies property (C1).

Subcase 4b Suppose S is a DELETEPAIR's SCX, and x is the degree violation assigned to P by ρ' . Then, P is in a cleanup phase for the deleted key, which was in u before S . Therefore, x (at the replacement leaf n) is on the search

path for this key, so (B) holds. As in the previous subcase, $\text{location}(\rho'(x)) = \text{entry}$ since $P = \rho'(x)$ has just entered its cleanup phase. So property (B) implies property (C1).

Subcase 4c If x is at a node that was added to the data structure by S (and is not covered by the above two cases), then $\rho'(x)$ is $\rho(y)$ for some violation y that has been removed from the tree by S , as described in the above two tables. Let k be the key such that process $\rho(y) = \rho'(x)$ is in the cleanup phase for k . By property (B), y was on the search path for k before S . By inspection of the tables, and Lemma 8.1.3, any search path that went through y 's node in configuration C goes through x 's node in configuration C' . (We designed the tables to have this property.) Thus, since y was on the search path for k in configuration C , x is on the search path for k in configuration C' , satisfying property (B).

If (C2) is true for violation y in configuration C , then (C2) is true for x in configuration C' (since any node that is finalized remains finalized forever, and its child pointers do not change).

So, for the remainder of the proof of subcase 4c, suppose (C1) is true for y in configuration C . Let $l = \text{location}(\rho(y))$ in configuration C . Then y is on the search path for k from l in configuration C .

First, suppose S removes l from the data structure.

- If y is a violation at node l in configuration C , then it makes (C2) true for x in configuration C' .
- Otherwise, since both l and its descendant, the parent of the node that contains y , are removed by S , the entire path between these two nodes is removed from the data structure by S . So, all nodes along this path are finalized by S because Constraint 3 is satisfied. Thus, the violation y makes (C2) true for x in configuration C' .

Now, suppose S does not remove l from the data structure. In configuration C , the search path from l for k contains y . By inspection of the tables defining ρ' , and Lemma 8.1.3, any search path from l that went through y 's node in configuration C goes through x 's node in configuration C' . So, (C1) is true in configuration C' .

Subcase 4d If x is at a node that was already in the data structure in configuration C , then $\rho'(x) = \rho(x)$. Let k be the key such that this process is in the cleanup phase for k . Since x was on the search path for k in configuration C and S did not remove x from the data structure, x is still on the search path for k in configuration C' (by Lemma 8.1.3). This establishes property (B).

If (C2) is true for x in configuration C , then it also holds for x in C' , for the same reason as in Subcase 4c. So, suppose (C1) is true for x in configuration C . Let $l = \text{location}(\rho(x))$ in configuration C . Then, (C1) says that x is on the search path for k from l in configuration C . If S does not change any of the child pointers on this path between l and x , then x is still on the search path from $\text{location}(\rho'(x)) = l$ in configuration C' , so property (C1) holds for x in C' . So, suppose S does change the child pointer of some node on this path from old to new . Then the search path for k from l in configuration C goes through old to some node f in the fringe set F_R of S and then onward to the node containing violation x . By Lemma 8.1.3, the search path for k from l in configuration C' goes through new to the same node f , and then onward to the node containing the violation x . Thus, property (C1) is true for x in configuration C' . ■

Corollary 8.7 *The number of violations in the data structure is bounded by the number of incomplete INSERT and DELETE operations.*

Lemma 8.8 *Consider a relaxed (a,b) -tree T rooted that contains n nodes and c violations. Suppose T' is any (a,b) -tree that results from performing sufficient rotations on T to eliminate all violations. Then, the following claims hold for any leaf u in T , and any leaf $u' \in T'$.*

1. $l(u) \leq rl(u) + c$

$$2. \ rl(u) \leq rl(u') + c$$

$$3. \ rl(u') = l(u')$$

Proof: **Claim 1:** Immediate from the definitions of l and rl .

Claim 2: The rebalancing steps in a relaxed (a,b) -tree maintain the invariant that all leaves have the same relaxed level. Thus, every pair of leaves u_1 and u_2 in T satisfy $rl(u_1) = rl(u_2)$. The only time the relaxed level of a leaf changes is when a ROOTUNTAG update changes the *tag* bit of the root from one to zero, uniformly incrementing the relaxed level of every leaf by one. ROOTUNTAG also decreases the number of violations in the tree by one. Thus, eliminating c violations can increase the relaxed level of every leaf by at most c . The claim then follows from the fact that T' is produced by performing rotations to eliminate c violations.

Claim 3: Immediate from definitions of l and rl , and the fact that, since T' is an (a,b) -tree, no node is tagged. ■

Corollary 8.9 *Our implementation of a relaxed (a,b) -tree containing n keys has height $O(c + \log_a n)$, where c is the number of incomplete INSERT and DELETE operations.*

Proof: Let r , T , r' and T' be defined as in Lemma 8.8. By Corollary 8.7, T contains at most c violations. Thus, we immediately obtain $l(r) \leq l(r') + 2c$ from Lemma 8.8. Since T' is an (a,b) -tree, $l(r') \in O(\log_a n)$. Therefore, $l(r) \in O(c + \log_a n)$. ■

8.6 Experiments

We implemented the relaxed (a,b) -tree in C++, using a fast memory reclamation scheme called DEBRA that is described in Chapter 11. Since this implementation was done in C++, a direct comparison with the Java implementation of the chromatic tree described earlier would not make sense. (Any performance differences due to the algorithms would be conflated with the differences between C++ and Java.) So, for comparison, we also implemented an unbalanced version of the chromatic tree (BST) in C++, which simply does not perform any rebalancing steps.² BST used the same memory reclamation scheme as the relaxed (a,b) -tree.

We ran a small set of experiments on an Intel E7-4830 v3 with 12 cores and 2 hyperthreads per core, for a total of 24 hardware contexts. Each core has a private 32KB L1 cache and 256KB L2 cache (which is shared between hyperthreads on a core). All cores share a 30MB L3 cache. The machine has 128GB of RAM and runs Ubuntu 14.04 LTS.

All code was compiled with the GNU C++ compiler (G++) 4.8.4 with build target x86_64-linux-gnu and compilation options `-std=c++0x -mcx16 -O3`. Thread support was provided by the POSIX Threads library. We used the scalable allocator jemalloc 4.2.1 [54], which greatly improved performance for both BST and the relaxed (a,b) -tree.

We pinned one thread to each hardware context. For thread counts up to 12, we pinned at most one thread to each core, so hyperthreading is engaged only for thread counts 13 through 24.

In our experiments, we fixed the parameters for the (a,b) -tree at $a = 6$ and $b = 16$. With $b = 16$, each node occupies four consecutive cache lines. Since (a,b) -trees require $b \geq 2a - 1$, with $b = 16$, we must have $a \leq 8$. We

²We subsequently implemented the full chromatic tree in C++, but we did not have that implementation when these experiments were run. Subsequent experiments with random insertions and deletions of uniformly random keys indicated that the performance of BST is the same, or slightly better than that of the chromatic tree, for the workloads shown here. This is because random operations on uniform keys yield a fairly balanced tree, so BST enjoys the benefits of balance without paying the overhead of performing rebalancing steps.

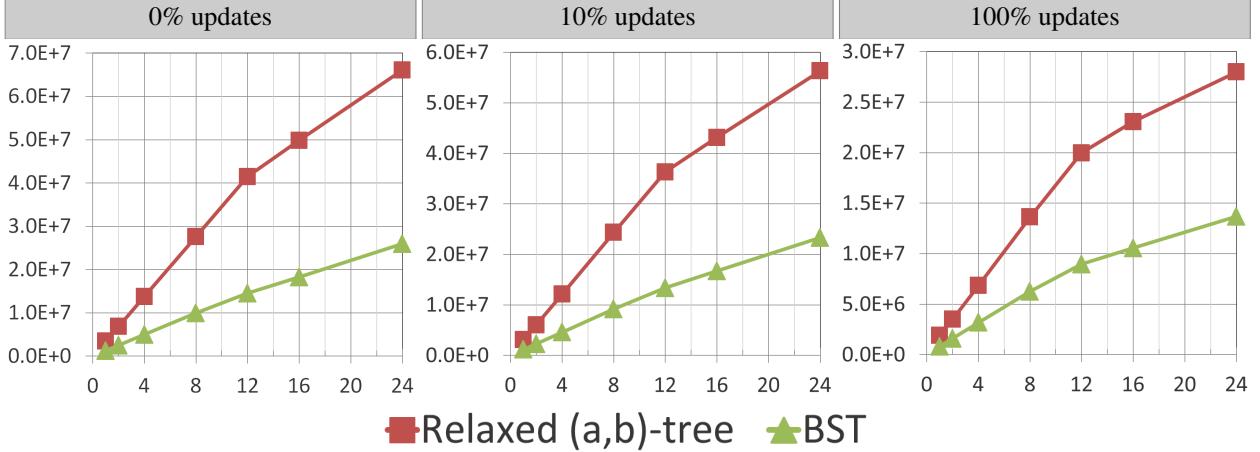


Figure 8.10: Performance results for a microbenchmark comparing BST and the relaxed (a, b) -tree. The x-axis represents the number of concurrent threads. The y-axis represents operations per microsecond.

chose to make a slightly smaller than 8 in order to exploit a performance tradeoff: a smaller minimum degree slightly increases depth, but decreases the number of degree violations that are created in an execution. With $a = 6$, after a leaf is created by an OVERFLOW update (containing $b/2 = 8$ keys), three keys must be deleted from the leaf (with no intervening insertions into the leaf) before a degree violation is created. (In contrast, with $a = 8$, the first deletion from a leaf created by OVERFLOW causes a degree violation, and triggers rebalancing.)

We compared the performance of BST and the relaxed (a, b) -tree using a simple randomized microbenchmark. For each algorithm $A \in \{\text{BST, relaxed } (a, b)\text{-tree}\}$ and update rate $U \in \{100, 10, 0\}$, we ran five timed *trials* for several thread counts n . Each trial proceeded in two phases: *prefilling* and *measuring*. In the prefilling phase, n concurrent threads performed 50% *Insert* and 50% *Delete* operations on keys drawn uniformly randomly from $[0, 10^6]$ until the size of the tree converged to a steady state (containing approximately $10^6/2$ keys). Next, the trial entered the measuring phase, during which threads began counting how many operations they performed. (These counts were eventually summed over all threads and reported in our graphs.) In this phase, each thread performed $(U/2)\%$ *Insert*, $(U/2)\%$ *Delete* and $(100 - U)\%$ *Find* operations on keys drawn uniformly from $[0, 10^6]$ for ten seconds.

As a way of validating correctness in each trial, each thread maintained a *checksum*. Each time a thread inserted a new key, it added the key to its checksum. Each time a thread deleted a key, it subtracted the key from its checksum. Observe that, at the end of any correct trial, the sum of all thread checksums must be equal to the sum of keys in the tree.

The results for this benchmark appear in Figure 8.10. In these experiments, the key range is fairly large, and the relaxed (a, b) -tree has a significant performance advantage over BST. With 100% updates, it approximately doubles the performance of the BST, and the performance gap grows wider as the proportion of operations that are searches increases. This performance advantage comes from the smaller depth and improved cache locality offered by the fat nodes in the (a, b) -tree.

We briefly explain why cache performance is better in a tree with fat nodes than in a BST. In a modern system, whenever a process reads a memory address, the cache coherence protocol actually reads the entire cache line containing the address (typically 64 bytes) and stores it in the cache. In fact, in recent Intel systems, there is a cache *prefetcher* that additionally fetches an *adjacent* cache line, as well as a *streamer* that may fetch even more subsequent,

consecutive cache lines (up to an entire page of memory). Thus, each read from memory loads *at least* two full cache lines from memory (and often three or four).

For example, reading a key in a BST node might also cause the value and child pointers to be loaded into the cache at the same time, if they happen to reside in the same cache line (or the adjacent one). In this particular experiment, a BST node was 40 bytes, so each time a node was accessed, at least $128 - 40 = 88$ bytes of unrelated information was also loaded into the cache. This unrelated information clutters the cache, negatively impacting its utilization. Furthermore, when a process is searching in the tree, each time a pair of cache lines are retrieved from memory, they likely contain only one relevant key, and the next relevant key cannot be accessed until more cache lines are loaded, which introduces significant latency. This is sometimes called the *pointer chasing* problem.

In contrast, fat nodes *harness* prefetching behaviour by storing several cache lines worth of relevant information consecutively. Consequently, cache utilization is improved, and accessing a key has the side effect of loading several other relevant keys and/or pointers from memory, at the same time (reducing the impact of pointer chasing).

Preliminary experiments suggest that there is an even larger performance gap with larger trees, and that the relaxed (a,b) -tree remains competitive with the BST even when the key range is two orders of magnitude smaller than in the workload shown here. In light of these results, future code libraries should consider including concurrent data structures with fat nodes, like the relaxed (a,b) -tree.

Chapter 9

B-slack trees: space efficient B-trees

B-trees are balanced trees designed for block-based storage media. Internal nodes contain between $b/2$ and b child pointers, and one less key. Leaves contain between $b/2$ and b keys. All leaves have the same depth, so access times are predictable. If memory can be allocated on a per-byte basis, nodes can simply be allocated the precise amount of space they need to store their data, and no space is wasted. However, typically, all nodes have the same, fixed capacity, and some of the capacity of nodes is wasted. As much as 50% of the capacity of each node is wasted in the worst case. This is particularly problematic when data structures are being implemented in hardware, since memory allocation and reclamation schemes are often very simplistic, allowing only a single block size to be allocated (to avoid fragmentation). Furthermore, since hardware devices must include sufficient resources to handle the worst-case, good expected behaviour is not enough to allow hardware developers to reduce the amount of memory included in their devices. To address this problem, we introduce *B-slack trees*, which are a variant of B-trees with substantially better worst-case space complexity. We also introduce relaxed B-slack trees, which are a variant of B-slack trees that are more amenable to concurrent implementation.

The development of B-slack trees was inspired by a collaboration with a manufacturer of internet routers, who wanted to build a concurrent router based on a tree. In such embedded devices, memory is limited, so it is important to use it as efficiently as possible. A suitable tree would have a simple algorithm for updates, small space complexity, fast searches, and searches that would not be blocked by concurrent updates. Updates were expected to be infrequent. One naive approach is to rebuild the entire tree after each update. Keeping an old copy of the tree while rebuilding a new copy would allow searches to proceed unhindered, but this would double the space required to store the tree.

B-slack trees are leaf-oriented trees with many desirable properties. The average degree of nodes is high, exceeding $b - 2$ for trees of height at least three. Their space complexity is better than all of their competitors. Consider a dictionary implemented by a leaf-oriented search tree, in which, along with each key, a leaf stores a pointer to associated data. Suppose that each key and each pointer to a child or to data occupies a single word. Then, $\frac{2b}{b-3}n$ is an upper bound on the number of words needed to store a B-slack tree with $n > b^3$ keys. For large b , this tends to $2n$, which is optimal. Section 9.3 gives a more complex upper bound, which is much better when b is small. B-slack trees have logarithmic height, and the number of rebalancing steps performed after a sequence of m updates to a B-slack tree of size n is amortized $O(\log(n+m))$ per update. Furthermore, the number of rebalancing steps needed to rebalance the tree can be reduced to amortized *constant* per update at the cost of slightly increased space complexity.

The rest of this paper is organized as follows. Section 9.1 surveys related work. Section 9.2 introduces B-slack trees and relaxed B-slack trees. Height, average degree, space complexity and rebalancing costs of relaxed B-slack trees (and, hence, of B-slack trees) are analyzed in Section 9.3. Section 9.4 describes the variant of B-slack trees with amortized constant rebalancing. Section 9.5 compares the worst-case space complexity of B-slack trees to competing data structures. Single-threaded experimental results appear in Section 9.6. Section 9.7 discusses some implementation issues surrounding rebalancing. Section 10.1 describes an approach for obtaining a lock-free implementation of a B-slack tree using the template described in Chapter 5.

9.1 Related work

B-trees were initially proposed by Bayer and McCreight in 1970 [18]. Insertion into a full node in a B-tree causes it to split into two nodes, each half full. Deletion from a half-full node causes it to merge with a neighbour. Arnow, Tenenbaum and Wu proposed P-trees [12], which enjoy moderate improvements to *average* space complexity over B-trees, but waste 66% of each node in the worst case.

A number of generalizations of B-trees have been suggested that achieve much less waste if no deletions are performed. Bayer and McCreight also proposed B*-trees in [18], which improve upon the worst-case space complexity of B-trees. At most a third of the capacity of each node in a B*-tree is wasted. This is achieved by splitting a node only when it and one of its neighbours are both full, replacing these two nodes by three nodes. Küspert [80] generalized B*-trees to trees where each node contains between $\lfloor \frac{bm}{m+1} \rfloor$ and b pointers or keys, where $m \leq b - 1$ is a design parameter. Such a tree behaves just like a B*-tree everywhere except at the leaves. An insertion into a full leaf causes keys to be shifted among the nearest $m - 1$ siblings to make room for the inserted key. If the $m - 1$ nearest siblings are also full, then these m nodes are replaced by $m + 1$ nodes which evenly share keys. Large values of m yield good worst-case space complexity. However, with large m , updates would be complicated and inefficient in a concurrent setting.

Baeza-Yates and Per-åke Larson introduced B+trees with partial expansions [15]. Several node sizes are used, each a multiple of the block size. An insertion to a full node causes it to expand to the next larger node size. With three node sizes, at most 33% of each node can be wasted, and worst-case utilization improves with the number of block sizes used. However, this technique simply pushes the complexity of the problem onto the memory allocator. Memory allocation is relatively simple for one block size, but it quickly becomes impractical for simple hardware to support larger numbers of block sizes.

Culik, Ottmann and Wood introduced strongly dense multiway trees (SDM-trees) [39]. An SDM-tree is a node-oriented tree in which all leaves have the same depth, and the root contains at least two pointers. Apart from the root, every internal node u with fewer than b pointers has at least one sibling. Each sibling of u has b pointers if it is an internal node and b keys if it is a leaf. Insertion can be done in $O(b^3 + (\log n)^{b-2})$ time. Deletion is not supported, but the authors mention that the insertion algorithm could be modified to obtain a deletion algorithm, and the time complexity of the resulting algorithm “would be at most $O(n)$ and at least $O((\log n)^{b-1})$. ” Besides the long running times for each operation (and the lack of better amortized results), the insertion algorithm is very complex and involves many nodes, which makes it poorly suited for hardware implementation. Furthermore, in a concurrent setting, an extremely large section of the tree would have to be modified atomically, which would severely limit concurrency.

Srinivasan introduced a leaf-oriented B-tree variant called an *Overflow tree* [116]. For each parent of a leaf, its children are divided into one or more groups, and an overflow node is associated with each group. The tree satisfies the B-tree properties and the additional requirement that each leaf contains at least $b - 1 - s$ keys, where $s \geq 2$ is a design parameter and b is the maximum degree of nodes. Inserting a key into a full leaf causes the key to be inserted into the overflow node instead; if the overflow node is full, the entire group is reorganized. Deleting from a leaf is the same as in a B-tree unless it will cause the leaf to contain too few keys, in which case, a key is taken from the overflow node; if a key cannot be taken from the overflow node, the entire group is reorganized. Each search must look at an overflow node. The need to atomically modify and search two places at once makes this data structure poorly suited for concurrent implementation.

Hsuang introduced a class of node-oriented trees called *H-trees* [71], which are a subclass of B-trees parameterized by γ and δ . These parameters specify a lower bound on the number of grandchildren of each internal node (that has grandchildren), and a lower bound on the number of keys contained in each leaf, respectively. Larger values of δ and γ yield trees that use memory more efficiently. When δ and γ are as large as possible, each leaf contains at least $b - 3$ keys, and each internal node has zero or at least $\lfloor \frac{b^2+1}{2} \rfloor$ grandchildren. The paper presents $O(\log n)$ insertion and deletion algorithms for node-oriented H-trees. The algorithms are very complex and involve many cases. H-trees have a minimum average degree of approximately $b/\sqrt{2}$ for internal nodes, which is much smaller than the $b - 2$ of B-slack trees (for trees of height at least three).

Section 9.5 describes families of B-trees, H-trees and Overflow trees which require significantly more space than B-slack trees.

Rosenberg and Snyder introduced *compact B-trees* [110], which can be constructed from a set of keys using the minimum number of nodes possible. No compactness preserving insertion or deletion procedures are known. The authors suggested using regular B-tree updates and periodically compacting a data structure to improve efficiency. However, experiments in [11] showed that starting with a compact B-tree and adding only 1.6% more keys using standard B-tree operations reduced storage utilization from 99% to 67%. Thus, to maintain reasonable space complexity, the expensive tree rebuilding/compaction algorithm would have to be executed prohibitively often.

An impressive paper by Brönnimann et al. [26] presented three ways to transform an arbitrary sequential dictionary into a more space efficient one. One of these ways will be discussed here; of the other two, one is extremely complex and poorly suited for concurrent hardware implementation, and the other pushes the complexity onto the memory allocator.

This transformation takes any sequential tree data structure and modifies it by replacing each key in the sequential data structure with a *chunk*, which is a group of $b - 2$, $b - 1$ or b keys, where b is the memory block size. All chunks in the data structure are also kept in a doubly linked list to facilitate iteration and movement of keys between chunks. For instance, a BST would be transformed into a tree in which each node has zero, one or two children, and $b - 2$, $b - 1$ or b keys. All keys in chunks in the left subtree of a node u would be smaller than all keys in u 's chunk, and all keys in chunks in the right subtree of u would be larger than all keys in u 's chunk. A search for key k behaves the same as in the sequential data structure until it reaches the only chunk that can contain k , and searches for k within the chunk. An insertion first searches for the appropriate chunk, then it inserts the key into this chunk. Inserting into a full chunk requires shifting the keys of the b closest neighbouring chunks to make room. If these b chunks are full, then b consecutive keys from these chunks are first placed in a new chunk, and the remaining $b(b - 1)$ keys are distributed amongst the original b chunks. Deletion is similar. Each operation in the resulting data structure runs in $O(f(n) + b^2)$ steps, where $f(n)$ is the number of steps taken by the sequential data structure to perform the same operation.

After this transformation, a B-tree with maximum degree b requires $2n + O(n/b)$ words to store n keys and pointers to data. In the worst-case, each chunk wastes $2/b$ of its space, which is somewhat worse than in B-slack trees. Furthermore, supporting fast searches can introduce significant complexity to the hardware design. Suppose this transformation is applied to a B-tree. Then a node in the transformed B-tree contains up to $b - 1$ chunks, each of which contains $b - 2$, $b - 1$ or b keys, and occupies one block of memory. Thus in order to determine which child pointer should be followed, a search must load up to $b - 1$ blocks. In order for searches to be fast, hardware must therefore be able to quickly load up to $b - 1$ blocks from memory (a total of $\theta(b^2)$ memory words). This represents a significant challenge for hardware design.

As we saw in Chapter 8, B-trees with relaxed balance have previously been proposed. In particular, several of the updates to B-slack trees are derivative of updates to the relaxed (a, b) -tree. However, limited work has been done on improving the space complexity of relaxed balance data structures. Larsen and Fagerberg improved the space complexity of the relaxed (a, b) -tree *while the tree is out of balance* [73], but its worst-case space complexity remains unchanged.

9.2 B-slack trees

A B-slack tree is a variant of a B-tree. Each node stores its keys in sorted order, so binary search can be used to determine which child of an internal node should be visited next by a search, or whether a leaf contains a key. Let p_0, p_1, \dots, p_m be the sequence of pointers contained in an internal node, and k_1, k_2, \dots, k_m be its sequence of keys. For each $1 \leq i \leq m$, the subtree pointed to by p_{i-1} contains keys strictly smaller than k_i , and the subtree pointed to by p_i contains keys greater than *or equal to* k_i . We say that the *degree of an internal node* is the number of non-NIL pointers it contains, and the *degree of a leaf* is the number of keys it contains. This unusual definition of degree simplifies our discussion. The degree of node v is denoted $\deg(v)$. If the maximum possible degree of a node is b , and its degree is $b - x$, then we say it contains x units of *slack* (or simply x *slack*).

A B-slack tree is a leaf-oriented search tree with maximum degree $b > 4$ in which:

- P1:** every leaf has the same depth,
- P2:** internal nodes contain between 2 and b pointers (and one less key),
- P3:** leaves contain between 0 and b keys, and
- P4:** for each internal node u , the total slack contained in the *children* of u is at most $b - 1$.

P4 is the key property that distinguishes B-slack trees from other variants of B-trees. It limits the aggregate space wasted by a number of nodes, as opposed to limiting the space wasted by each node. Alternatively, P4 can be thought of as a lower bound on the sum of the degrees of the children of each internal node. Formally, for each internal node with children v_1, v_2, \dots, v_l , $\deg(v_1) + \deg(v_2) + \dots + \deg(v_l) \geq lb - (b - 1) = lb - b + 1$. This interpretation is useful to show that all nodes have large subtrees. For instance, it implies that a node u with two internal children must have at least $b + 1$ grandchildren. If these grandchildren are also internal nodes, we can conclude that u must have at least $b^2 - b + 2$ great grandchildren.

A tree that satisfies P1, and in which every node has degree $b - 1$, is an example of a B-slack tree. Another example of a B-slack tree is a tree of height two, where b is even, the root has degree two, its two children have degree $b/2$ and $b/2 + 1$, respectively, and the grandchildren of the root are leaves with degree b , except for two, one in the left subtree of the root, and one in the right subtree, that each have degree one. This tree contains the smallest number of keys of any B-slack tree of height two.

9.2.1 Relaxed B-slack trees

A relaxed balance search tree decouples updates that rebalance (or reorganize the keys of) the tree from updates that modify the set of keys stored in the tree [87]. The advantages of this decoupling are twofold. First, updates to a relaxed balance version of a search tree are smaller, so a greater degree of concurrency is possible in a multithreaded setting. Second, for some applications, it may be useful to temporarily disable rebalancing to allow a large number of updates to be performed quickly, and to gradually rebalance the tree afterwards.

A relaxed B-slack tree is a relaxed balance version of a B-slack tree that has weakened the properties. A weight of zero or one is associated with each node. These weights serve a purpose similar to the colors red and black in a red-black tree. We define the *relaxed depth* of a node to be one less than the sum of the weights on the path from the root to this node. A relaxed B-slack tree is a leaf-oriented search tree with maximum degree $b > 4$ in which:

- P0':** every node with weight zero contains exactly two pointers,
- P1':** every leaf has the same relaxed depth,
- P2':** internal nodes contain between 1 and b pointers (and one less key), and

P3 : leaves contain between 0 and b keys

To clarify the difference between B-slack trees and relaxed B-slack trees, we identify several types of *violations* of the B-slack trees properties that can be present in a relaxed B-slack tree. We say that a *weight violation* occurs at a node with weight zero, a *slack violation* occurs at a node that violates P4, and a *degree violation* occurs at an internal node with only one child (violating P2). Observe that P1 is satisfied in a relaxed B-slack tree with no weight violations. Likewise, P2 is satisfied in a relaxed B-slack tree with no degree violations, and P4 is satisfied in a relaxed B-slack tree with no slack violations. Therefore, a relaxed B-slack tree that contains no violations is a B-slack tree. Rebalancing steps can be performed to eliminate violations, and gradually transform any relaxed B-slack tree into a B-slack tree.

9.2.2 Updates to relaxed B-slack trees

We now describe the algorithms for inserting and deleting keys in a relaxed B-slack tree (in a way that maintains P0', P1', P2' and P3). We use the **Insert**, **Delete** and **Overflow** updates introduced by Larsen and Fagerberg in their work on relaxed (a,b) -trees [84]. These updates appear in Figure 9.1. There, weights appear to the right of nodes, and shaded regions represent slack. If u is a node that is not the root, then we let $\pi(u)$ denote the parent of u . The insertion and deletion algorithms always ensure that all leaves have weight one. We also study how these updates change the amount of slack in nodes, and how they create, move or eliminate violations.

Deletion. First, a search is performed to find the leaf u where the deletion should occur. If the leaf does not contain the key to be deleted, then the deletion terminates immediately, and the tree does not change. If the leaf contains the key to be deleted, then the key is removed from the sequence of keys stored in that leaf. Deleting this key may create a slack violation.

Insertion. To perform an insertion, a search is first performed to find the leaf u where the insertion should occur. If u contains some slack, then the key is added to the sequence of keys in u , and the insertion terminates. Otherwise, u cannot accommodate the new key, so Overflow is performed. Overflow replaces u by a subtree of height one consisting of an internal node with weight zero, and two leaves with weight one. The b keys stored in u , plus the new key, are evenly distributed between the children of the new internal node. If u was the root before the insertion, then the new internal node becomes the new root. Otherwise, u 's parent $\pi(u)$ before the insertion is changed to point to the new internal node instead of u . After Overflow, there is a weight violation at the new internal node. Additionally, since the new internal node contains $b - 2$ slack, whereas u contained no slack, there may be a slack violation at $\pi(u)$.

Delete, Insert and Overflow maintain the properties of a relaxed B-slack tree. They will also maintain the properties of a B-slack tree, provided that rebalancing steps are performed to remove any violations that are created.

9.2.3 Rebalancing steps

The rebalancing steps are also based on the work of Larsen and Fagerberg. In fact, **Root-Zero**, **Root-Replace**, **Absorb** and **Split** are the same as in [84]. However, the **Compress** and **One-Child** operations are newly introduced by this work. These operations ensure that P4 and P2 are satisfied, respectively.

If there is a degree violation at the root, then Root-Replace is performed. If there is no degree violation at the root, but there is a weight violation at the root, then Root-Zero is performed. If there is a weight violation at an internal node that is not the root, then Absorb or Split is performed. Suppose there are no weight violations. If there is a degree violation at a node u and no degree or slack violation at $\pi(u)$, then One-Child is performed. If there is a slack violation

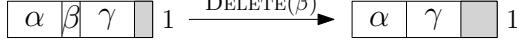
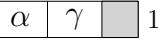
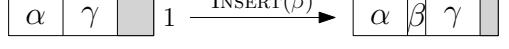
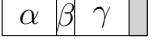
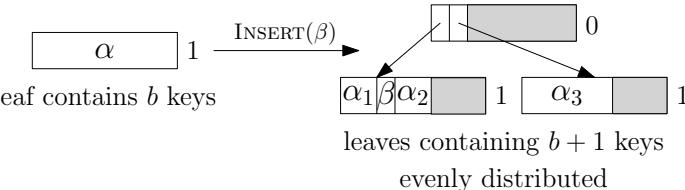
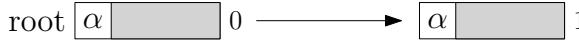
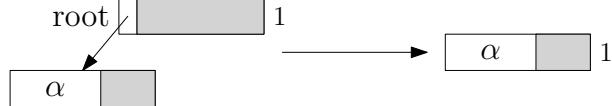
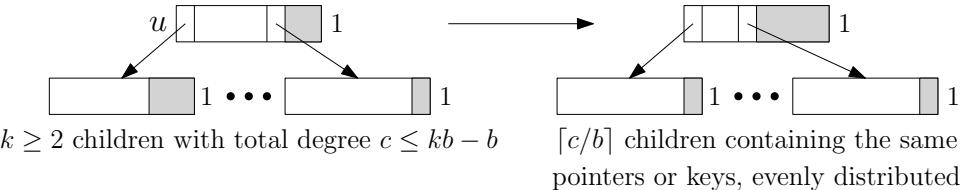
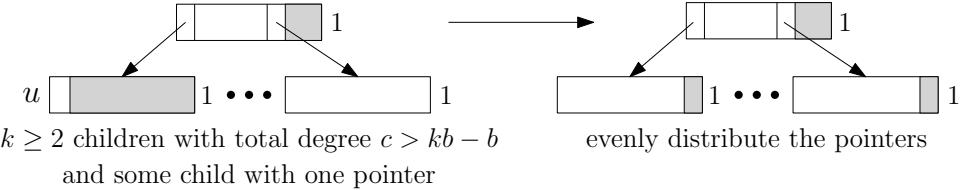
Delete	 $\xrightarrow{\text{DELETE}(\beta)}$ 
Insert	 $\xrightarrow{\text{INSERT}(\beta)}$ 
Overflow	<p>leaf contains b keys</p>  <p>leaves containing $b + 1$ keys evenly distributed</p>
Root-Zero	
Root-Replace	 <p>if the root has only one child, make that child the new root</p>
Absorb	 <p>initially, $\pi(u)$ contains less than b pointers</p>
Split	 <p>initially, $\pi(u)$ contains b pointers; afterwards, all pointers in alpha, beta, gamma are evenly distributed between the two children</p>
Compress	 <p>$k \geq 2$ children with total degree $c \leq kb - b$</p> <p>$[c/b]$ children containing the same pointers or keys, evenly distributed</p>
One-Child	 <p>$k \geq 2$ children with total degree $c > kb - b$</p> <p>evenly distribute the pointers and some child with one pointer</p>

Figure 9.1: Updates to B-slack trees (and relaxed B-slack trees). Nodes with weight zero contain exactly two pointers.

at a node u and no degree violation at u , then Compress is performed. Figure 9.1 illustrates these rebalancing steps. The goal of rebalancing is to eliminate all violations, while maintaining the relaxed B-slack tree properties.

Root-Zero. Root-Zero changes the weight of the root from zero to one, eliminating a weight violation, and incrementing the relaxed depth of every node. If P1' held before Root-Zero, it holds afterwards.

Root-Replace. Root-Replace replaces the root r by its only child u , and sets u 's weight to one. This eliminates a degree violation at r , and any weight violation at u . If u had weight zero before Root-Replace, then the relaxed depth of every leaf is the same before and after Root-Replace. Otherwise, the relaxed depth of every leaf is decremented by Root-Replace. In both cases, if P1' held before Root-Replace, it holds afterwards.

Absorb. Let u be a non-root node with weight zero. Absorb is performed when $\pi(u)$ contains less than b pointers. In this case, the two pointers in u are moved into $\pi(u)$, and u is removed from the tree. Since the pointer from $\pi(u)$ to u is no longer needed once u is removed, $\pi(u)$ now contains at most b pointers. The only node that was removed is u and, since it had weight zero, the relaxed depth of every leaf remains the same. Thus, if P1' held before Absorb, it also holds afterwards. Absorb eliminates a weight violation at u , but may create a slack violation at $\pi(u)$.

Split. Let u be a non-root node with weight zero. Split is performed when $\pi(u)$ contains exactly b pointers. In this case, there are too many pointers to fit in a single node. We create a new node v with weight one, and evenly distribute all of the pointers and keys of u and $\pi(u)$ (except for the pointer from $\pi(u)$ to u) between u and v . Now $\pi(u)$ has two children, u and v . The weight of u is set to one, and the weight of $\pi(u)$ is set to zero. As above, this does not change the relaxed depth of any leaf, so P1' still holds after Split. Split moves a weight violation from u to $\pi(u)$ (closer to the root, where it can be eliminated by a Root-Zero or Root-Replace), but may create slack violations at u and v .

Compress. Compress is performed when there is a slack violation at an internal node u , there is no degree violation at u , and there are no weight violations at u or any of its $k \geq 2$ children. Let $c \leq kb - b$ be the number of pointers or keys stored in the children of u . Compress evenly distributes the pointers or keys contained in the children of u amongst the first $\lceil c/b \rceil$ children of u , and discards the other children. This will also eliminate any degree violations at the children of u if $c > 1$. After the update, u satisfies P4. Compress does not change the relaxed depth of any node, so P1' still holds after. Compress removes at least one child of u , so it increases the slack of u by at least one, possibly creating a slack violation at $\pi(u)$. (However, it decreases the total amount of slack in the tree by at least $b - 1$.) Thus, after a Compress, it may be necessary to perform another Compress at $\pi(u)$. Furthermore, as Compress distributes keys and pointers, it may move nodes with different parents together, under the same parent. Even if two parents initially satisfied P4 (so the children of each parent contain a total of less than b slack), the children of the combined parent may contain b or more slack, creating a slack violation. Therefore, after a Compress, it may also be necessary to perform Compress at some of the children of u .

One-Child. One-Child is performed when there is a degree violation at an internal node u , there are no weight violations at u or any of its siblings, and there is no violation of any kind at $\pi(u)$. Let k be the degree of $\pi(u)$. Since there is no slack violation at $\pi(u)$, there are a total of $c > kb - b = b(k - 1)$ pointers stored in u and its siblings. Since u has only one child pointer, each of its other $k - 1$ siblings must contain b pointers. One-Child evenly distributes the keys and pointers of the children of $\pi(u)$. One-Child does not change the relaxed depth of any node, so P1' still holds after. One-Child eliminates a degree violation at u , but, like Compress, it may move children with different parents together under the same parent, possibly creating slack violations at some children of $\pi(u)$. So, it may be necessary to perform Compress at some of the children of $\pi(u)$.

All of these updates maintain P0', P2' and P3. While rebalancing steps are being performed to eliminate the violation created by an insertion or deletion, there is at most one node with weight zero.

We prove that a rebalancing step can be applied in any relaxed B-slack tree that is not a B-slack tree.

Lemma 9.1 *Let T be a relaxed B-slack tree. If T is not a B-slack tree, then a rebalancing step can be performed.*

Proof: If T is not a B-slack tree, it contains a weight violation, a slack violation or a degree violation. If there is weight violation, then Root-Zero, Absorb or Split can be performed. Suppose there are no weight violations. Let u be the node at the smallest depth that has a slack or degree violation. Suppose u has a degree violation. If u is the root, then Root-Replace can be performed. Otherwise, $\pi(u)$ has no violation, so One-Child can be performed. Suppose u does not have a degree violation. Then, u must have a slack violation, and Compress can be performed. ■

9.3 Analysis

This section provides a detailed analysis of B-slack trees that store n keys, by giving: an upper bound on the height of the tree, a lower bound on the average degree of nodes (and, hence, utilization), and an upper bound on the space complexity. We first give a brief outline of the proofs and results, and then provide the full details.

Arbitrary B-slack trees are difficult to analyze, so we begin by studying a class of trees called b -overslack trees. A b -overslack tree has a root with degree two, and satisfies P1, P2 and P3, but instead of P4, the children of each internal node contain a total of exactly b slack. Thus, a b -overslack tree is a relaxed B-slack tree, but not a B-slack tree. Consider a b -overslack tree T of height h that contains n keys. We prove that the total degree at depth $\delta \leq h$ in T is $d(\delta) = 2^{-\delta}(\alpha^\delta + \gamma^\delta)$, where $\alpha = b + \sqrt{b^2 - 4b}$ and $\gamma = b - \sqrt{b^2 - 4b}$. Since the total degree at the lowest depth is precisely the number of keys in the tree, every b -overslack tree of height h contains exactly $d(h)$ keys. Furthermore, when $h \geq 3$, we also have $(b-2)^h < d(h) \leq b^h$. Therefore, for $n > b^3$ (which implies height at least three), h satisfies $\lceil \log_b n \rceil \leq h \leq \lceil \log_{b-2} n \rceil$. We also prove that the average degree of nodes in T is $\frac{b \cdot d(h-1) - b + 2}{b \cdot d(h-2) - b + 3}$, which is greater than $b-2$ for $h \geq 3$.

We next prove some connections between overslack trees and B-slack trees. First, we show that each b -overslack tree of height h has a smaller total degree of nodes at each depth than any B-slack tree of height h . We do this by starting with an arbitrary B-slack tree of height h , and repeatedly removing pointers and keys from the children of each internal node that satisfies P4 (taking care not to violate P1, P2 or P3), until we obtain a b -overslack tree. It follows that each b -overslack tree of height h contains fewer keys than any B-slack tree of height h . Consequently, every b -overslack tree with n keys has height at least as large as any B-slack tree with n keys. We next prove that every b -overslack tree of height h has a smaller average node degree than any B-slack tree of height h . As above, the proof starts with an arbitrary B-slack tree of height h , and removes pointers and keys from nodes until the tree becomes a b -overslack tree. However, in this proof, every time we remove a pointer, we must additionally show that the average degree of nodes in the tree decreases.

We then compute the *space complexity* of a B-slack tree containing n keys, which is the number of words needed to store it. Consider a leaf-oriented tree with maximum degree b . For simplicity, we assume that each key and each pointer to a child or data occupies one word in memory. Thus, a leaf occupies $2b$ words, and an internal node occupies $2b - 1$ words. A memory block size of $2b$ is assumed. Let \bar{D} be the average degree of nodes. Then, $U = \bar{D}/b$ is the proportion of space that is utilized (which we call the *average space utilization* of the tree), and $1 - U$ is the proportion of space that is wasted. The space complexity is $2bF$, where F is the number of nodes in the tree. Suppose the tree contains n keys. By definition, the sum of the degrees of all nodes is $F - 1 + n$, since each node, except the root, has a pointer into it and the degree of a leaf is the number of keys it contains. Additionally, $F\bar{D}$ is equal to the sum of degrees

of all nodes, so $F = (n - 1)/(\bar{D} - 1)$. Therefore, the space complexity is $2b(n - 1)/(\bar{D} - 1)$. In order to compute an upper bound on the space complexity for a B-slack tree of height h , we simply need a lower bound on \bar{D} . Above, we saw that $\bar{D} > b - 2$ for B-slack trees of height at least three. It follows that a B-slack tree with $n > b^3$ keys has space complexity at most $2b(n - 1)/(b - 3) < 2n \frac{b}{b-3}$. (Recall that $2n$ is optimal under these assumption.) A slightly tighter upper bound is $2b(n - 1)/\left(\frac{b \cdot d(h-1) - b + 2}{b \cdot d(h-2) - b + 3} - 1\right)$.

Section 9.5 describes pathological families of B-trees, Overflow trees and H-trees, and compares the space complexity of example trees in these families with the worst-case upper bound on the space complexity of a B-slack tree. By studying these families, we obtain lower bounds on the space complexity of these trees that are above the upper bound for B-slack trees.

We also study the number of rebalancing steps necessary to maintain balance in a relaxed B-slack tree. Consider a relaxed B-slack tree obtained by starting from a B-slack tree containing n keys and performing a sequence of i insertions and d deletions. We prove that such a relaxed B-slack tree will be transformed back into a B-slack tree after at most $2i(2 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} (n+i)/2 \rfloor) + d/(b-1)$ rebalancing steps, irrespective of which rebalancing steps are performed, and in which order. Hence, insertions perform amortized $O(\log(n+i))$ rebalancing steps and deletions perform an amortized constant number of rebalancing steps.

9.3.1 Analysis of overslack trees

For our analysis, it is helpful to generalize the definition of b -overslack trees to the following. A (b,k) -overslack tree is the same as b -overslack tree, except that the root has degree k , instead of degree two.

We now compute the total degree of nodes at each depth in a (b,k) -overslack tree. As we will see, for each k , the total degree of nodes at a given depth δ is the same in every (b,k) -overslack tree of height at least δ .

Lemma 9.2 *The total degree of nodes at depth δ in a (b,k) -overslack tree of height h is:*

$$d(\delta, k) = \begin{cases} k & \text{if } \delta = 0 \\ kb - b & \text{if } \delta = 1 \\ b(d(\delta-1, k) - d(\delta-2, k)) & \text{if } 1 < \delta \leq h \end{cases}$$

Proof: In the following, we use $c(u)$ to denote the set of children of node u . Let T be a (b,k) -overslack tree. The proof is by induction on δ . The base cases $\delta = 0$ and $\delta = 1$ are immediate from the definition of a (b,k) -overslack tree. Consider $\delta > 1$. Let T_δ be the set of nodes at depth δ in T . Then,

$$d(\delta, k) = \sum_{v \in T_\delta} \deg(v) = \sum_{u \in T_{\delta-1}} \sum_{v \in c(u)} \deg(v).$$

Since T is an overslack tree,

$$\sum_{v \in c(u)} \deg(v) = \deg(u)b - b = b(\deg(u) - 1), \text{ so}$$

$$d(\delta, k) = \sum_{u \in T_{\delta-1}} b(\deg(u) - 1) = b(d(\delta-1, k) - |T_{\delta-1}|) = b(d(\delta-1, k) - d(\delta-2, k)).$$

■

Since $d(\delta, k)$ is a linear homogeneous recurrence relation with constant coefficients, we use the technique described in Section 2.1.1(a) of [58] to obtain the following closed form solution.

Lemma 9.3 $d(\delta, k) = 2^{-\delta}(k_1(b + \sqrt{b^2 - 4b})^\delta + k_2(b - \sqrt{b^2 - 4b})^\delta)$, where $k_1 = \frac{bk - 2b}{2\sqrt{b^2 - 4b}} + \frac{k}{2}$ and $k_2 = k - k_1$.

Corollary 9.4 $d(\delta, 2) = 2^{-\delta}((b + \sqrt{b^2 - 4b})^\delta + (b - \sqrt{b^2 - 4b})^\delta)$.

Since $b + \sqrt{b^2 - 4b}$ asymptotically approaches $2b$ as b increases, and $b - \sqrt{b^2 - 4b}$ approaches zero, $d(\delta, 2)$ approximately grows like b^δ for large b . Simple algebra establishes the following bounds on $d(\delta, 2)$.

Corollary 9.5 For $h \geq 3$, $(\frac{b}{2})^h < d(h, 2) \leq b^h$.

For the following lemma, we used symbolic mathematics software to obtain the partial derivatives of $d(\delta, k)$ with respect to δ and k , and prove that they are positive.

Lemma 9.6 $d(\delta, k)$ is an increasing function of δ and k .

Lemma 9.7 The total degree of nodes in every (b, k) -overslack tree of height h is

$$D(h, k) = \begin{cases} k & : h = 0 \\ k + b(d(h-1, k) - 1) & : h > 0 \end{cases}$$

Moreover, $D(h, k)$ is increasing in h and k .

Proof: By Lemma 9.2, $D(h, k) = \sum_{\delta=0}^h d(\delta, k) = k + (kb - b) + \sum_{\delta=2}^h b(d(\delta-1, k) - d(\delta-2, k))$. This telescoping sum reduces to $D(h, k) = k + (kb - b) + b(d(h-1, k) - d(0, k)) = k + b(d(h-1, k) - 1)$. Lemma 9.6 implies that $D(h, k)$ is increasing in h and k . ■

We now consider the average degree of nodes at each depth in a $(b, 2)$ -overslack tree. In any $(b, 2)$ -overslack tree, the two children of the root must share exactly $2b - b = b$ pointers or keys. Let us build intuition with an example. Consider a $(b, 2)$ -overslack tree in which the children of the root evenly share b pointers or keys. The grandchildren of the root must share a total of exactly $((b/2)b - b) + ((b/2)b - b) = b^2 - 2b$ pointers or keys. Thus, the average degree of nodes at depth zero is two, at depth one is $b/2$, and at depth two is $b - 2$. We prove that every $(b, 2)$ -overslack tree T of height h has the smallest average node degree of any (b, k) -overslack tree of height h . We first derive expressions for the average degree of nodes in an overslack tree, and the average degree at each depth.

Since the total degree of nodes at depth δ is $d(\delta, k)$, and the total number of nodes at depth δ is $d(\delta-1, k)$, we obtain the following.

Lemma 9.8 The average degree of nodes at depth δ in any (b, k) -overslack tree of height h is:

$$\bar{d}(\delta, k) = \begin{cases} k & : \delta = 0 \\ \frac{d(\delta, k)}{d(\delta-1, k)} & : 0 < \delta \leq h \end{cases}$$

We used our mathematics software to obtain the partial derivatives of $\bar{d}(\delta, k)$ with respect to δ and k . We proved the following two lemmas by showing that $\frac{\partial}{\partial \delta} \bar{d}(\delta, k) > 0$ for $2 \leq k \leq b - 2$, $\frac{\partial}{\partial \delta} \bar{d}(\delta, k) < 0$ for $k > b - 2$, and $\frac{\partial}{\partial k} \bar{d}(\delta, k) > 0$.

Lemma 9.9 $\bar{d}(\delta, k)$ is an increasing function of δ for $2 \leq k \leq b - 2$, and is a decreasing function of δ for $k \geq b - 1$.

Lemma 9.10 $\bar{d}(\delta, k)$ is an increasing function of k .

Let T be a B-slack tree of height h . Since every node in T except for the root is pointed to by exactly one child pointer, the number of nodes in T is the total degree of all nodes at depths zero through $h - 1$, plus one for the root, which is exactly $D(h - 1, k) + 1$. Thus, we obtain the following.

Lemma 9.11 The average degree of nodes in any (b, k) -overslack tree of height h is:

$$\bar{D}(h, k) = \begin{cases} k & : h = 0 \\ \frac{D(h, k)}{D(h-1, k)+1} & : h > 0 \end{cases}$$

Lemma 9.12 $\bar{D}(h, k)$ is an increasing function of h for $2 \leq k \leq b - 2$, and is a decreasing function of h for $k \geq b - 1$.

Proof: The average node degree at depth $\delta \in \{0, 1, \dots, h\}$ is the same in every (b, k) -overslack tree of height at least h . By definition, $\bar{D}(h, k)$ must be a weighted average of the terms $\bar{d}(0, k), \bar{d}(1, k), \dots, \bar{d}(h, k)$. Thus, $\bar{D}(h+1, k)$ is a weighted average of $\bar{D}(h, k)$ and $\bar{d}(h+1, k)$. Suppose $k \leq b - 2$. By Lemma 9.9, $\bar{d}(h+1, k) > \max_{i \leq h} \bar{d}(i, k)$. Therefore, $\bar{d}(h+1, k) > \bar{D}(h, k)$, which implies that $\bar{D}(h+1, k) > \bar{D}(h, k)$. Now, suppose $k \geq b - 1$. By Lemma 9.9, $\bar{d}(h+1, k) < \min_{i \leq h} \bar{d}(i, k)$. Therefore, $\bar{d}(h+1, k) < \bar{D}(h, k)$, which implies that $\bar{D}(h+1, k) < \bar{D}(h, k)$. ■

We used our mathematics software to obtain the partial derivative of $\bar{D}(h, k)$ with respect to k , and proved that it is positive, yielding the following result.

Lemma 9.13 $\bar{D}(h, k)$ is an increasing function of k .

We proved a simple lower bound on $\bar{D}(h, k)$ using our mathematics software.

Lemma 9.14 $\bar{D}(h, k) > b - 2$ for $h \geq 3$ (and $b \geq 5$).

9.3.2 Relating B-slack trees to overslack trees

In this section, we first prove that a $(b, 2)$ -overslack tree has the greatest height of any B-slack tree containing the same number of keys, and a smaller average degree of nodes than any B-slack tree with the same height. Then, we compute upper and lower bounds on the space used to store a B-slack tree with n keys.

Proposition 9.15 Let u be an internal node in a B-slack tree. If the total slack contained in the children of u is less than b , then some child of u has degree at least three.

Proof: Suppose the total slack contained in the children v_1, v_2, \dots, v_l of u is less than b . Then, $\deg(v_1) + \deg(v_2) + \dots + \deg(v_l) > lb - b$. By P2, $l \geq 2$. Since $b > 4$, it follows that $lb - b = b(l-1) > 4(l-1) \geq 2l$. Therefore, the average degree of the children of u is $(\deg(v_1) + \deg(v_2) + \dots + \deg(v_l))/l > \frac{lb-b}{l} > 2$. ■

Lemma 9.16 Every $(b, 2)$ -overslack tree of height h has a smaller total degree of nodes, at each depth, than any B-slack tree of height h .

Proof: Let T be a B-slack tree of height h . We can transform T into a $(b, 2)$ -overslack tree by removing keys and pointers. Removing a key, or a pointer from a node that has at least three pointers, does not affect P1, P2 or P3. Let u be any internal node whose children share a total of less than b slack. We arbitrarily remove a key or pointer from the child, v , of u with the largest degree. By Proposition 9.15, v must have degree at least three. We can repeat this process until T is a $(b, 2)$ -overslack tree. \blacksquare

Corollary 9.17 Every $(b, 2)$ -overslack tree with n keys has a larger height than any B-slack tree with n keys.

Corollary 9.18 Any B-slack tree of height h contains more keys than every $(b, 2)$ -overslack tree of height h , and, hence, more than $d(h, 2)$ keys.

Lemma 9.19 Every $(b, 2)$ -overslack tree of height h has a smaller average node degree than any B-slack tree of height h .

Proof: We first describe how to transform a B-slack tree into an overslack tree of the same height while decreasing the average node degree. Observe that, since an overslack tree satisfies P1, P2 and P3, any B-slack tree will become an overslack tree if pointers and keys are removed until, for each internal node u , the children of u share a total of b slack. The proof is by induction on the height of the tree. Let u be the root of a B-slack tree T .

In the base case, the children of u are leaves. Arbitrarily removing keys from the children of u until the children contain a total of exactly b slack will transform T into an overslack tree while decreasing the average degree of nodes.

Now, suppose the children of u are internal. By the inductive hypothesis, we can transform each subtree rooted at a child of u into an overslack tree. After these transformations, for every internal node in every subtree rooted at a child of u , the children of this internal node contain a total of b slack. If the children of u contain a total of b slack, then T is an overslack tree. Otherwise, we would like to remove some grandchild of u , to increase this slack. As we argued in the proof of Lemma 9.16, removing a pointer from a node that has the largest degree amongst its siblings yields a B-slack tree. However, we must carefully choose which grandchild to remove so that we decrease the average degree of nodes. Let v be the child of u that is the root of the tree with the largest average degree. By Lemma 9.13, v has the largest degree amongst its siblings. By Lemma 9.15, v must have at least three pointers, so removing one of its children does not violate P2. It is easy to verify that removing one of v 's children will not violate P1 or P3. We remove the child of v that is the root of the tree with the largest average degree. Since this tree has the largest average degree of any tree rooted at a child of v , removing it decreases the average degree of T . (This is because every other subtree rooted at a child of v has the same or smaller average degree, and removing this child of v decreases the degree of v .) We can repeatedly apply this transformation until the children of u contain a total of b slack, at which point T is an overslack tree.

We now prove the main result. Given a B-slack tree T , we first transform it into an overslack tree. Then, if the root of T has more than two children, we transform T into a $(b, 2)$ -overslack tree by keeping the two children that are the roots of the trees with the largest average degrees, and throwing away the rest. \blacksquare

Since the average degree of nodes represents the fraction of space that is utilized, this implies that every $(b, 2)$ -overslack tree of height h wastes a larger proportion of space than any B-slack tree of the same height. We can also obtain a lower bound on the average degree (and, hence, the fraction of space that is utilized) for any B-slack tree containing n keys.

Lemma 9.20 A B-slack tree with $n \geq 2$ keys has average degree greater than $\bar{D}(\lceil \log_b n \rceil - 1, 2)$.

Proof: Let T be a B-slack tree of height h containing n keys. By Lemma 9.19, T has average degree greater than $\bar{D}(h, 2)$. Lemma 9.12 implies that $\bar{D}(h, 2)$ increases with h , so it suffices to find a lower bound on h . Since b is the maximum possible degree for any node in T , h is at least $\lceil \log_b n \rceil - 1$. ■

We can now compute bounds on $S(n)$, the space complexity of a B-slack tree containing n keys. Recall from Section 9.2 that $S(n) = 2b(n-1)/(\bar{D}-1)$. By Lemma 9.20, $D^* \geq \bar{D}(\lceil \log_b n \rceil - 1, 2)$. Let $s = \bar{D}(\lceil \log_b n \rceil - 1, 2)$. Then,

$$\frac{n-1}{b-1} \leq F \leq \frac{n-1}{s-1} \text{ and } \frac{2b(n-1)}{b-1} \leq S(n) \leq \frac{2b(n-1)}{s-1}.$$

9.3.3 Amortized logarithmic rebalancing

In the following, we assume that the tree is initially a B-slack tree. After a sequence of insertions and deletions, the tree is a relaxed B-slack tree. The goal of this section is to establish an upper bound on the number of rebalancing steps needed to transform this relaxed B-slack tree back into a B-slack tree.

We assume that the updates shown in Figure 9.1 are performed sequentially. In a concurrent setting, locks or lock-free methodologies such as the template in Chapter 5 can be used to ensure that updates appear to atomically operate on mutually exclusive sets of nodes (so that the effect will be the same as if the updates were performed sequentially in some order).

Our analysis follows the approach taken in [84]. Consider any arbitrary B-slack tree T . Initially, we associate every key in the tree with the leaf that contains it. When a key is inserted into a leaf u , we associate the key with u . After a key is deleted from u , the key is still associated with u . If the node u is deleted, then all keys associated with u are instead associated with another node. Two cases arise. If u is deleted by a Root-Replace, then all keys associated with u are instead associated with the only child of u . Otherwise, u is deleted by Absorb or Compress, and all keys associated with u are instead associated with the node that was the parent of u before the Absorb or Compress.

Let σ be a sequence of updates to T , and w be an internal node in the tree after the updates in σ have been performed. We define the *A-multiset* of w to be the multiset of all keys associated with nodes in the subtree rooted at w . Therefore, the A-multiset of the root contains $n+i$ keys, where i is the number of insertions in σ and n is the size of T .

We also define the *relaxed height* of a node u in a relaxed B-slack tree. Suppose we formed a *new* relaxed B-slack tree T' by detaching the subtree rooted at u from the relaxed B-slack tree that contains it. The relaxed height of u , denoted $rh(u)$, is then the relaxed depth of the leaves in T' .

The following lemma relates the relaxed height of a node to the number of keys in its A-multiset.

Lemma 9.21 *Consider a B-slack tree T containing at least two keys, and a sequence of updates to it. Then, let u be any node in the resulting tree. If u is the root, then its A-multiset contains at least $2\lfloor \frac{b}{2} \rfloor^{rh(u)-\text{weight}(u)}$ keys. Otherwise, its A-multiset contains at least $\lfloor \frac{b}{2} \rfloor^{rh(u)}$ keys.*

Proof: The proof is by induction on the sequence of updates performed on T .

Base case. Let u be any node in T , T_u be the tree rooted at u , and h_u be the height of T_u . Any subtree of a B-slack tree is a B-slack tree, so T_u is a B-slack tree. By Corollary 9.18, T_u must contain at least $d(h_u, 2)$ keys. By Corollary 9.5,

$d(h_u, 2) > (\frac{b}{2})^{h_u}$. Since T_u is a B-slack tree, every node has weight one, so the relaxed height of each node is equal to its height and $h_u = rh(u)$. Therefore, $d(h_u, 2) > (\frac{b}{2})^{rh(u)} \geq \lfloor \frac{b}{2} \rfloor^{rh(u)} \geq 2 \lfloor \frac{b}{2} \rfloor^{rh(u)-1} = 2 \lfloor \frac{b}{2} \rfloor^{rh(u)-weight(u)}$.

Inductive step. Suppose the claim holds before an update U . We prove it holds after U . Let $rh'(u)$ be the relaxed height of a node u after U .

Suppose U is Delete. Then, each A-multiset remains the same, and every node has the same relaxed height before and after U .

Suppose U is Insert. Then, each A-multiset either gains one new key, or remains the same, and every node has the same relaxed height before and after U .

Suppose U is Root-Replace. Let p be the old root, and u be its only child. If u has weight one before U , then $rh'(u) = rh(p) - 1$. Otherwise, u has weight zero before U , so $rh'(u) = rh(p)$. Any keys associated with p before U are associated with u after U , so the A-multiset of the root is the same before and after U .

Suppose U is Root-Zero. Let r be the relaxed height of the root before U . By the inductive hypothesis, the A-multiset of the root contains at least $2 \lfloor \frac{b}{2} \rfloor^r$ keys before U . Moreover, U does not change any A-multiset. After U , the relaxed height of the root increases to $r+1$ because the weight of the root changes from zero to one, so the A-multiset of the root contains at least $2 \lfloor \frac{b}{2} \rfloor^r = 2 \lfloor \frac{b}{2} \rfloor^{(r+1)-1} = 2 \lfloor \frac{b}{2} \rfloor^{rh(\text{root})-weight(\text{root})}$.

Suppose U is Absorb. Let u be the child before U and p be its parent. In this case, u is removed by U , and all of its associated keys are instead associated with p . The A-multiset, weight and relaxed height of p are all the same before and after U .

Suppose U is Split. Let u be the child before U and p be its parent. In this case, U creates a new child, v , of p and moves all of p 's pointers (except for its pointers to u and v) into u and v , so that u and v each contain at least $\lfloor \frac{b+1}{2} \rfloor \geq \frac{b}{2}$ pointers. Observe that $rh'(u) = rh'(v) = rh'(p) = rh(p)$. Each pointer in u or v after U points to a node with relaxed height $rh(p) - 1$. By the inductive hypothesis, the A-multiset of every such node contains at least $\lfloor \frac{b}{2} \rfloor^{rh(p)-1}$ keys. Therefore, the A-multisets of u and v each contain at least $\frac{b}{2} \lfloor \frac{b}{2} \rfloor^{rh(p)-1} \geq \lfloor \frac{b}{2} \rfloor^{rh(p)} = \lfloor \frac{b}{2} \rfloor^{rh'(u)} = \lfloor \frac{b}{2} \rfloor^{rh'(v)}$ keys, and the A-multiset of p contains at least $\lfloor \frac{b}{2} \rfloor^{rh'(u)} + \lfloor \frac{b}{2} \rfloor^{rh'(v)} = 2 \lfloor \frac{b}{2} \rfloor^{rh'(p)}$ keys.

Suppose U is Overflow. Let u be the leaf that is full. In this case, U creates a new leaf r and an internal node p with weight zero and pointers to u and r , and moves half of the keys from u into r . After U , the A-multiset of p contains at least $b+1$ keys. Since u and r are leaves, $rh(p) = 1$, so $b+1 \geq 2 \lfloor \frac{b}{2} \rfloor^{rh(p)}$. The A-multisets of u and r each contain at least $\frac{b}{2}$ keys. Since $rh(u) = rh(r) = 1$, $\frac{b}{2} \geq \lfloor \frac{b}{2} \rfloor^{rh(u)} = \lfloor \frac{b}{2} \rfloor^{rh(r)}$.

Suppose U is Compress or One-Child. Let p be the upper node and k be its degree. Observe that U does not change the weight or relaxed height of any node, and does not remove any key from the A-multiset of p . After U , p has $\lceil \frac{c}{b} \rceil$ children that evenly share c pointers or keys. Thus, each child contains at least $\lfloor \frac{c}{\lceil c/b \rceil} \rfloor \geq \lfloor \frac{c}{c/b+1} \rfloor = \lfloor \frac{cb}{c+b} \rfloor = \lfloor \frac{b}{1+b/c} \rfloor$ pointers or keys. If U is One-Child, then $c > kb - b$ and $k \geq 2$, so $c > b$ and $\lfloor \frac{b}{1+b/c} \rfloor > \lfloor \frac{b}{2} \rfloor$. If U is Compress, then two cases arise. If p has at least two children after U , then $\lceil \frac{c}{b} \rceil \geq 2$, so $c/b+1 \geq 2$ and $b/c \leq 1$. Therefore, each child of p contains at least $\lfloor \frac{b}{1+b/c} \rfloor \geq \lfloor \frac{b}{2} \rfloor$ pointers or keys after U . Otherwise, after U , the single child of p contains all of the pointers and keys of the children that were removed, so its A-multiset is at least as large as it was before U . The claim then follows immediately from the inductive hypothesis. ■

Corollary 9.22 Consider a relaxed B-slack tree that results from performing a sequence of operations, i of which are insertions, on a B-slack tree containing n keys. The relaxed height of the root, and, hence, any node in this relaxed B-slack tree is at most $\lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor + 1$.

Lemma 9.23 After a sequence of operations, i of which are insertions, on a B-slack tree containing n keys, the total number of Absorb and Root-Zero updates that can be performed is at most i , and the number of Split updates that can be performed is at most $i(1 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor)$.

Proof: Absorb or Split is performed when a node has weight zero. Overflow is the only update that increases the number of zero weights in the tree, and at most i Overflows occur, so there are at most i nodes with weight zero in the tree. Absorb and Root-Zero each decrease the number of zero weights in the tree by one, so at most i of these updates can be performed. Root-Zero, Root-Replace, Absorb and Split are the only updates that can change the relaxed height of a node with weight zero. Root-Zero, Root-Replace and Absorb each change a zero weight to one, and Split moves a zero weight from a node with relaxed height r to a node with relaxed height $r+1$. Therefore, each zero weight will remain at a node with the same relaxed height until it is moved by Split or changed to one by Root-Zero, Root-Replace or Absorb. Since the relaxed height of any node in the tree is at most $\lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor + 1$, each of the i zero weights in the tree can be moved by Split at most $\lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor + 1$ times. ■

Lemma 9.24 Let T be a relaxed B-slack tree of height h that is obtained by performing any sequence of i insertions and d deletions on an initially empty relaxed B-slack tree. At most $2i(4 + \frac{3}{2} \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) + 2d/(b-1)$ rebalancing steps can be applied to T .

Proof: Let c be the total degree of the children of a parent where Compress or One-Child is performed. We first bound the number of One-Child updates that can be performed. If a node has exactly one pointer, we say a *pointer violation* occurs at that node. One-Child is performed only when a pointer violation occurs at a child of the parent and $c > kb - b$. Since $c > kb - b$ and $k \geq 2$, $c \geq b+1$, so the parent will have at least two children after One-Child. Furthermore, each child of the parent will have degree at least $\lfloor b/2 \rfloor$. Thus, One-Child removes every pointer violation at a child of the parent, and does not create any pointer violation. Root-Replace removes a pointer violation at the root, decreasing the number of pointer violations in the tree by one. However, Compress can *increase* the number of pointer violations in the tree by one if $c \leq b$. No other update changes the number of pointer violations in the tree. Therefore, the total number of One-Child and Root-Replace updates that can be performed is bounded above by the number of Compress updates.

We bound the number of Compress updates by studying the change in the total amount of slack in the tree that is caused by each type of update. It is convenient to ignore the slack in any node with a zero weight value, since Compress cannot affect any such node. Compress redistributes a total of $c < kb - b$ pointers or keys from k nodes to $\lceil c/b \rceil$ nodes. Since $c \leq kb - b = b(k-1)$, $\lceil c/b \rceil \leq k-1$, Compress will remove at least one node from the tree. Removing this node removes b slack, and increases slack at the parent by one. Thus, Compress reduces the total amount of slack in the tree by at least $b-1$ (and by even more, if more than one node is removed). Split is applied precisely when the parent of a node with weight value zero contains exactly b pointers (and no slack). Since the node with weight value zero contains exactly two pointers, $b+1$ pointers are moved into the nodes with weight value one in Figure 9.1, so Split increases the total slack in the tree by exactly $b-1$. By Lemma 9.23, Split can be performed at most $i(1 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor)$ times, so the total amount of slack created by Split is at most $i(1 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor)(b-1)$. It is easy to verify that Insert, Insert-Distribute and Absorb each decrease the total slack by one, and that Delete and Insert-Overflow increase the total slack by one and $2(b-1)$, respectively. Thus, the total slack created by Delete and Insert-Overflow updates is $d + 2i(b-1)$, so the total slack in T is at most $i(1 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor)(b-1) + d + 2i(b-1) = i(b-1)(3 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) + d$. Therefore, at most $i(3 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) + d/(b-1)$ Compress updates can occur.

By Lemma 9.23 at most i Absorb and Root-Zero updates and $i(1 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor)$ Split updates can occur. Since the total number of Root-Replace and One-Child updates is at most the number of Compress updates, the number of Root-Replace, One-Child and Compress updates that can occur is at most $2i(3 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) + 2d/(b-1)$. Therefore, at most $2i(3 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) + 2d/(b-1) + i + i(1 + \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) = 2i(4 + \frac{3}{2} \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) + 2d/(b-1)$ rebalancing steps can be applied to T . \blacksquare

This result implies that the number of rebalancing steps needed to rebalance the tree after a sequence of deletions is amortized constant, and after a sequence of insertions is amortized logarithmic in: the size of the tree the last time it was a B-slack tree plus the number of insertions that have occurred since then. Section 9.4 explains how B-slack trees can be modified to obtain amortized constant rebalancing by slightly increasing the amount of slack shared amongst the children of an internal node.

9.4 B-slack trees with amortized constant rebalancing

The main challenge in achieving amortized constant rebalancing is ensuring that long sequences of Split and Compress operations occur infrequently. Split can necessitate other Splits higher in the tree and many Compresses. Compress can necessitate many other Compresses. This makes Split and Compress particularly problematic.

Split occurs only when an internal node is full. If a Compress at an internal node leaves some slack in each of its children, then Splits will not immediately occur at the children. With this in mind, we make some small modifications. P4 is replaced with P4', which says that, for each internal node u of degree k , the total slack contained in the children of u is at most $b+k-1$ (so the worst-case slack per node is only one greater than in a standard B-slack tree). A slack violation then occurs at any internal node that violates P4'. The children of an internal node of degree k where a slack violation occurs will have total degree less than $kb - (b+k-1) = (k-1)(b-1)$. Thus, Compress is performed only at internal nodes whose children have total degree $c \leq (k-1)(b-1)$, and One-Child is performed only at internal nodes whose children have total degree $c > (k-1)(b-1)$. This threshold is chosen so that Compress is only performed when it can remove one node and still leave each child with one slack (so that each child can accommodate one more key before necessitating a Split). We then change Compress so that it evenly distributes the c pointers or keys of its children amongst $\lceil \frac{c}{b-1} \rceil$ nodes, instead of $\lceil \frac{c}{b} \rceil$. This way, each child is guaranteed to have at least one slack afterwards.

We prove that the number of rebalancing steps is amortized constant using the potential method. The potential of a node u , denoted $\phi(u)$, captures the intuition that a node is bad if it contains too much slack, is full, or has weight zero.

$$\phi(u) = \begin{cases} b - \deg(u) & \text{if } \deg(u) < b \text{ and } u \text{ has weight one} \\ b & \text{otherwise (i.e., } \deg(u) = b \text{ or } u \text{ has weight zero)} \end{cases}$$

The potential of a tree T , denoted $\Phi(T)$, is the sum of potentials of its nodes.

We now study how $\Phi(T)$ is changed by deletion, insertion, and each rebalancing step. Let u be a node with degree k . Recall that leaves never have weight zero.

Delete. If u is full, then $\phi(u)$ changes from b to 1. Otherwise, $\phi(u)$ changes from $b-k$ to $b-(k-1)$. So, $\phi(u)$ increases by at most one.

Insert. If the insertion fills u , then $\phi(u)$ changes from 1 to b . Otherwise, $\phi(u)$ changes from $b-k$ to $b-(k+1)$. So, $\phi(u)$ increases by at most $b-1$.

Overflow. A full node with potential b turns into a node with weight zero, which has potential b , and two nodes with weight one that share a total of $b - 1$ slack. Thus, b potential is replaced by $b + b - 1 = 2b - 1$ potential, which is an increase of $b - 1$.

Absorb. Let u be the node with weight zero. Its parent $\pi(u)$ has weight one. Beforehand, u has degree two and $\pi(u)$ contains $j \leq b - 1$ pointers, so it is not full. Absorb decreases potential by b by eliminating u . It also moves a pointer from u to $\pi(u)$. If $j = b - 1$, then $\phi(\pi(u))$ changes from 1 to b , increasing potential by $b - 1$, for a net decrease of one. Otherwise, $\phi(\pi(u))$ changes from $b - j$ to $b - (j - 1)$, for a net decrease in potential of $b - 1$.

Split. Let u be the node with weight zero. Its parent $\pi(u)$ has weight one, but it is full, so it has potential b beforehand. Afterwards, it has weight zero, so its potential does not change. Before the Split, u has potential b , and it is split into two nodes, each of weight one, that share a total of $b - 1$ slack. After the Split, the sum of their potentials is $b - 1$. Thus, the potential of the tree is decreased by one.

Compress. Let u be a node with $k \geq 2$ children that have total degree at most $(k - 1)(b - 1)$. The modified version of Compress will leave at least one slack at each child of u , so the total potential of u 's children will be the total amount of slack they contain, which decreases by at least b , since at least one of the children is removed. Removing a child of u also increases the slack at u by one, which increases $\phi(u)$ by one (unless u was full, in which case it decreases $\phi(u)$). For each child of u that is removed by Compress, b slack is eliminated at the children of u , and one slack is added at the parent. Therefore, the total potential of the tree decreases by $b - 1$ for each child removed by Compress. Since at least one child is removed, the total potential of the tree decreases by at least $b - 1$.

One-Child. Since One-Child evenly distributes keys, it cannot create any more full nodes than existed beforehand. It does not affect weights, and it does not remove any key or pointer. So, One-Child does not affect the total potential of the tree.

Since $\Phi(T)$ is increased by one for Delete and $b - 1$ for Insert (and Overflow), and no other operation increases it, after i insertions and d deletions, $\Phi(T) \leq (b - 1)i + d$. We can use $\Phi(T)$ to bound the number of rebalancing steps that can be performed on T . Let C, A, S, R_0, R_r be the number of Compresses, Absorbs, Splits, Root-Zeros and Root-Replaces, respectively. We immediately obtain $(b - 1)C + A + S + 2(R_0 + R_r) \leq \Phi(T) \leq (b - 1)i + d$. Astute readers will notice that One-Child has not yet made an appearance. By the same argument as in Lemma 9.24, the number of One-Childs is at most C . Therefore, the number of rebalancing steps is constant per update.

In fact, we can achieve tighter bounds if we are more careful. By the same argument as in Lemma 9.23, $A + R_0 \leq i$. Additionally, the same argument used to show that the number of One-Childs is at most C applies to Root-Replace, so $R_r \leq C$. Therefore, on average, there is at most one Absorb or Root-Zero per insertion, at most one Compress (and One-Child) and Root-Replace per insertion, and at most one Split per insertion. Similarly, on average, there is at most one Absorb, Split, Root-Replace or Root-Zero per deletion, and at most one Compress (and One-Child) per $b - 1$ deletions.

The increase in space complexity associated with these changes is very small. Since the worst-case slack per node is only one greater than in a B-slack tree, the minimum average degree of this modified B-slack tree is at most one less than in a B-slack tree. So, worst-case lower bound on utilization changes from \bar{D}/b to $(\bar{D} - 1)/b$, and the space complexity upper bound changes from $2b(n - 1)/(\bar{D} - 1) < 2b(n - 1)/(b - 3)$ to $2b(n - 1)/((\bar{D} - 1) - 1) < 2b(n - 1)/(b - 4)$.

Max degree	B-tree	Overflow tree	H-tree	B-slack tree	Optimal
8	$\geq 5.333n$	$\geq 5.066n$	$\geq 3.840n$	$< 2.789n$	$2.000n$
16	$\geq 4.571n$	$\geq 3.120n$	$\geq 2.685n$	$< 2.301n$	$2.000n$
32	$\geq 4.266n$	$\geq 2.492n$	$\geq 2.307n$	$< 2.145n$	$2.000n$

Figure 9.2: Space complexity of example trees, and worst-case bound for B-slack trees.

9.5 Space complexity of competing trees

In this section, we study the space complexity of some pathological families of B-trees, Overflow trees and H-trees. The maximum degree of nodes is b , the block size is $2b$, and all trees are leaf-oriented.

- **B-tree.** The root has degree two, and all other nodes have degree $b/2$.
- **Overflow tree.** The root has degree two, the internal nodes have degree $b/2$, and the leaves have degree $b - 3$. Overflow groups are chosen to be as large as possible, to minimize wasted space. Specifically, for each parent u of a leaf, u 's children are all in a single group, with one shared overflow node. Thus, each overflow node is shared by $b/2$ leaves (which contain a total of $(b - 3)(b/2) = b^2/2 - 3b/2$ keys).
- **H-tree.** Parameters γ and δ are chosen to be as large as possible, to minimize wasted space. The root has degree two, the internal nodes have degree $\lceil b/\sqrt{2} \rceil$, and the leaves have degree $b - 2$. (H-trees are node-oriented, which would significantly inflate their space complexity on a system with only one block size. The space complexity bounds shown in Figure 9.2 ignore this, and are thus quite charitable. To actually achieve such good space complexity bounds for H-trees, one would have to completely redesign the data structure to be leaf-oriented.)

We assume that a key and a pointer each occupy a single word in memory. For each family, and each choice of maximum degree in $\{8, 16, 32\}$, we consider the minimum height tree from the family containing at least 10^6 keys, and computed its space complexity. Observe that the resulting space complexity values are *lower bounds* on the worst-case space complexity for these data structures. These space complexity values appear in Figure 9.2, along with *very pessimistic upper bounds* on the space complexity for any B-slack tree ($b \in \{8, 16, 32\}$) containing at least 50,000 keys. (The aforementioned upper bounds actually apply to overslack trees, which allow the slack shared amongst the children of a node to be one greater than in a B-slack tree. Additionally, despite the fact that the space complexity of B-slack trees *improves as the number of keys grows*, these upper bounds only assume the tree contains at least 50,000 keys, in contrast to the other data structures, which contain at least 10^6 keys.) Therefore, these results are actually quite charitable to the other data structures.

Nevertheless, the advantage of B-slack trees is clear. The optimal space to store n keys and pointers to associated data is $2n$. H-trees, the closest competitor to B-slack trees, use more than double the space beyond what is optimal. If these trees were modified to implement a set instead of a dictionary (by eliminating data and allowing leaves to contain up to $2b$ keys), then the optimal space would become n , and it is expected that relative differences in space complexity between the trees would increase further.

9.6 Sequential experiments

Java implementation The B-slack tree was implemented as a sequential data structure in Java. Each update U shown in Figure 9.1 was implemented as follows. A process performing U creates new a node for each node on the right hand side of the depiction of U in Figure 9.1, and replaces the nodes on the left hand side by these new nodes. The nodes that were replaced by U are eventually reclaimed by Java’s automatic garbage collection. If a Delete, Insert or Overflow creates a violation, then the process invokes a **Cleanup** procedure to perform rebalancing steps until the tree no longer contains any violations.

For simplicity, we implemented Cleanup as a recursive procedure. Cleanup takes the node u where a violation occurs as its argument, and attempts to perform a rebalancing step to fix the violation at u . If this rebalancing step creates any new violations, or replaces any nodes with existing violations and moves the violations to new nodes, then recursive invocations of Cleanup are performed to eliminate these violations. If an invocation of Cleanup sees that the node whose violation it was supposed to fix has already been replaced, then it knows the invocation of Cleanup that replaced it will make a recursive call to fix the violation, wherever it was moved. Thus, this invocation of Cleanup can simply return. The downside of a recursive implementation of Cleanup is that stack overflow may occur if rebalancing steps create a large number of violations. Other ways to implement the Cleanup procedure are discussed in Section 9.7.

Java is not an ideal language for implementing the B-slack tree, since it gives very little control over memory layout. The purpose of our Java implementation is simply to serve as a guide for any implementers who are interested in porting the B-slack tree tree to other languages. Each node is implemented as an array of keys, and an array of child pointers. Unlike in C/C++, in Java, arrays cannot be embedded directly in a node. Instead, nodes contain pointers to arrays, which are located elsewhere in memory. Thus, even after a process has loaded (a cache line that contains) a node, accessing a key or child pointer of that node still requires performing additional loads from potentially distant locations in main memory. This makes the implementation somewhat inefficient. Furthermore, the implementation does not directly satisfy the space complexity upper bounds computed in Section 9.3. However, if we ignore these complications and pretend that keys and pointers are embedded directly inside nodes, then we can still use this implementation to study structural properties of B-slack trees in practice.

Methodology We performed randomized experimental trials for several b values, simulated workloads, and tree sizes. Each trial was divided into two phases. In the first phase, a B-slack tree was created and initialized by inserting and deleting (with 50% probability each) keys drawn uniformly randomly from $[0, \text{size}]$, until the tree stabilized, containing approximately $\text{size}/2$ keys. In the second phase, one million random insertions and deletions were performed with some specified probabilities, and each rebalancing step was recorded. We considered probabilities: 50% insertion and 50% deletion (50i-50d), 90% insertion and 10% deletion (90i-10d), and 10% insertion and 90% deletion (10i-90d). At the end of each trial, average degree and space complexity were computed (under the assumption that keys and pointers were actually embedded directly in nodes).

Results We discuss a small selection of the results. For 50i-50d, $b = 16$ and $\text{size} = 2^{20} = 1,048,576$, there were approximately 1.2 rebalancing steps per successful update, the average degree was approximately 15.5, and the space complexity was less than $2.209n$. In fact, even with a rather small size of $2^{12} = 4,096$, the average degree was approximately 15.4, and the space complexity was less than $2.226n$, which is substantially better than the theoretical average degree lower bound of 12.7 and space complexity upper bound of $2.726n$. This suggests that the performance

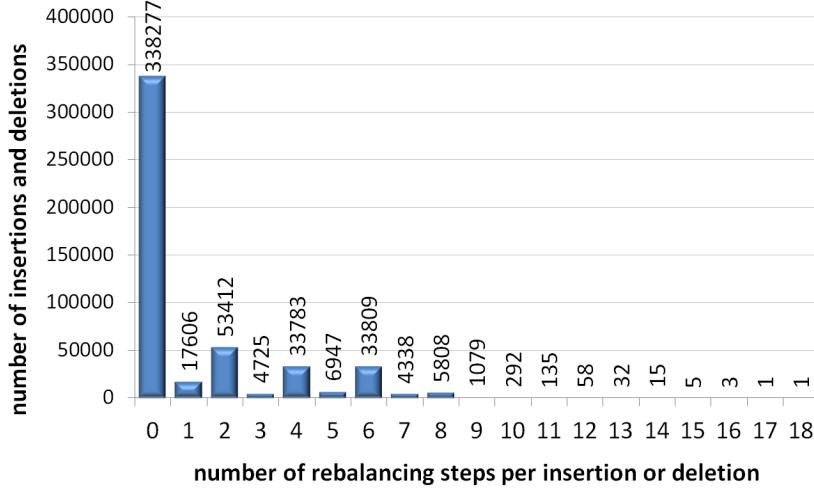


Figure 9.3: Histogram showing the frequency of updates that perform a certain number of rebalancing steps in a randomized trial.

of B-slack trees is much better in practice than what is suggested by our theoretical results. For 50i-50d, $b = 32$, and $\text{size} = 2^{20}$, there were approximately 1.1 rebalancing steps per successful update, the average degree was approximately 31.5, and the space complexity was less than $2.097n$. For 10i-90d, $b = 16$ and $\text{size} = 2^{20}$, there was less than one rebalancing step per successful update. Even with 90% of updates being deletion, the average degree remained the same as in the 50i-50d case, and the space complexity was less than $2.213n$. For 90i-10d, there were approximately 1.2 rebalancing steps per successful update. These results suggest that little rebalancing is required for random updates on uniform keys.

Distribution of rebalancing It is also interesting to understand how many rebalancing steps are necessitated by each insertion or deletion. So, we plotted a histogram for one trial with parameters 50i-50d, $b = 16$ and $\text{size} = 2^{20} = 1,048,576$ in Figure 9.3. (Results for other trials, values of b and simulated workloads are similar.) The x -axis shows the number of rebalancing steps that were performed by a single insertion or deletion, and the y -axis shows how many insertions or deletions performed x rebalancing steps. The total number of successful insertions and deletions was 500,326, and the height of the tree was four when the trial finished. The most rebalancing steps performed by an insertion or deletion was 18, 67.6% of successful insertions and deletions performed no rebalancing, 97.6% performed six or less rebalancing steps, and 99.9% performed less than ten rebalancing steps.

9.7 Implementation issues for rebalancing

One way to implement rebalancing is to explicitly maintain a collection of pointers to internal nodes where rebalancing steps must be performed. After an update creates a violation, a rebalancing step is performed to fix that violation. Every time a rebalancing step creates a violation, a pointer to the node where the violation occurs (and possibly also a pointer to its parent) is added to the collection. An update does not terminate until it has emptied the queue and performed rebalancing steps to fix all violations in the tree.

Observe that violations can only occur at internal nodes. The number of internal nodes is quite small compared to

n (close to $n/(b-2)^2$ in B-slack trees of height at least three). So, even if a collection contains *every* internal node, the worst-case collection size may be reasonable for some applications. Since the amortized number of rebalancing steps per update is small, most updates will result in a small collection. We recommend using a small, fixed-size queue and switching to a more computationally expensive algorithm if the queue becomes full. For instance, when a process tries to enqueue a pointer and the queue is full, it can simply discard that pointer, and continue the algorithm, recording the fact that the a pointer was discarded. Eventually, after enough rebalancing steps are performed, the queue becomes empty, and the process can traverse the tree to find any outstanding violations, repopulate the queue, and continue the algorithm. Although this approach is very expensive once the queue becomes full, it will not significantly increase the average running time of updates if the queue rarely becomes full. The experimental results in Section 9.6 indicate that this approach could be practical, even with a very small bound on queue size.

Chapter 10

Relaxed B-slack trees implemented with the template

```

type Node
  ▷ Fields used by LLX/SCX algorithm
  info      ▷ pointer to SCX-record
  marked    ▷ marked bit
  ▷ User-defined fields
  weight    ▷ weight bit (immutable)
  leaf      ▷ leaf bit (immutable)
  searchKey ▷ an auxiliary key for rebalancing (immutable)
  d         ▷ degree of the node (immutable)
  k1,k2,...,kd ▷ keys (immutable)
  p1,p2,...,pd ▷ pointers (mutable)

```

Figure 10.1: Data definition for a node in a relaxed B-slack tree.

10.1 Implementation

We present an implementation of a relaxed B-slack tree using the template described in Chapter 5. This implementation is very similar to the implementation of the (a,b) -tree in Chapter 8, except for the way that rebalancing is performed.

Let b be the maximum degree of nodes. We represent each node by a Data-record with b mutable pointers, and b immutable keys, as well as immutable fields d , $weight$, $leaf$ and $searchKey$. (See Figure 10.1.) The field d contains the number of pointers that are used. The $weight$ field contains the node's weight. The bit $leaf$ indicates whether the node is a leaf. The field $searchKey$ contains an auxiliary key that can be used to locate the node during rebalancing. If a node contains at least one key, then $searchKey$ is just the first key, k_1 , which might seem redundant. However, when a node contains no keys, $searchKey$ provides a process with a key that it can use to search for the node (which would otherwise be difficult to locate). Internal nodes have one fewer key than pointers, so the key k_d is unused in an internal node. Leaves have exactly as many pointers as keys.

To avoid special cases when the tree is empty, we add a sentinel node at the top of the tree. The sentinel node $entry$ always has one child and no keys. (Every search that passes through it will simply follow that one child pointer.) The sole child of this sentinel node is initially an empty leaf (with $d = 0$ and no keys or pointers). The actual relaxed B-slack tree is rooted at the child of the sentinel node. For convenience, we use $root$ to refer to the current child of the sentinel node.

Detailed pseudocode for GET (and its auxiliary subroutine SEARCH), INSERT and DELETE is given in Figure 8.3, 10.2 and 10.3. The implementations of GET and SEARCH are identical to the same procedures in the relaxed (a,b) -tree implementation. The implementations of INSERT and DELETE are also very similar to those of the relaxed (a,b) -tree. The difference lies in the way that these procedures trigger rebalancing. In our discussion of INSERT and DELETE, we focus on the differences between the INSERT and DELETE procedures for the relaxed B-slack tree and relaxed (a,b) -tree. The way we perform rebalancing in the relaxed B-slack tree is quite different from rebalancing in the relaxed (a,b) -tree, and we carefully discuss the relaxed B-slack tree rebalancing algorithm.

10.1.1 Insertion

Pseudocode for INSERT appears in Figure 10.2. As in the relaxed (a,b) -tree, INSERT takes two arguments: key and $value$. If key is not in the dictionary, then INSERT inserts the key-value pair $\langle key, value \rangle$, and returns \perp . Otherwise,

```

1  INSERT(key,value)
2    ▷ Returns  $\perp$  if key was not in the dictionary. Otherwise, this returns the value previously associated with key.
3    loop
4      ▷ Start of operation attempt
5      ▷ Search for key in the tree
6       $\langle -, p, l \rangle := \text{SEARCH}(key)$ 
7
8      ▷ Template iteration 0 (parent of leaf)
9       $result_p := \text{LLX}(p)$ 
10     if  $result_p \in \{\text{FAIL}, \text{FINALIZED}\}$  then continue ▷ goto next iteration (to retry)
11     if  $l \notin result$  then ▷ CONFLICT: verify p still points to l
12       continue ▷ goto next iteration (to retry)
13   Let  $p.p_i$  be the child pointer of p that pointed to l at the previous line
14
15   ▷ Template iteration 1 (leaf)
16    $result_l := \text{LLX}(l)$ 
17   if  $result_l \in \{\text{FAIL}, \text{FINALIZED}\}$  then continue ▷ goto next iteration (to retry)
18
19   ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
20    $V := \langle p, l \rangle$ 
21    $R := \langle l \rangle$ 
22    $fld :=$  a pointer to  $p.p_i$ 
23   if l contains key then ▷ Replace the value associated with an existing key
24     Let oldValue be the value associated with key in l
25      $n :=$  new copy of l that contains  $\langle key, value \rangle$  instead of  $\langle key, oldValue \rangle$ 
26     overflow := FALSE
27   else if  $l.d < b$  then ▷ Insert a new key-value pair into a full leaf
28      $n :=$  new copy of l that has the new key-value pair inserted
29      $oldValue := \perp$ 
30     overflow := FALSE
31   else ▷  $l.d = b$  ▷ Insert a new key-value pair into a non-full leaf
32      $n :=$  pointer to a subtree of three newly created nodes: one internal node and two leaves, configured as in
33     Overflow in Figure 9.1 (so that the key-value pairs in  $kv(l) \cup \{\langle key, value \rangle\}$  are evenly distributed
34     between the leaves, and the internal node has weight zero). The searchKey of each new node is its first
35     key.
36      $oldValue := \perp$ 
37     overflow := TRUE
38   ▷ If n will replace root, then we also perform Root-Zero as part of the same atomic update
39   if p = entry then n.weight = 1
40
41    $success := \text{SCX}(V, R, fld, n)$ 
42   ▷ End of operation attempt
43   if success then
44     if overflow then ▷ After an Overflow, we may need to rebalance
45       FIXWEIGHT(n)
46       FIXSLACK(p)
47       return oldValue ▷ Check for weight violation at n
48   else continue ▷ Check for slack violation at p ▷ goto next iteration (to retry)

```

Figure 10.2: Pseudocode for INSERT. Here, *b* is the maximum degree of nodes.

INSERT replaced the existing key-value pair $\langle key, oldValue \rangle$ in the dictionary with $\langle key, value \rangle$ and returns $oldValue$.

INSERT differs from the relaxed (a,b) -tree's INSERT procedure in three ways. First, instead of having a helper function TRYINSERT that is invoked repeatedly in a loop, the body of the TRYINSERT function is inlined directly in INSERT. Second, when new nodes are created for an Overflow update, we set the *searchKey* field of each new node to the node's first key. Note that each new node contains at least one key, and can be located by searching for its first key (using the SEARCHNODE procedure in Figure 10.8, which is described below). Third, after performing a successful SCX, instead of invoking CLEANUP to perform any needed rebalancing, two rebalancing procedures called FIXWEIGHT and FIXSLACK are invoked. These procedures are discussed in Section 10.1.3.

```

46  DELETE(key)
47  ▷ Returns  $\perp$  if key was not in the dictionary. Otherwise, this returns the value associated with key.
48  loop
49    ▷ Start of operation attempt
50    ▷ Search for key in the tree
51     $\langle -, p, l \rangle := \text{SEARCH}(key)$ 

53    ▷ Template iteration 0 (parent of leaf)
54     $result_p := \text{LLX}(p)$ 
55    if  $result_p \in \{\text{FAIL}, \text{FINALIZED}\}$  then continue ▷ goto next iteration (to retry)
56    if  $l \notin result$  then ▷ CONFLICT: verify p still points to l
57      continue ▷ goto next iteration (to retry)
58    Let  $p.p_i$  be the child pointer of p that pointed to l at the previous line

60    ▷ Template iteration 1 (leaf)
61     $result_l := \text{LLX}(l)$ 
62    if  $result_l \in \{\text{FAIL}, \text{FINALIZED}\}$  then continue ▷ goto next iteration (to retry)

64    ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
65     $V := \langle p, l \rangle$ 
66     $R := \langle l \rangle$ 
67     $fld :=$  a pointer to  $p.p_i$ 
68    if l does not contain key then ▷ The tree does not contain key
69      return  $\perp$ 
70    else
71      Let oldValue be the value associated with key in l
72       $n :=$  new copy of l that does not contain  $\langle key, oldValue \rangle$ 
73       $success := \text{SCX}(V, R, fld, n)$ 
74      ▷ End of operation attempt
75      if success then
76         $\text{FIXSLACK}(p)$  ▷ Check for slack violation at p
77        return oldValue
78      else continue ▷ goto next iteration (to retry)

```

Figure 10.3: Pseudocode for DELETE.

10.1.2 Deletion

Pseudocode for DELETE appears in Figure 10.3. As in the relaxed (a,b) -tree, DELETE takes one argument, key . If key is not in the dictionary, then DELETE simply returns \perp . Otherwise, DELETE removes the existing key-value pair $\langle key, oldValue \rangle$ from the dictionary and returns $oldValue$.

DELETE differs from the relaxed (a,b) -tree's DELETE procedure in three ways. First, instead of having a helper function TRYDELETE that is invoked repeatedly in a loop, the body of the TRYDELETE function is inlined directly in DELETE. Second, when the new node is created for a DELETE, we set its $searchKey$ field to the $searchKey$ of the node l that it is replacing. Note that, since l could be reached by searching for $l.searchKey$ (using the SEARCHNODE procedure) before the DELETE, so can the new node after the DELETE. Third, after performing a successful SCX, instead of invoking CLEANUP to perform any needed rebalancing, the rebalancing procedure FIXSLACK is invoked.

Update U by P	Violations that P becomes responsible for after U	
Delete	possible $\langle slack, parent \rangle$	
Insert	none possible	
Overflow	$\langle weight, n \rangle$, and possible $\langle slack, parent \rangle$	
Root-Zero	$\langle degree, n \rangle$, and $\langle slack, n \rangle$	if $\langle degree, root \rangle \in \beta$ if $\langle slack, root \rangle \in \beta$
Root-Replace	$\langle degree, n \rangle$, and $\langle slack, n \rangle$	if $\langle degree, \text{the child of } root \rangle \in \beta$ if $\langle slack, \text{the child of } root \rangle \in \beta$
Absorb	possible $\langle slack, n \rangle$	
Split	$\langle weight, n \rangle$, and possible $\langle slack, n \rangle$, and possible $\langle slack, n.p_1 \rangle$, and possible $\langle slack, n.p_2 \rangle$, and possible $\langle slack, parent \rangle$	
Compress	possible $\langle slack, \text{any child of } n \rangle$, and possible $\langle slack, parent \rangle$, and $\langle degree, n \rangle$	if the children of n have total degree $c \leq b$
One-Child	possible $\langle slack, \text{any child of } n \rangle$	

Figure 10.4: Violations that a process P becomes responsible for after performing an update U . Here, n is the topmost node shown on the right-hand side of U 's diagram in Figure 9.1, $parent$ is the node whose child pointer is changed by U (i.e., the parent of n after U), and $root$ is the root of the relaxed B-slack tree (not the sentinel node $entry$).

10.1.3 The rebalancing algorithm

Our algorithm for rebalancing a relaxed B-slack tree is considerably different from the algorithm for rebalancing the relaxed (a,b) -tree. In the relaxed (a,b) -tree, each insertion or deletion can increase the number of violations in the tree by at most one. Whenever a process creates a new violation, it takes responsibility for that violation, and performs rebalancing steps to fix it. Each rebalancing step that the process performs will either fix the violation (eliminating it and decreasing the number of violations in the tree), or move the violation from a node to its parent. Since a violation can move only from a node to its parent, a process can always find the violation it is responsible for by searching for the key it inserted or deleted to create the violation. Thus, a process performs rebalancing by repeatedly searching for a key, and performing rebalancing steps to fix any violations it sees, until it performs a search and sees no violations.

However, in a relaxed B-slack tree, a rebalancing step can create several new violations, which are not necessarily on the search path to any one key. Therefore, processes cannot simply search for a single key, fixing any violations they see.

We think of a violation as a pair $\langle \text{type}, \text{node} \rangle$, where type is *slack*, *degree* or *weight*, and node is a node *in the tree* where the violation occurs. Consider any update U to a relaxed B-slack tree (see Figure 9.1). Observe that U changes a child pointer of a node parent , removing a connected set R of nodes from the tree and inserting a connected set N of nodes into the tree. Let β be the set of violations in the tree before U , and β' be the set of violations in the tree after U . The process that performs U takes responsibility for all violations in $\beta' \setminus \beta$. All of the violations at nodes in R are no longer in the tree after U , and any violations at nodes in N are in the tree after U , but were not in the tree before U . Additionally, U may create a slack violation at parent . Furthermore, observe that U can create violations only at nodes in $N \cup \{\text{parent}\}$, and can create only a slack violation at parent . Therefore, $\beta' \setminus \beta$ is the set of violations at nodes in N , as well as a possible slack violation at parent . (Technically, in our implementation, a process if P performs an update that *could* create a slack violation at parent , then it will take responsibility for any slack violation at parent , even if that violation was in β . So, P *actually* takes responsibility for $(\beta' \setminus \beta) \cup \{\text{any slack violation at } \text{parent} \text{ just after } U\}$.) In Figure 10.4, we precisely characterize the set of violations that a process takes responsibility for after each update to a relaxed B-slack tree. The information in this figure was obtained in a straightforward way by inspecting the updates to relaxed B-slack trees in Figure 9.1.

Note that, after U removes the nodes in R from the tree, any processes that were previously responsible for violations occurring at nodes in R are no longer responsible for them. This approach turns out to be fairly practical. If a process is trying to fix a violation at a node u , and it fails to perform a rebalancing step because u has been removed from the tree, then it can simply proceed as if it had successfully fixed the violation (since the violation has either been eliminated, or another process is now responsible for it).

Consistent with Figure 10.4, after a process performs an Overflow update in `INSERT`, it invokes `FIXWEIGHT` to fix the weight violation at the topmost node n in N , and then invokes `FIXSLACK` to fix any possible slack violation at the parent p of n . Similarly, after a process performs a successful `DELETE`, it invokes `FIXSLACK` to fix any possible slack violation at p .

10.1.4 Fixing weight violations

Pseudocode for `FIXWEIGHT` appears in Figure 10.5. The procedure takes a single argument, node , which points to a node that is known (or suspected) to contain a weight violation. At a high level, `FIXWEIGHT` repeatedly: locates node in the tree, determines whether it contains a weight violation, and attempts to fix the weight violation. It continues to do this until it either successfully fixes the weight violation, or node is removed by another process (so this process is no longer responsible for fixing any violations at node).

We now describe `FIXWEIGHT` in detail. An invocation I of `FIXWEIGHT` begins by checking whether node has weight zero. If not, then there is no weight violation at node , so I simply returns. So, suppose node has weight zero (meaning there is a weight violation at node).

The next step is an (optional) optimization that avoids searching for node if it has already been removed from the tree. This optimization entails invoking `LLX(node)` and checking whether node is `FINALIZED`. It is easy to verify that our lock-free relaxed B-slack tree implementation satisfies the following property: a node is finalized precisely when it is removed from the tree. (This is equivalent to saying that every node removed from the tree by an invocation

```

79 FIXWEIGHT(node)
80   if node.weight = 1 then return
81   ▷ Assert: node is internal (because leaves have weight one), and is not entry or root (because both have weight one)
82   ▷ Optimistically check to see if node was already removed (so we are no longer responsible for its violations)
83   if LLX(node) = FINALIZED then return
84   loop
85     ▷ Start of operation attempt
86     ▷ Search for node and fix any weight violation at node
87     result := SEARCHNODE(node)
88     ▷ If result = FAIL, then another update removed node, so we are no longer responsible for its violations.
89     if result = FAIL then return else  $\langle gp, ix_p, p, ix_l, l \rangle := result$ 
90
91     ▷ We cannot fix a weight violation at l if there is a weight violation at p, so we first check p
92     if p.weight = 0 then
93       ▷ End of operation attempt
94       FIXWEIGHT(p)
95       continue ▷ go to next iteration (to retry)
96
97     ▷ Template iteration 0 (grandparent of node)
98     resultgp := LLX(gp)
99     if resultgp  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then continue ▷ goto next iteration (to retry)
100    if resultgp.pixp  $\neq p$  then continue ▷ CONFLICT: verify gp still points to p
101    ▷ Template iteration 1 (parent of node)
102    resultp := LLX(p)
103    if resultp  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then continue ▷ goto next iteration (to retry)
104    if resultp.pixl  $\neq l$  then continue ▷ CONFLICT: verify p still points to l
105    ▷ Template iteration 2 (node)
106    resultl := LLX(l)
107    if resultl  $\in \{\text{FAIL}, \text{FINALIZED}\}$  then continue ▷ goto next iteration (to retry)
108
109    ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
110    V :=  $\langle gp, p, l \rangle$ ; R :=  $\langle p, l \rangle$ ; fld := a pointer to gp.pixp
111    if p.d + l.d - 1  $\leq b$  ▷ If the contents of l and p fit in a single node
112      n := pointer to a new internal node configured as in Absorb in Figure 9.1. The searchKey of the new node is its first key.
113      success := SCX(V, R, fld, n)
114      ▷ End of operation attempt
115      if success then
116        FIXSLACK(n)
117        return
118      else ▷ The contents of l and p cannot fit in a single node
119        n := pointer to a subtree of three newly created nodes, configured as in Split in Figure 9.1. The searchKey of each new node is its first key.
120        ▷ If n will replace root, then we also perform Root-Zero as part of the same atomic update
121        if gp = entry then n.weight = 1
122        success := SCX(V, R, fld, n)
123        ▷ End of operation attempt
124        if success then
125          if gp  $\neq entry then FIXWEIGHT(n)
126          for each x  $\in \{n, left, right, gp\} do FIXSLACK(x)
127          return$$ 
```

Figure 10.5: Pseudocode for FIXWEIGHT. Here, *b* is the maximum degree of nodes.

of $\text{SCX}(V, R, \text{fld}, \text{new})$ appears in R .) So, if $\text{LLX}(\text{node})$ returns FINALIZED, then I can simply return.

Suppose the invocation of $\text{LLX}(\text{node})$ does not return FINALIZED. Then, I enters a loop wherein it will repeatedly attempt to locate and fix a violation at node until either it succeeds, or node is removed from the tree by another process. Each iteration of this loop follows the tree update template (described in Chapter 5) until just before it invokes FIXWEIGHT or FIXSLACK (described in Section 10.1.5).

The first step in the loop is to search for node by invoking a procedure called SEARCHNODE , which appears in Figure 10.8. SEARCHNODE is similar to SEARCH , except that (1) it uses the searchKey of node to locate it in the tree, (2) the search stops as soon as it encounters node , and (3) it returns FAIL if it does not encounter node . SEARCHNODE either returns FAIL, or $\langle gp, ix_p, p, ix_{\text{node}}, \text{node} \rangle$ such that, during the search, p was read from $gp.p_{ix_p}$ and node was subsequently read from $p.p_{ix_{\text{node}}}$. It is straightforward to prove the following using the techniques in Chapter 8. If SEARCHNODE returns FAIL, then node was removed from the tree at some point before SEARCHNODE terminated. Thus, if I 's invocation of $\text{SEARCHNODE}(\text{node})$ returns FAIL, then the process executing I is no longer responsible for any violations at node , so I can simply return. So, suppose I 's invocation of $\text{SEARCHNODE}(\text{node})$ does not return FAIL.

Fixing any weight violation at p first. Then, I will attempt to fix the weight violation at node by performing Absorb or Split. However, neither of these updates can be performed if the parent p of node has weight zero. So, if p has weight zero, I invokes $\text{FIXWEIGHT}(p)$ to fix it and then skips to the next iteration of the loop (to retry fixing the weight violation at node).

Note that I does not follow the template if it performs this invocation of $\text{FIXWEIGHT}(p)$. We briefly explain why this is the case. I does not invoke SCX , so if it were to follow the template, its final UPDATENOTNEEDED procedure would have to return TRUE. If UPDATENOTNEEDED were to return TRUE, then the template operation would have to be successful and return RESULT. However, the template operation will continue to the next iteration of the loop and perform another attempt after invoking FIXWEIGHT (which conceptually corresponds to the template operation returning FAIL and trying again). In contrast, when I invokes FIXWEIGHT at line 125, it does so after performing a successful SCX , and it subsequently returns successfully. (So, in that case, we can think of the template operation as having terminated successfully before invoking FIXWEIGHT .) As we will see, deviating from the template adds an additional step to the progress proof.

Performing LLXs. Suppose p has weight one. Then, I invokes $\text{LLX}(gp)$. Conceptually, this step represents the beginning of iteration 0 of the tree update template. After invoking $\text{LLX}(gp)$, I verifies that the gp still points to p . This step conceptually represents the CONFLICT procedure in the template. If the LLX returns FAIL or FINALIZED, or if gp does not point to p , then I skips to the next iteration of the loop (to retry). So, suppose I does not skip to the next iteration.

Then, I invokes $\text{LLX}(p)$. This represents the beginning of iteration 1 of the template. Next, I verifies that p still points to node (as part of the template's CONFLICT procedure). If the LLX returns FAIL or FINALIZED, or if p does not point to l , then I skips to the next iteration of the loop. So, suppose I does not skip to the next iteration.

Finally, I invokes $\text{LLX}(l)$ (where $l = \text{node}$). This represents iteration 2 of the template. If the LLX returns FAIL or FINALIZED then I skips to the next iteration of the loop. So, suppose it does not.

Computing SCX-ARGUMENTS, performing SCX and rebalancing. Next, I computes the arguments V, R, fld and n for an invocation of $\text{SCX}(V, R, \text{fld}, n)$ by using locally stored values and immutable fields of nodes. Both Absorb and Split change a pointer of gp (at index ix_p) from p to point to some new node n , removing p and l from the tree. Since

p and l are removed, $R = \langle p, l \rangle$. Since gp is changed, it must be in V . Since R is a subsequence of V , $V = \langle gp, p, l \rangle$. In order to compute n , I must decide whether it should perform Absorb or Split. So, it checks whether the contents of l and p can fit in a single node.

Suppose the contents of l and p can fit in a single node. Then, I attempts to perform Absorb by creating a single new internal node n (configured as shown in Absorb in Figure 9.1), and invoking SCX. (The node l must be internal, since there was a weight violation at l , and weight violations cannot occur at leaves. Thus, the new node created by this update must be internal.) The $searchKey$ of n is simply its first key. If the invocation of SCX returns TRUE, then, in accordance with Figure 10.4, I invokes FIXSLACK to fix any slack violation at n .

Now, suppose the contents of l and p cannot fit in a single node. Then, I attempts to perform Split by creating a subtree of three new nodes (configured as shown in Split in Figure 9.1) rooted at n , and invoking SCX. As we argued in the previous case, l must be internal, so the new nodes must all be internal, as well. The $searchKey$ of each new node is its first key. If $gp = entry$ (and the SCX is successful), then the SCX will replace the $root$ of the relaxed B-slack tree with n . In this case, I also performs Root-Zero as part of the same atomic update (by setting $n.weight = 1$ before invoking SCX). If the invocation of SCX returns TRUE, then, in accordance with Figure 10.4, I invokes FIXWEIGHT to fix a weight violation at n (unless it eliminated the violation by performing Root-Zero as part of the SCX), and then invokes FIXSLACK to fix possible slack violations at gp and each new node.

10.1.5 Fixing degree and slack violations

Pseudocode for FIXSLACK appears in Figure 10.6. The procedure takes a single argument, $node$, which points to a node that is suspected to contain a degree or slack violation. FIXSLACK is quite similar to FIXWEIGHT. It repeatedly: locates $node$ in the tree, determines whether it contains a degree or slack violation, and attempts to fix the violation. It continues to do this until it either successfully fixes the violation, or $node$ is removed by another process.

We give a detailed description of FIXSLACK. An invocation I of FIXSLACK begins by invoking a procedure called NOFIXNEEDED, which appears in Figure 10.8. If NOFIXNEEDED returns TRUE, then I simply returns. This is an (optional) optimization that avoids searching for $node$ if it can determine there is no violation to fix. NOFIXNEEDED is described in detail in Section 10.1.6. Suppose NOFIXNEEDED returns FALSE.

Then, I enters a loop wherein it will repeatedly attempt to locate and fix a degree or slack violation at $node$ (using Compress or One-Child) until either it succeeds, or $node$ is removed from the tree by another process. Each iteration of this loop follows the tree update template (described in Chapter 5) until just before it invokes FIXWEIGHT or FIXSLACK.

As in FIXWEIGHT, the first step in the loop is to search for $node$ by invoking SEARCHNODE. Recall that SEARCHNODE either returns FAIL, or $\langle gp, ix_p, p, ix_{node}, node \rangle$ such that, during the search, p was read from $gp.p_{ix_p}$ and $node$ was subsequently read from $p.p_{ix_{node}}$. If it returns FAIL, then the process executing I is no longer responsible for any violations at $node$, so I can simply return. So, suppose SEARCHNODE does not return FAIL. Then, the local variable l is the same as $node$.

Using Root-Replace to fix a degree violation at the root

Next, I checks whether $gp = NIL$ and $node.d = 1$. If so, then $p = entry$ and $node$ is the $root$. So, I will attempt to perform Root-Replace to eliminate the degree violation. It does this by performing an invocation I' of DOROOTREPLACE(p, ix_l, l). The pseudocode for DOROOTREPLACE appears in Figure 10.7.

```

128 FIXSLACK(node)
129   if NOFIXNEEDED(node) then return                                ▷ Optimistically check if this procedure is needed
130   loop
131     ▷ Start of operation attempt
132     ▷ Search for node and fix any degree or weight violation at node
133     result := SEARCHNODE(node)
134     ▷ If result = FAIL, then another update removed node, so we are no longer responsible for its violations.
135     if result = FAIL then return else  $\langle gp, ix_p, p, ix_l, l \rangle := result$ 

137   ▷ Determine whether Root-Replace should be performed
138   if gp = NIL and node.d = 1 then                                     ▷ Degree violation at node = root
139     result := DOROOTREPLACE(p, ix_l, l)
140     ▷ End of operation attempt
141     if result = RETRY then continue                                     ▷ goto next iteration (to retry)
142     for each x  $\in$  result do FIXSLACK(x)
143     return

145   if node.d > 1 ▷ If there is no degree violation at node, then we are trying to perform Compress
146     Take one extra step in the search, so that p becomes the topmost node in the Compress diagram in
147     Figure 9.1

148   ▷ Note: If there is a degree violation at node, then l = node. Otherwise, p = node.
149   result := DOCOMPRESS(gp, ix_p, p, ix_l, l)
150   ▷ End of operation attempt
151   if result = RETRY then continue                                     ▷ goto next iteration (to retry)
152   for each x  $\in$  result do FIXSLACK(x)
153   return

```

Figure 10.6: Pseudocode for FIXSLACK.

Performing LLXs. I' begins by invoking $\text{LLX}(p)$. Conceptually, this step represents the beginning of iteration 0 of the tree update template. After invoking $\text{LLX}(p)$, I' verifies that the *p* still points to *l*. This step conceptually represents the CONFLICT procedure in the template. If the LLX returns FAIL or FINALIZED, or if *p* does not point to *l*, then I' returns RETRY, which causes I to skip to the next iteration of loop (to retry). So, suppose I' does not return RETRY.

Then, I' invokes $\text{LLX}(l)$. This represents the beginning of iteration 1 of the template. If the LLX returns FAIL or FINALIZED, then I' return RETRY. So, suppose I' does not return RETRY.

Computing SCX-ARGUMENTS. Next, I' computes the arguments *V, R, fld* and *n* for an invocation of $\text{SCX}(V, R, fld, n)$ by using locally stored values and immutable fields of nodes. Root-Replace changes a pointer of *p* (at index *ix_l*) from *l* (which is *root*) to point to some new node *n*, removing *l* from the tree. Since *l* is removed, *R* = $\langle l \rangle$. Since *p* is changed, it must be in *V*. Since *R* is a subsequence of *V*, *V* = $\langle p, l \rangle$. Recall that the degree of *l* = *node* was seen to be one before DOROOTREPLACE was invoked by I . Since the degree of a node never changes, *l* still has degree one. Let *c* be the single child of *l* that was returned by the invocation of $\text{LLX}(l)$ performed by I' . I' creates *n* by copying *c* and setting the *weight* of the new copy to one. If there was a weight violation at *c*, then this update eliminates it.

Performing SCX and subsequent rebalancing. Next, I' invokes $\text{SCX}(V, R, fld, n)$ to replace *l* with *n*. If the invoca-

```

154 DOROOTREPLACE( $p, ix_l, l$ )
155   ▷ Template iteration 0 (parent of  $node$ )
156    $result_p := \text{LLX}(p)$ 
157   if  $result_p \in \{\text{FAIL}, \text{FINALIZED}\}$  then return RETRY
158   if  $result_p.p_{ix_l} \neq l$  then return RETRY           ▷ CONFLICT: verify  $p$  still points to  $l$ 
159
160   ▷ Template iteration 1 ( $node$ )
161    $result_l := \text{LLX}(l)$ 
162   if  $result_l \in \{\text{FAIL}, \text{FINALIZED}\}$  then return RETRY
163   Let  $c$  be the single child pointer in  $result_l$ 
164
165   ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
166    $V := \langle p, l \rangle; R := \langle l \rangle; fld := p.p_{ix_l}$ 
167    $n :=$  a pointer to a newly created copy of  $c$  with weight one
168   ▷ Note: if  $c.weight = 0$  then this update also eliminates a weight violation
169   if  $\text{SCX}(V, R, fld, n)$  then return  $n$ 
170   return RETRY
171
172 DOCOMPRESS( $gp, ix_p, p, ix_l, l$ )
173   ▷ Template iteration 0 (grandparent of  $node$ )
174    $result_{gp} := \text{LLX}(gp)$ 
175   if  $result_{gp} \in \{\text{FAIL}, \text{FINALIZED}\}$  then return RETRY
176   if  $result_{gp}.p_{ix_p} \neq p$  then return RETRY           ▷ CONFLICT: verify  $gp$  still points to  $p$ 
177
178   ▷ Template iteration 1 (parent of  $node$ )
179    $result_p := \text{LLX}(p)$ 
180   if  $result_p \in \{\text{FAIL}, \text{FINALIZED}\}$  then return RETRY
181   if  $result_p.p_{ix_l} \neq l$  then return RETRY           ▷ CONFLICT: verify  $p$  still points to  $l$ 
182
183   ▷ Perform LLXs on the nodes in  $result_p$ 
184    $failed := \text{FALSE}$ 
185   for  $i = 1..p.d$ 
186     ▷ Template iteration  $i + 1$  (child  $i$  of  $p$ )
187     if  $\text{LLX}(result_p.p_i) \in \{\text{FAIL}, \text{FINALIZED}\}$  then return RETRY
188
189   ▷ Before fixing a degree or slack violation at  $node$ , we must fix any weight violations at  $p$  or its children
190   if  $\text{FIXALLWEIGHTVIOLATIONS}(p, result_p)$  then return RETRY ▷ Retry if we fixed a weight violation //  

191
192   ▷ Determine whether there is a slack violation at  $node = p$ , or a degree violation at  $node = l$ 
193    $pGrandDegree := 0$ 
194   for  $i = 1..p.d$  do  $pGrandDegree := pGrandDegree + result_p.p_i.d$ 
195    $slack = p.d * b - pGrandDegree$            ▷ Total slack shared amongst the nodes of  $result_p$ 
196   if  $slack < b$  and  $node.d > 1$  then return  $\emptyset$       ▷ UPDATEROOTNEEDED: no violation at  $node$ 
197
198   ▷ Computing SCX-ARGUMENTS from locally stored values (and immutable fields)
199    $V := \langle gp, p, result_p.p_1, result_p.p_2, \dots \rangle; R := \langle p, result_p.p_1, result_p.p_2, \dots \rangle; fld := gp.p_{ix_p}$ 
200    $\langle \text{doRootReplace}, n, nChildren \rangle := \text{CREATECOMPRESSEDNODES}(gp, p, result_p, pGrandDegree)$ 
201   if  $\text{SCX}(V, R, fld, n)$  then
202     if  $\text{doRootReplace}$  then return  $\{n\}$            ▷ Did Compress/One-Child and Root-Replace
203     else return  $\{gp, n\} \cup nChildren$            ▷ Did Compress/One-Child
204   return RETRY

```

Figure 10.7: Pseudocode for DOCOMPRESS and DOROOTREPLACE. Here, b is the maximum degree of nodes.

```

205 SEARCHNODE(node)
206   ▷ If node is in the tree, this returns  $\langle gp, ix_p, p, ix_{node}, node \rangle$  such that, during the search, p was read from  $gp.p_{ix_p}$ 
207   and node was subsequently read from  $p.p_{ix_{node}}$ . Otherwise, this returns FAIL.
208   key := node.searchKey                                ▷ We locate node using node.searchKey
209   gp :=  $\perp$ ; p := entry; l := p.p1          ▷ Save the last three nodes encountered
210   ixp := 0; ixl := 1                            ▷ Invariant: l was read from  $p.p_{ix_l}$ 
211   while l  $\neq$  node and not l.leaf
212     ixp := ixl                                ▷ Locate appropriate child pointer to follow
213     while ixl < l.d and key  $\geq l.k_{ix_l}$ 
214       ixl := ixl + 1                          ▷ Follow the child pointer
215       gp := p; p := l; l := l.pix_l
216   ▷ Check whether we reached node
217   if l  $\neq$  node then
218     return FAIL ▷ Another update removed l, so we are no longer responsible for any violation at l.
219   return  $\langle gp, ix_p, p, ix_l, l \rangle$ 

220 NOFIXNEEDED(node)
221   if node.leaf return TRUE                                ▷ Degree and slack violations cannot occur at leaves
222
223   ▷ Optimistically check to see if node was already removed (so we are no longer responsible for its violations)
224   result := LLX(node)
225   if result = FINALIZED then return TRUE
226   if result = FAIL then return FALSE                      ▷ Cannot determine whether there is a violation
227   if node.weight = 0 then return FALSE                  ▷ Weight violations must be fixed before others
228   if node.d = 1 then return FALSE                      ▷ Found a degree violation at node
229
230   ▷ Use the snapshot to (try to) determine whether there is a slack violation at node
231   slack := 0; numLeaves := 0
232   for i = 1..node.d
233     slack := slack + (b - result.pi.d)           ▷ Weight violations must be fixed before others
234     if result.pi.weight = 0 then return FALSE
235     if slack  $\geq b$  then return FALSE                  ▷ Possible slack violation at node
236   return TRUE                                         ▷ No violation to fix

237 FIXALLWEIGHTVIOLATIONS(p, resultp)
238   for i = 1..p.d
239     if resultp.pi.weight = 0 then
240       ▷ End of operation attempt
241       foundWeightViolation := TRUE
242       FIXWEIGHT(resultp.pi)
243     if p.weight = 0 then
244       ▷ End of operation attempt
245       foundWeightViolation := TRUE
246       FIXWEIGHT(p)
247   return foundWeightViolation                                ▷ If we fixed a weight violation, retry the search

```

Figure 10.8: Pseudocode for SEARCHNODE, NOFIXNEEDED and FIXALLWEIGHTVIOLATIONS. Here, *b* is the maximum degree of nodes.

```

248 CREATECOMPRESSEDNODES( $gp, p, result_p, pGrandDegree$ )
249   ▷ Determine how to divide keys and values as evenly as possible
250    $numNewChildren := \lceil pGrandDegree/b \rceil$ 
251    $ceilNodes := pGrandDegree \bmod numNewChildren$ 
252    $floorNodes := numNewChildren - ceilNodes$ 
253    $ceilDegree := \lceil pGrandDegree/numNewChildren \rceil$ 
254    $floorDegree := \lfloor pGrandDegree/numNewChildren \rfloor$ 

255   ▷ Note: since there are no weight violations at any nodes in  $result_p$ , they are either all leaves or all internal
256    $pChildrenAreLeaves := (result_p.p_1.leaf)$  ▷ Determine which is the case

257   ▷ Create new node(s)
258   if  $gp = entry$  and  $numNewChildren = 1$  then
259     ▷ If we would replace  $root$  by a node with degree one, then we perform Root-Replace as well
260      $n :=$  pointer to a newly created node that contains all of the keys and pointers of the nodes in  $result_p$ . This node
261       is a leaf if  $pChildrenAreLeaves = \text{TRUE}$ , and is internal otherwise. The  $searchKey$  of  $n$  is the same as the
262          $searchKey$  of  $p$  (since  $n$  will replace  $p$ ).
263     return  $\langle \text{TRUE}, n, \emptyset \rangle$ 
264   else
265      $n :=$  pointer to a subtree consisting of a newly created parent and  $numNewChildren$  newly created children,
266       configured as in Compress in Figure 9.1. The keys and pointers of the nodes in  $result_p$  are divided such
267       that  $ceilNodes$  of the new children have degree  $ceilDegree$ , and  $floorNodes$  of the new children have
268       degree  $floorDegree$ . These children are leaves if  $pChildrenAreLeaves = \text{TRUE}$ , and are internal
269       otherwise. If  $numNewChildren = 1$ , then the  $searchKey$  of each new node is the same as the  $searchKey$  of
270        $p$ , otherwise the  $searchKey$  of each node is its first key.
271     return  $\langle \text{FALSE}, n, \{ \text{children of } n \} \rangle$ 

```

Figure 10.9: Pseudocode for CREATECOMPRESSEDNODES. Here, b is the maximum degree of nodes.

tion of SCX returns FALSE, then I' simply returns RETRY. So, suppose the SCX returns TRUE. Then, according to Figure 10.4, the process performing I' may now be responsible for a degree or slack violation at n . So, I' returns n to I . I then invokes FIXSLACK(n) to fix any degree or slack violation at n .

Using Compress to fix both degree and slack violations

Now, suppose $gp \neq \text{NIL}$ or $node.d \neq 1$, so I will not perform Root-Replace. Instead, it will attempt to perform Compress or One-Child. Although we conceptually view Compress and One-Child as different updates, we can actually use Compress instead of One-Child to fix degree violations. According to Figure 9.1, Compress applies when u has $k \geq 2$ children with total degree $c \leq kb - b$ (so there is a slack violation at u). Suppose we follow the tree update template to implement Compress by replacing u and its children with a new internal node and $\lceil c/b \rceil$ new children (where the keys and pointers in the children of u before the update are evenly distributed amongst the new children after the update). According to Figure 9.1, One-Child applies when $\pi(u)$ has $k \geq 2$ children with total degree $c > kb - b$ (so there is a degree violation at u , but no slack violation at $\pi(u)$). Each of the k children of $\pi(u)$ can have degree at most b , so $kb \geq c > kb - b$, we have $\lceil c/b \rceil = k$. So, whenever the preconditions for One-Child are satisfied, if we simply pretend there is a slack violation at $\pi(u)$, and perform Compress (as if we were going to fix

that violation), then we will create a new internal node and $\lceil c/b \rceil = k$ new children (which evenly share the keys and pointers that were contained in the children of $\pi(u)$ before the update). This is exactly what One-Child does.

Observe that a Compress update where $\pi(u)$ is the topmost node replaced by the update will fix either a slack violation at $\pi(u)$ or a degree violation at u . So, if the violation to be fixed at $node$ is a degree violation, then we fix it by performing Compress where the *parent* of $node$ is the topmost node replaced by the update. In the pseudocode, $l = node$, and p is its parent. So, we fix the violation by performing Compress where p is the topmost node replaced by the update. However, if there is no degree violation at $node$, then the violation we are interested in fixing at $node$ must be a slack violation. In this case, we fix the violation by performing Compress where $node$ (instead of its parent) is the topmost node replaced by the update. In order to facilitate the use of the same code for both cases, if the violation is a slack violation, then I takes one more step in the search. After this extra step, the variable p points to $node$, and l points to some child of $node$. So, as in the previous case, we fix the violation by performing Compress where p is the topmost node replaced by the update.

I performs the necessary Compress update by performing an invocation I' of $\text{DOCOMPRESS}(gp, ix_p, p, ix_l, l)$. Pseudocode for DOCOMPRESS appears in Figure 10.7.

Performing LLXs. I' begins by invoking $\text{LLX}(gp)$. Conceptually, this step represents the beginning of iteration 0 of the tree update template. After invoking $\text{LLX}(gp)$, I' verifies that the gp still points to p . This step conceptually represents the CONFLICT procedure in the template. If the LLX returns FAIL or FINALIZED, or if gp does not point to p , then I' returns RETRY (which causes I to skip to the next iteration of its loop and retry). So, suppose I' does not return RETRY.

Then, I' invokes $\text{LLX}(p)$. This represents the beginning of iteration 1 of the template. Next, I' verifies that p still points to l (as part of the template's CONFLICT procedure). If the LLX returns FAIL or FINALIZED, or if p does not point to l , then I' returns RETRY. So, suppose I' does not return RETRY.

I' then performs LLXs on each node pointed to by the snapshot that was returned by its $\text{LLX}(p)$. These steps correspond to template iterations 2 through $p.d + 1$, where $p.d$ is the degree of p . If any of these invocations of LLX return FAIL or FINALIZED, then I' returns RETRY. So, suppose I' does not return RETRY.

Fixing nearby weight violations first. Compress cannot be performed if there are any weight violations at p or any of its children. So, I' invokes $\text{FIXALLWEIGHTVIOLATIONS}(p, result_p)$ to fix any weight violations at these nodes before proceeding (see Figure 10.8). $\text{FIXALLWEIGHTVIOLATIONS}$ checks for weight violations at p or the nodes in $result_p$, and invokes FIXWEIGHT to fix each violation it finds. $\text{FIXALLWEIGHTVIOLATIONS}$ returns TRUE if it performs any invocation of FIXWEIGHT and FALSE otherwise.

Note that I does not follow the template if this invocation of $\text{FIXALLWEIGHTVIOLATIONS}$ performs an invocation of FIXWEIGHT . As we will see, deviating from the template adds an additional step to the progress proof. If the invocation of $\text{FIXALLWEIGHTVIOLATIONS}(p, result_p)$ returns TRUE, then I' returns RETRY. So, suppose I' does not return RETRY.

Computing UPDATENOTNEEDED. Now that I' has performed $\text{LLX}(gp)$ and $\text{LLX}(p)$, and confirmed that there are no weight violations at p or any of its children, it checks whether a Compress update is necessary. Conceptually, this is part of the UPDATENOTNEEDED procedure in the template. To do this, I' uses the snapshot of p 's children that was returned by its invocation of $\text{LLX}(p)$ to compute the total slack shared amongst the children of p . I' then checks whether there is a slack violation at p , or degree violation at l . (Observe that a slack violation at p and a degree violation at l can each be fixed by a Compress where p is the topmost node replaced by the update.) If there is no

slack or degree violation, then I' simply returns \emptyset (which will cause I to return, successfully, without performing any additional rebalancing). So, suppose I' finds a slack or degree violation.

Computing SCX-ARGUMENTS. Next, I' computes the arguments V, R, fld and n for an invocation of $\text{SCX}(V, R, fld, n)$ by using locally stored values and immutable fields of nodes. Compress changes a pointer of gp (at index ix_p) from p to point to some new node n , removing p and its children from the tree. Since p and its children are removed, $R = \langle p, result_p.p_1, result_p.p_2, \dots \rangle$. Since gp is changed, it must be in V . Since R is a subsequence of V , $V = \langle gp, p, result_p.p_1, result_p.p_2, \dots \rangle$. In order to create the new nodes needed for the Compress update, I' invokes a procedure called `CREATECOMPRESSEDNODES`, which appears in Figure 10.9.

Creating the new node(s). At a high level, `CREATECOMPRESSEDNODES` creates a small subtree of new nodes that would result by applying Compress (and possibly also Root-Replace) to p and its children. These new nodes will be used by `DOCOMPRESS` to replace p and its children in the tree.

More specifically, `CREATECOMPRESSEDNODES` takes four arguments: a pointer gp , a pointer p , a set of pointers $result_p$ and a natural number $pGrandDegree$. These arguments correspond to the variables with the same names in `FIXSLACK`. An invocation I'' of $\text{CREATECOMPRESSEDNODES}(gp, p, result_p, pGrandDegree)$ begins by determining how to divide the $pGrandDegree$ keys and values in the nodes of $result_p$ as evenly as possible into $numNewChildren = \lceil pGrandDegree/b \rceil$ new nodes (where b is the maximum degree of nodes). Specifically, it divides them into $ceilNodes$ nodes with degree $ceilDegree$, and $floorNodes$ with degree $floorDegree$.

Next, I'' determines whether the nodes in $result_p$ are leaves or internal nodes. Observe that, since there are no weight violations at any of the nodes in $result_p$, property P1' of relaxed B-slack trees implies that these nodes are either all leaves or all internal nodes. Therefore, I'' checks whether the first node in $result_p$ is a leaf, and stores the result in a variable $pChildrenAreLeaves$. If so, all nodes in $result_p$ are leaves. Otherwise, all nodes in $result_p$ are internal nodes.

If $gp = entry$ and $numNewChildren = 1$, then Compress would replace the *root* of the tree by a new node with only one child n , necessitating a subsequent Root-Replace update. In this case, I'' will simply cause Root-Replace to be performed at the same time as the Compress update by returning the new child n as the replacement for p . This new node n contains all of the keys and pointers of the nodes in $result_p$. It is a leaf if the nodes in $result_p$ are leaves, and an internal node otherwise. The *searchKey* of n is the same as the *searchKey* of p . Since p was reachable by searching for $p.\text{searchKey}$ before this update, and n replaces node p in the tree (and none of the ancestors of p are replaced or changed by the update, except to replace p by n), n will also be reachable by searching for $p.\text{searchKey}$ after this update. Finally, I'' returns $\langle \text{TRUE}, n, \emptyset \rangle$, which indicates that the newly created node n is the result of performing both Compress and Root-Replace.

Now, suppose $gp \neq entry$ or $numNewChildren \neq 1$. In this case, I'' creates a new parent for the $numNewChildren$ newly created children, and returns points to this parent, and the new children. The new children are leaves if the nodes in $result_p$ are leaves, and are internal nodes otherwise. Two subcases arise.

First, suppose $numNewChildren = 1$. Then, the *searchKey* fields of the single new child and its new parent n are both the same as the *searchKey* of p . By the same argument as in the previous case, n is reachable by searching for $p.\text{searchKey}$ after this update. Moreover, since n has only one child, clearly that child is reachable by searching for the same *searchKey*.

Second, suppose $numNewChildren \neq 1$. Then, the *searchKey* of each new node is its first key. We briefly argue that each new node contains at least one key. Since $numNewChildren \neq 1$, and P2' says that each internal node has

at least one pointer, we obtain $\text{numNewChildren} > 1$. Thus, the new parent node n contains at least one key. Since I'' divides the key and pointers as evenly as possible, and into as few new child nodes as possible, each new child has degree at least $\lfloor b/2 \rfloor$. So, each new child contains at least $\lfloor b/2 \rfloor - 1$ keys. Since $b \geq 5$, $\lfloor b/2 \rfloor - 1 \geq 2$.

Finally, I'' returns $\langle \text{FALSE}, n, \{\text{children of } n\} \rangle$, which indicates that n is the result of performing Compress.

Performing SCX and subsequent rebalancing. After the invocation of CREATECOMPRESSEDNODES by I' returns $\langle \text{doRootReplace}, n, n\text{Children} \rangle$, I' invokes $\text{SCX}(V, R, fId, n)$ to replace the subtree rooted at p with the subtree rooted at n . If the invocation of SCX returns FALSE, then I' simply returns RETRY. So, suppose the SCX returns TRUE. Then, I' checks whether $\text{doRootReplace} = \text{TRUE}$. If so, I' atomically performed both Compress and Root-Replace, so n is the only new node created by the update. Consequently, by Figure 10.4, I need only check for degree or slack violations at n . Thus, I' returns $\{n\}$, which will cause I to invoke $\text{FIXSLACK}(n)$. However, if $\text{doRootReplace} = \text{FALSE}$, then I must check for a degree violation at n , and slack violations at gp and all new children of n . Thus, I' returns $\{gp, n\} \cup n\text{Children}$. I will then invoke $\text{FIXSLACK}(x)$ for each $x \in \{gp, n\} \cup n\text{Children}$.

As a minor point, as we described above, whenever a rebalancing update would necessitate a subsequent Root-Zero update, we always perform Root-Zero as part of the same atomic update. Consequently, in our implementation, there is *never* a weight violation at the *root*.

10.1.6 An optimization to avoid unnecessary searches while rebalancing

An invocation I' of NOFIXNEEDED begins by checking whether node is a leaf. If it is, then I' returns TRUE, since degree and slack violations cannot occur at leaves. So, suppose node is internal. Then, I' performs $\text{LLX}(\text{node})$ and checks whether it returns FINALIZED. If so, node has been removed from the tree by another process (and we are no longer responsible for any violations at node), so I' returns TRUE. Otherwise, if the LLX returns FAIL, then I' cannot determine whether there is a violation at node , so it returns FALSE. So, suppose the LLX does not return FINALIZED or FAIL (which means it returns a snapshot of the pointers of node).

Next, I' checks if node has weight zero. If so, there is a weight violation at node . Since a degree or slack violation at node cannot even be fixed if there is a weight violation at node , the weight violation must be fixed before proceeding, so I' returns FALSE. So, suppose there is no weight violation at node . Next, I' checks if there is a degree violation at node , and returns FALSE if there is. So, suppose there is no degree violation at node .

I' uses the snapshot of node 's children that was returned by the $\text{LLX}(\text{node})$ to check for a slack violation at node . Specifically, it uses the degree of each of these children to compute the total slack shared amongst these children, and it also checks whether there are weight violations at any of these children. If I' finds a weight violation, it must be fixed before proceeding (for the same reasons we gave in the case where there is a weight violation at node), so I' returns FALSE. Suppose I' does not find any weight violations. Then, if the total slack is less than b , there was no degree violation at node when the $\text{LLX}(\text{node})$ occurred, so I' returns TRUE. Otherwise, there may be a slack violation at node , so I' returns FALSE.

10.2 Correctness proof

For convenience, we define a few terms. Consider an iteration of the outer loop in INSERT, DELETE, FIXWEIGHT or FIXSLACK. Within this iteration, we call the sequence of steps between the comments “Start of operation attempt” and “End of operation attempt” an **operation attempt**. More precisely, an operation attempt begins when a process

performs a step just after a “Start of operation attempt” comment, and ends when it reaches an “End of operation attempt” comment, exits the loop, goes to the next iteration, or starts a new operation attempt. Each invocation of `INSERT` or `DELETE` performs a (non-empty) sequence of operation attempts. Each invocation of `FIXWEIGHT` or `FIXSLACK` either returns before entering the loop, or performs a sequence of operation attempts. As we will see below, some operation attempts follow the template, and some do not. The terms **search path** and **range** are defined as they were in the proof of the chromatic tree (in Section 6.3).

Lemma 10.1 *Our implementation of a relaxed B-slack tree satisfies the following claims.*

1. *Each operation attempt A follows the tree update template and satisfies all constraints specified by the template, unless: (1) A occurs in `FIXWEIGHT`, and it invokes `FIXWEIGHT` at line 94, or (2) A occurs in `FIXSLACK`, and it performs an invocation of `FIXWEIGHT` at line 242 or line 246.*
2. *The node entry has weight one, no keys, and a single child pointer to a node root \neq entry with weight one.*
3. *If a node v is in the data structure in some configuration C and v was on the search path for key k in some earlier configuration C', then v is on the search path for k in C. (Equivalently, updates to the tree do not shrink the range of any node in the tree.)*
4. *If an invocation of `SEARCH(k)` reaches a node v, then there was some earlier configuration during the search when v was on the search path for k.*
5. *The `SEARCH` procedure used by `INSERT` and `DELETE` satisfies DTP.*
6. *All operation attempts in `INSERT` and `DELETE` that perform a successful SCX are atomic (including their search phases). All operation attempts in `FIXWEIGHT` and `FIXSLACK` that perform a successful SCX have atomic update phases.*
7. *The tree rooted at the child of entry is always a relaxed B-slack tree.*

Proof: We prove these claims together by induction on the sequence of steps in an execution. That is, we assume that all of the claims hold before an arbitrary step in the execution, and prove they hold after the step.

Claim 1: This claim follows almost immediately from inspection of the code. The only subtlety is showing that no process invokes `LLX(r)`, `FIXWEIGHT(r)` or `FIXSLACK(r)` where $r = \text{NIL}$. Suppose the inductive hypothesis holds just before an invocation I of `LLX(r)`, `FIXWEIGHT(r)` or `FIXSLACK(r)`.

We first observe that, if an invocation of `SEARCH` or `SEARCHNODE` that returns $\langle gp, ix_p, p, ix_l, l \rangle$, then $p, l \neq \text{NIL}$. This follows by inductive Claim 2 and inspection of the code.

Suppose I is an invocation of `FIXWEIGHT(r)`. If I occurs in `INSERT`, then r is a newly created node. If I occurs in `FIXWEIGHT`, then $r = p$ or $r = n$. In the first case, $r \neq \text{NIL}$ since it was returned by `SEARCHNODE`. In the second case, r is a newly created node. If I occurs in `FIXALLWEIGHTVIOLATIONS`, then $r = p$ or $r = result_p.p_i$, where $result_p$ is a result of an `LLX(p)`. In the first case, $r \neq \text{NIL}$ since it was returned by `SEARCHNODE`. In the second case, r was returned by an `LLX(p)` where $p \neq \text{NIL}$, so $r \neq \text{NIL}$ by inductive Claim 7.

Suppose I is an invocation of `FIXSLACK(r)`. If I occurs in `INSERT` or `DELETE`, then $r = p$, where p was returned by `SEARCH` (and hence is non-NIL).

If I occurs in `FIXWEIGHT`, then r is either gp or a newly created node. We argue that $gp \neq \text{NIL}$. The pointer gp was returned by `SEARCHNODE(node)`, which also returned l and p . Since this invocation of `SEARCHNODE` does not return `FAIL`, $l = node$. Also observe that $node = l$ was seen to have weight zero before `SEARCHNODE` was invoked. The *weight* field of a node cannot change, so l always has weight zero. By Claim 2, *entry* and its child always have weight one. So, l cannot be *entry* or its child. It follows that $gp \neq \text{NIL}$.

If I occurs in FIXSLACK, then r is either gp or a newly created node. We argue that $gp \neq \text{NIL}$. By inspection of the pseudocode, if $r = gp$, then I was invoked at line 152. Thus, r is either the first or second pointer returned by SEARCHNODE, depending on whether line 146 is executed. Two subcases arise.

Case 1: suppose $\text{node}.d > 1$ at line 145, so line 146 is executed. By Claim 2, $\text{entry}.d = 1$, so $\text{node} \neq \text{entry}$. Since SEARCHNODE does not return FAIL, the variable l returned by SEARCHNODE is the same as node . Thus, after line 146, $p = \text{node} \neq \text{NIL}$ and $l \neq \text{NIL}$. Since $p \neq \text{entry}$, we have $gp \neq \text{NIL}$.

Case 2: suppose $\text{node}.d \leq 1$ at line 145, so line 146 is not executed. By Claim 7, P2' is satisfied, so $\text{node}.d = 1$. Since we only execute this case if we previously saw that $gp \neq \text{NIL}$ or $\text{node}.d \neq 1$, and we know that $\text{node}.d = 1$, we obtain $gp \neq \text{NIL}$.

Now, suppose I is an invocation of LLX(r). Then, I can occur in INSERT, DELETE, FIXWEIGHT, NOFIXNEEDED, DOROOTREPLACE or DOCOMPRESS.

If I occurs in INSERT or DELETE, then r must be one of the nodes p or l returned by SEARCH. As we argued above, $l, p \neq \text{NIL}$.

If I occurs in FIXWEIGHT(node), then r must either be node , or one of the nodes gp , p or l returned by SEARCHNODE(node). We have already argued above that the argument node is non-NIL. If r is p or l , then $r \neq \text{NIL}$ (as argued above). The argument for $gp \neq \text{NIL}$ is identical to the argument above for the case where I is an invocation of FIXWEIGHT that occurs in FIXWEIGHT.

If I occurs in NOFIXNEEDED(node), then $r = \text{node}$. By inspection of the pseudocode, NOFIXNEEDED(node) is invoked only by FIXSLACK(node), and we have argued above that the argument node to FIXSLACK is non-NIL.

If I occurs in DOROOTREPLACE(p, ix_l, l), then $r = p$ or $r = l$. Observe that DOROOTREPLACE is invoked only by FIXSLACK(node), and the arguments p and l to DOROOTREPLACE were returned by an invocation of SEARCHNODE(node) (along with another pointer gp). As we argued above, $p, l \neq \text{NIL}$.

If I occurs in DOCOMPRESS(gp, ix_p, p, ix_l, l), then $r = gp$ or $r = p$ or $r = l$ or r is one of the pointers returned by an invocation of LLX(p) performed by the DOCOMPRESS. The argument that $l, p \neq \text{NIL}$ is identical to the argument in the previous case. If r was returned by an LLX(p), then $r \neq \text{NIL}$ by inductive Claim 7. The argument that $gp \neq \text{NIL}$ is identical to the case where I is an invocation of FIXSLACK that occurs in FIXSLACK.

Claim 2: The only step that can modify the tree (and, hence, affect this claim) is an invocation S of SCX performed by an invocation I of INSERT, DELETE, FIXWEIGHT or FIXSLACK. Proving this claim entails (1) arguing that I cannot replace the entry point, or modify it in any way except by changing its single child pointer and (2) if I replaces root , then it replaces it with a new node that has weight one.

Suppose the inductive hypothesis holds just before S . By inductive Claim 1, I follows the tree update template up until it performs S . By Lemma 5.1 and Lemma 5.5, the update phase of I is performed atomically. Thus, by inspection of the pseudocode, I atomically performs Delete, Insert, Overflow, Root-Replace, Absorb, Split or Compress, or a combination of Root-Zero and Overflow or Split, or a combination of Root-Replace and Compress.

We first argue that entry cannot be replaced. All of these transformations simply change a single child pointer to replace one or more nodes. However, each node that is replaced has a parent, which entry does not.

Now, suppose I replaces root . The only updates that could replace root by a new node with weight zero are Overflow and Split. However, whenever an Overflow or Split would do so, Root-Zero is also performed as part of the same atomic update (see line 36 and line 121) so that the new root has weight one.

Claim 3: The proof of this claim is very similar to the proof of Lemma 8.1.3. Initially, the claim is trivially true (since

the tree only contains *entry*, which is on every search path). In order for v to change from being on the search path for k in configuration C' to no longer being on the search path for k in configuration C , the tree must change between C' and C . Thus, there must be a successful SCX S between C' and C . Moreover, this is the only kind of step that can affect this claim. We show S preserves the property that v is on the search path for k .

By inductive Claim 1 and inspection of the pseudocode, S is performed by a template operation. Thus, by Lemma 5.1, S changes a pointer of a node from *old* to *new*, removing a connected set R of nodes (rooted at *old*) from the tree, and inserting a new connected set N of nodes. If v is not a descendant of *old* immediately before S , then this change cannot remove v from the search path for k . So, suppose v is a descendant of *old* immediately prior to S .

Since v is in the data structure in both C' and C , it must be in the data structure at all times between C' and C by Lemma 5.3. Therefore, v is a descendant of *old*, but S does not remove v from the tree. Recall that the fringe F_R is the set of nodes that are children of nodes in R , but are not themselves in R (see Figure 5.1 and Figure 5.2). By definition, v must be a descendant of a node $f \in F_R$. Moreover, since v is on the search path for k just before S , so is f . We argue, for each possible tree modification in Figure 9.1, that if any node in F_R is on the search path for k prior to S , then it is still on the search path for k after S . We proceed by cases.

Case 1: Suppose S performs Insert, Overflow or Delete. Since S replaces a leaf with either a new leaf, or a new internal node and two new leaves, the fringe set is empty. Thus, the claim is vacuously true.

Case 2: Suppose S performs Root-Replace or Root-Zero update. Then, S does not change the range of any node in the fringe set, so the claim holds.

Case 3: Suppose S performs Absorb. By inductive Claim 7, the tree is a relaxed B-slack tree before S . Since leaves always have weight one in a relaxed B-slack tree, the node u in the depiction of Absorb in Figure 9.1 must be internal. Thus, one can think of each of the nodes u and $\pi(u)$ as a sequence of alternating pointers and keys, starting and ending with a pointer. Consequently, α , β and γ (in Figure 9.1) can be thought of as sequences of alternating pointers and keys, where α starts with a pointer and ends with a key, β starts and ends with pointers, and γ starts with a key and ends with a pointer. The fringe F_R is the set of nodes pointed to by α , β and γ . Observe that the keys in u and $\pi(u)$ partition the range of $\pi(u)$, and this partition defines the range of each node in F_R . Specifically, the partition begins with the left endpoint of the range of $\pi(u)$, then continues with the alternating pointers and keys of α , β and γ , and finally ends with the right endpoint of the range of $\pi(u)$. The update does not change this partition, so the range of each node in F_R is the same before and after the update. Therefore, if a node in F_R is on the search path to key before the update, it is still on the search path after the update. The cases for Split, Compress and One-Child follow the exact same reasoning.

Claim 4: The proof of this claim is identical to the proof of Lemma 8.1.4 (except for the text substitution “relaxed (a,b) -tree” \rightarrow “relaxed B-slack tree”).

Claim 5: The proof of this claim is very similar to the proof of Lemma 8.1.5. Initially, the claim holds vacuously (since no steps have been taken). Suppose an invocation S of $\text{SEARCH}(k)$ in INSERT terminates and returns m . (The proof for DELETE is similar.) Consider any configuration C , after S returns m , in which all of the nodes in m are in the tree and their fields agree with the values in m . Suppose the inductive hypothesis up until C . We prove that an invocation S' of $\text{SEARCH}(k)$ in INSERT would return m if S' were performed atomically just after configuration C .

The value $m = \langle -, p, l \rangle$ returned by S contains a leaf l and its parent p . By inductive Claim 4, p and l were each on the search path at some point during S (which is before C). Since p and l are in the tree in C , inductive Claim 3 implies that they are on the search path for k in C . Therefore, S' will visit each of them. Conceptually, m also encodes the

fact that p points to l . This fact is checked at line 11 as part of the CONFLICT procedure. By our assumption (that the fields of the nodes in m in configuration C agree with their values in m), p is also the parent of l when S' is performed. Consequently, S' will also return $m = \langle -, p, l \rangle$.

Claim 6: By inductive Claim 5 and Theorem 5.7, all operation attempts that occur in INSERT and DELETE and execute a successful SCX are atomic. By Lemma 5.5, all operation attempts that occur in FIXWEIGHT and FIXSLACK and execute a successful SCX have atomic update phases.

Claim 7: Only successful invocations of SCX can affect this claim. Successful invocations of SCX are performed only in INSERT, DELETE and the rebalancing procedures: FIXWEIGHT and FIXSLACK. The claim holds in the initial state of the tree (which is described in Claim 2). We show that every successful invocation S of SCX preserves the claim. We proceed by cases.

Case 1: S is in an operation attempt A that occurs in an invocation of $\text{INSERT}(key, value)$ or $\text{DELETE}(key)$. By Claim 6, the entire operation attempt A is atomic (including the search phase). Thus, the invocation of SEARCH by A returns the unique leaf l on the search path for key . Consequently, A atomically performs one of the transformations INSERT, OVERFLOW or DELETE to replace l (and possibly some of its neighbouring nodes). Since A is entirely atomic, and it simply performs one of the relaxed B-slack tree updates, it is easy to verify that it preserves the claim.

Case 2: S is in an operation attempt A that occurs in an invocation of one FIXWEIGHT or FIXSLACK. By Claim 6, the update phase of A is atomic (but the search phase is not necessarily atomic). Therefore, A atomically performs one or two rebalancing transformations at some location in the tree (but not necessarily the same rebalancing transformations, at the same location, that it would perform if A 's search were also part of the atomic update). All of the rebalancing transformations preserve the claim (regardless of where in the tree they are performed). ■

We define the linearization points for relaxed B-slack tree operations as follows.

- $\text{GET}(key)$ is linearized at a time during the operation when the leaf reached was on the search path for key . (This time exists, by Lemma 10.1.4.)
- An INSERT is linearized at its successful SCX (if such an SCX exists). (Note: every INSERT that terminates performs a successful SCX.)
- A DELETE that returns \perp is linearized at a time during the operation when the leaf returned by its last invocation of SEARCH was on the search path for key . (This time exists, by Lemma 10.1.4.)
- A DELETE that does not return \perp is linearized at its successful SCX (if such an SCX exists). (Note: every DELETE that terminates, but does not return \perp , performs a successful SCX.)

It is easy to verify that every operation that terminates is linearized, and that each linearized operation has a linearization point that is during the operation.

Theorem 10.2 *The relaxed B-slack tree is a linearizable implementation of a dictionary with the operations GET, INSERT and DELETE.*

Proof: Lemma 10.1.6 proves that each successful SCX atomically performs one or two of transformations shown in Figure 9.1. By inspection of these transformations, the set of keys and associated values stored in leaves are not altered by any rebalancing steps. Moreover, the transformations performed by each linearized INSERT and DELETE maintain the invariant that the set of keys and associated values stored in leaves of the tree is exactly the set that should be in the dictionary. When an invocation of $\text{GET}(key)$ is linearized, the search path for key ends at the leaf returned by its

invocation of SEARCH. If that leaf contains key , GET returns the associated value, which is correct. If that leaf does not contain key , then, by Lemma 10.1.7, it is nowhere else in the tree, so GET is correct to return \perp . ■

10.3 Progress proof

Our goal is to prove that, if processes take steps infinitely often, then relaxed B-slack tree operations succeed infinitely often. At a high level, this follows from Theorem 5.10 (the final progress result for template operations), and Lemma 9.24, which states that at most $2i(4 + \frac{3}{2} \lfloor \log_{\lfloor \frac{b}{2} \rfloor} \frac{n+i}{2} \rfloor) + 2d/(b-1)$ rebalancing steps can be performed after i insertions and d deletions have been performed on an empty B-slack tree.

Theorem 5.10 applies only if processes perform infinitely many template operations, so we must prove that processes will perform infinitely many template operations if they take steps infinitely often. The subtlety is that some operation attempts might not follow the template. Specifically, as we saw in Lemma 10.1.1, an operation attempt A does not follow the template if: (1) A occurs in FIXWEIGHT, and it invokes FIXWEIGHT at line 94, or (2) A occurs in FIXSLACK, and it performs an invocation of FIXWEIGHT at line 242 or line 246. One can imagine a pathology in which non-blocking progress is violated, because processes perform only finitely many template operations, but take infinitely many steps in FIXWEIGHT and/or FIXSLACK. We first prove that this does not happen. Then, we prove the main result.

Observation 10.3 *Nodes are finalized precisely when they are removed from the data structure (and are never reinserted into the data structure).*

Lemma 10.4 *If processes take infinitely many steps in FIXWEIGHT and/or FIXSLACK, then infinitely many template operations are performed.*

Proof: Suppose, to derive a contradiction, that processes take infinitely many steps in FIXWEIGHT and/or FIXSLACK, but only finitely many template operations are performed. Since the tree is changed only by successful invocations of SCX, and SCX is invoked only by template operations, the tree eventually stops changing. Consequently, every invocation of SEARCH or SEARCHNODE terminates after a finite number of steps, unless the process executing it crashes.

Suppose there is some configuration C' after which there are no operation attempts in FIXWEIGHT or FIXSLACK. Then, every invocation of FIXWEIGHT after C' returns at line 80 or line 83, and every invocation of FIXSLACK after C' returns at line 129. By inspection of the code, these invocations of FIXWEIGHT and FIXSLACK are wait-free (since their loops are bounded, and our implementation of LLX is wait-free). Therefore, processes must perform infinitely many invocations of FIXWEIGHT and/or FIXSLACK. Since there are no operation attempts in FIXWEIGHT or FIXSLACK after C' , these invocations must be performed by operation attempts in INSERT and/or DELETE. Each operation attempt in INSERT or DELETE performs a finite number of invocations of FIXWEIGHT and/or FIXSLACK. Thus, there must be infinitely many operation attempts in INSERT and/or DELETE. However, by Lemma 10.1.1, these operation attempts are in fact template operations. Thus, infinitely many template operations are performed. Since this contradicts our assumption, there must be infinitely many operation attempts in FIXWEIGHT and/or FIXSLACK.

Since only finitely many template operations are performed, there must be some configuration C' after which no operation attempt is a template operation. Thus, by Lemma 10.1.1, after C' , every operation attempt in FIXWEIGHT

invokes FIXWEIGHT at line 94, and every operation attempt in FIXSLACK invokes FIXWEIGHT at line 242 or line 246. We consider two subcases.

Case 1: suppose processes begin infinitely many operation attempts in FIXWEIGHT. Let q be any process that takes infinitely many steps in FIXWEIGHT. Consider the first invocation I of FIXWEIGHT(node) by q that begins after C' . Since I is after C' , it performs another invocation I' of FIXWEIGHT($\pi(\text{node})$) at line 94 where $\pi(\text{node})$ was read from node.left or node.right in SEARCHNODE and $\pi(\text{node}).\text{weight}$ was seen to be zero just before I started I' . Since there are no template operations after C' , the tree does not change after C' , so $\pi(\text{node})$ is the parent of node . The invocation I' also performs another invocation of FIXWEIGHT($\pi(\pi(\text{node}))$) where $\pi(\pi(\text{node}))$ is the parent of $\pi(\text{node})$ and $\pi(\pi(\text{node})).\text{weight} = 0$, and so on. By Lemma 10.1.2 and Lemma 10.1.7, node is a descendent of entry , so q will eventually invoke FIXWEIGHT(entry). However, $\text{entry.weight} \neq 0$, so the invocation of FIXWEIGHT(entry) will not invoke FIXWEIGHT at line 94, which is a contradiction.

Case 2: suppose processes begin infinitely many operation attempts in FIXSLACK. We argue that processes also begin infinitely many operation attempts in FIXWEIGHT. Suppose not. Since there are no template operation after C' , every operation attempt that occurs in FIXSLACK and starts after C' must invoke FIXWEIGHT (in FIXALLWEIGHTVIOLATIONS). Furthermore, since we have assumed there are only finitely many operation attempts in FIXWEIGHT, eventually, each invocation of FIXWEIGHT must return before taking any step in the loop. Consequently, eventually, every invocation of FIXWEIGHT(node) must see $\text{node.weight} = 0$ or LLX(node) = FINALIZED. When a process executing FIXALLWEIGHTVIOLATIONS invokes FIXWEIGHT(node), it does so after seeing $\text{node.weight} = 1$. Thus, eventually, every invocation of FIXWEIGHT(node) must see LLX(node) = FINALIZED.

However, as we now argue, invocations of FIXWEIGHT can perform only finitely many invocations of LLX that return FINALIZED. Only finitely many invocations of LLX can be performed by operation attempts that begin before C' . Consider any invocation of FixWEIGHT performed by an operation attempt A that begins after C' . We prove that no invocation of LLX performed by A returns FINALIZED. Before invoking FIXWEIGHT(node), A performs an invocation of SEARCHNODE. This invocation either returns node , or it returns the parent p of node and the operation attempt then performs an invocation of LLX(p) that returns a pointer to node . Since the tree does not change after C' , in each case, node is in the tree during A . Consequently, node cannot be finalized. However, this contradicts our argument above that, eventually, every invocation of FIXWEIGHT(node) must see LLX(node) = FINALIZED. Thus, our assumption that processes only begin finitely many operation attempts in FIXWEIGHT must be invalid. Therefore, the proof of this case follows from the proof of Case 1.

Since both cases lead to a contradiction, our original assumption must be incorrect. Thus, infinitely many template operations must be performed. ■

Theorem 10.5 *The relaxed B-slack tree operations are non-blocking.*

Proof: To derive a contradiction, suppose there is some configuration C after which some processes continue to take steps but no successful relaxed B-slack tree operations occur. We first argue that eventually the tree stops changing. Since no successful relaxed B-slack tree operations occur after C , the only steps that can change the tree after C are successful invocations of SCX performed by invocations of FIXWEIGHT or FIXSLACK. By Lemma 9.24, at most $2i(4 + \frac{3}{2}\lfloor \log_{\frac{b}{2}} \frac{n+i}{2} \rfloor) + 2d/(b-1)$ rebalancing steps can be performed after i insertions and d deletions have been performed on an empty B-slack tree (and then no further rebalancing steps can be applied). Thus, eventually, the tree must stop changing.

This implies that every invocation of SEARCH, SEARCHNODE or GET terminates after a finite number of steps,

unless the process executing it crashes. Consequently, any invocation of GET performed after C will be successful (unless the process executing it crashes). Thus, eventually, no process takes a step in GET.

Suppose processes take infinitely many steps in FIXWEIGHT and/or FIXSLACK after C . Then, by Lemma 10.4, infinitely many template operations are performed. Thus, by Theorem 5.10, infinitely many of these template updates will succeed. However, this contradicts the above argument that the tree eventually stops changing.

Thus, eventually, processes take steps only in INSERT and/or DELETE. Consequently, infinitely many operation attempts are performed in INSERT and DELETE. By Lemma 10.1.1, infinitely many template operations are performed. Moreover, by Theorem 5.10, infinitely many of these template updates will succeed, which contradicts our argument that the tree eventually stops changing. ■

10.4 Balance proof

Lemma 10.6 *In all configurations, for each node u in the tree, u is on the search path to $u.\text{searchKey}$.*

Proof: In the initial configuration, there are only two nodes, *entry* and *root*, and both are on the search path to *every* key, so the invariant holds. We show that any step S by any process P preserves the invariant. Since the *searchKey* field of a node is immutable, the only steps that can affect this invariant are steps that insert nodes into the tree. The only step that can insert a node into the tree is a successful SCXs S (at line 38, line 73, line 113, line 122, line 169 or line 201). S atomically performs one or two of the relaxed B-slack tree updates in Figure 9.1.

It is straightforward to verify that, if S occurs at line 38, line 113 or line 122, then each node u inserted by S contains at least one key, and the *searchKey* of u is its first key. Thus, in each of these cases, Lemma 10.1.7 implies that u is on the search path to $u.\text{searchKey}$. If S occurs at line 73, then S replaces a leaf l with a new node n that has the same *searchKey*. Since we have assumed that l is on the search path to $l.\text{searchKey}$ before S , n is on the search path to $n.\text{searchKey} = l.\text{searchKey}$ after S . The argument is similar for the case where S occurs at line 169. If S occurs at line 201, three cases arise.

Case 1: S replaces an internal node p and its children with a single new node n (created at line 262) that has the same *searchKey* as p . Since we have assumed that p is on the search path to $p.\text{searchKey}$ before S , n is on the search path to $n.\text{searchKey} = p.\text{searchKey}$ after S .

Case 2: S replaces p and its children with a new node n that has a single child n_c (both created at line 265 when $\text{numNewChildren} = 1$), where n and n_c both have the same *searchKey* as p . Since we have assumed that p is on the search path to $p.\text{searchKey}$ before S , n is on the search path to $n.\text{searchKey} = p.\text{searchKey}$ after S . Since n_c is the only child of n , it is also on the search path to $n_c.\text{searchKey} = n.\text{searchKey}$ after S .

Case 3: for each node u inserted by S (created at line 265 when $\text{numNewChildren} > 1$), the *searchKey* of u is its first key. It is straightforward to verify that u contains at least one key (since $\text{numNewChildren} > 1$ and there is at most $b - 1$ slack shared amongst the new children). By Lemma 10.1.7, u is on the search path to $u.\text{searchKey}$. ■

Definition 10.7 *If a process P is between line 208 and line 214, then $\text{location}(P)$ is the value of P 's local variable l and $\text{target}(P)$ is the value of P 's local variable node . Otherwise, $\text{location}(P) = \text{target}(P) = \text{entry}$.*

Lemma 10.8 *In all configurations, for each process P , if $\text{target}(P)$ is not finalized, then $\text{target}(P)$ is on the search path to $\text{target}(P).\text{searchKey}$ starting from $\text{location}(P)$.*

Proof: In the initial configuration, $location(P) = target(P) = entry$ for each process P , so the invariant trivially holds. We show that any step S by any process P preserves the invariant. Since the $searchKey$ field of a node is immutable, the only steps that can affect this invariant are steps that modify child pointers in the tree, and steps that change $target(P)$ or $location(P)$. The only step S that can modify a child pointer in the tree is a successful SCXs at line 38, line 73, line 113, line 122, line 169 or line 201. The only step S that can change $target(P)$ or $location(P)$ is a read of a child pointer at line 208 or line 214.

Case 1: S occurs at line 208 in an invocation of $\text{SEARCHNODE}(node)$. Step S changes $location(P)$ to $entry.p_1$ and $target(P)$ to $node$. By Lemma 10.6, $node$ is on the search path to $node.searchKey$ starting from $entry$ (and, hence, starting from $entry.p_1 = location(P)$).

Case 2: S occurs at line 214. We use $l(P)$ to denote the value of $location(P)$ before S , and $l'(P)$ to denote the value of $location(P)$ after S . Observe that $l(P) \neq node$, since P would have exit the loop at line 210 otherwise. Since (1) we have assumed that $target(P) = node$ was on the search path to $node.searchKey$ from $l(P)$ before S , (2) $l(P) \neq node$, and (3) $l'(P)$ is a child of $l(P)$ that was chosen using $node.searchKey$, $node$ is still on the search path to $node.searchKey$ from $l'(P)$ after S .

Case 3: S is a successful SCX. S atomically performs one or two of the relaxed B-slack tree updates in Figure 9.1, removing a connected set R of nodes rooted at top from the tree, and replacing them with a set N of newly created nodes rooted at n . In order to affect the invariant, S must remove and replace at least one node on the path from $location(P)$ to $target(P)$. Three subcases arise.

Subcase 1: $target(P) \in R$. In this case, $target(P)$ is finalized after S , so the invariant trivially holds.

Subcase 2: $target(P) \notin R$ and $location(P) \in R$. Let F_R be the fringe of R , that is, the set of nodes that are not in R , but are pointed to by a node in R . Before S , the search path to $target(P).searchKey$ starting from $location(P)$ enters R by passing through top , and then exits R by passing through some node $u \in F_R$ (eventually passing through $target(P)$, which is either u or a descendant of U). Since $location(P)$ and the parent of u are both in R , and R is a connected set of nodes, all of the nodes on the path from $location(P)$ to the parent of u are removed from the tree and finalized by S . Consequently, they do not change after they are removed. Thus, the search path to $target(P).searchKey$ starting from $location(P)$ still passes through u after S . Since S does not modify any node on the path from u to $target(P)$, the invariant holds.

Subcase 3: $target(P) \notin R$ and $location(P) \notin R$. Since $location(P), target(P) \notin R$, and R is a connected set of nodes, one of which we have assumed is on the path from $location(P)$ to $target(P)$, $location(P)$ is a proper ancestor of all nodes in R , and some node $u \in F_R$ is an ancestor of $target(P)$. Before S , the search path to $target(P).searchKey$ starting from $location(P)$ enters R by passing through top , and then exits R by passing through u . It is straightforward to verify that, regardless of which transformations in Figure 9.1 are performed by S , after S , the search path to $target(P).searchKey$ starting from $location(P)$ enters N by passing through n , and then exits N by passing through u . (The transformations were designed to satisfy this property.) Since S does not modify any node on the path from u to $target(P)$, the invariant holds. ■

Corollary 10.9 *If an invocation of $\text{SEARCHNODE}(node)$ returns FAIL, then $node$ is finalized.*

Definition 10.10 *Consider an execution in which processes perform invocations of INSERT, DELETE and GET (and these procedures invoke other procedures). In any given configuration, each process is executing a particular procedure, which is either INSERT, DELETE or GET, or was invoked by another procedure. Each process has a call chain,*

which consists of a sequence of invocations I_1, I_2, \dots, I_k where I_{i+1} was invoked by I_i for all i . Observe that, for each process, $I_1 \in \{\text{INSERT}, \text{DELETE}, \text{GET}\}$, and for each $i > 1$, $I_i \notin \{\text{INSERT}, \text{DELETE}, \text{GET}\}$.

Definition 10.11 A process P is **performing weight cleanup for node** in configuration C if its call chain includes $\text{FIXWEIGHT}(node)$. Similarly, P is **performing slack cleanup for node** in configuration C if its call chain includes $\text{FIXSLACK}(node)$.

We now show that each violation in the data structure has a process that is responsible for removing it. Recall that a violation is defined as an ordered pair containing a type and a node.

Lemma 10.12 Consider any execution $C_0 \cdot S_1 \cdot C_1 \cdot S_2 \cdot C_2 \dots$, where $\mathcal{C} = \{C_0, C_1, \dots\}$ is the set of configurations, $\mathcal{S} = \{S_1, S_2, \dots\}$ is the set of steps, and \mathcal{P} is the set of processes that take steps. Let \mathcal{V} and \mathcal{N} be the sets of violations and nodes, respectively, that are ever in the tree. There exists a **responsibility** function $\rho : \mathcal{C} \times \mathcal{V} \rightarrow \mathcal{P} \times \mathcal{N}$ such that, for every configuration C_i , and every weight (resp., slack or degree) violation $x = \langle \text{type}, \text{loc} \rangle$ in the tree in configuration C_i , the following holds. Let $\rho(C_i, x) = \langle P, \text{node} \rangle$.

- (1) $\text{node} = \text{loc}$, and
- (2) P is performing INSERT or DELETE in configuration C_i , and is also performing weight (resp., slack) cleanup for $\text{node}.\text{searchKey}$, or will do so before its INSERT or DELETE terminates, and
- (3) If x is a violation in configuration C_{i-1} , and $\rho(C_{i-1}, x) = \langle Q, - \rangle$ and $\rho(C_i, x) = \langle P, - \rangle$, where $P \neq Q$, then S_i must be a successful SCX by P .

Proof: In the initial configuration, there are no violations, so the invariant is trivially satisfied. We show that any step S_i by any process P preserves the invariant. We assume that ρ satisfies the claim for configurations C_0, C_1, \dots, C_{i-1} and show that it satisfies the claim for the configuration C_i just after S_i . The only step that can cause P to stop performing cleanup for $\text{node}.\text{searchKey}$ is the termination of an invocation of $\text{FIXWEIGHT}(node)$ or $\text{FIXSLACK}(node)$ by P . The only steps that can add and remove nodes and violations are successful SCXs. No other steps S_i can cause the invariant to become false.

Case 1 S_i is the termination of an invocation of $\text{FIXWEIGHT}(node)$ by P . We let $\rho(C_i, x) = \rho(C_{i-1}, x)$ for each violation x in configuration C_i . Part (3) of the invariant is immediate. Since S_i is not a successful SCX, it does not add or remove nodes or violations, so the same violations exist, at the same nodes, in C_{i-1} and C_i . Thus, part (1) of the invariant holds.

By definition, P is performing weight cleanup for $\text{node}.\text{searchKey}$. Recall that the call chain for each process starts with a GET, INSERT or DELETE. If P 's call chain starts with GET, then P cannot be executing FIXWEIGHT , so P must be performing INSERT or DELETE. We show that ρ does not map any weight violation to node in configuration C_{i-1} (and, hence, in configuration C_i), so S_i cannot make part (2) of the invariant become false. S_i is a return statement at line 80, line 83, line 89, line 117 or line 127. Suppose S_i occurs at line 80. Then, $\text{node}.\text{weight} = 1$ in configuration C_{i-1} , so ρ does not map any weight violation to node in C_{i-1} . Now, suppose S_i occurs at line 83. Then, node is finalized before S_i (and it remains finalized thereafter, so it is finalized in configuration C_{i-1}), so ρ does not map any violation to node in C_{i-1} . Now, suppose S_i occurs at line 89. Then, SEARCHNODE searched for $\text{node}.\text{searchKey}$ and returned FAIL, so Corollary 10.9 implies that node is finalized before C_{i-1} (so it is finalized in C_{i-1}). Consequently, ρ does not map any violation to node in C_{i-1} . Now, suppose S_i occurs at line 117 or line 127. Then, node was finalized by the preceding SCX, so ρ does not map any violation to node in C_{i-1} .

Case 2 S_i is the termination of an invocation of $\text{FIXSLACK}(node)$ by P . We let $\rho(C_i, x) = \rho(C_{i-1}, x)$ for each violation x in configuration C_i . Part (3) of the invariant is immediate. Since S_i is not a successful SCX, it does not add or remove nodes or violations, so the same violations exist, at the same nodes, in C_{i-1} and C_i . Thus, part (1) of the invariant holds.

By definition, P is performing slack cleanup for $node.\text{searchKey}$. Recall that the call chain for each process starts with a GET, INSERT or DELETE. If P 's call chain starts with GET, then P cannot be executing FIXSLACK, so P must be performing INSERT or DELETE. We show that ρ does not map any slack or degree violation to $node$ in configuration C_{i-1} (and, hence, in configuration C_i), so S_i cannot make part (2) of the invariant become false. S_i is a return statement at line 129, line 135, line 143 or line 153. We start with the easy cases. If S_i occurs at line 135, then SEARCHNODE searched for $node.\text{searchKey}$ and returned FAIL before C_{i-1} , so Corollary 10.9 implies that $node$ is finalized before C_{i-1} . Consequently, in configuration C_{i-1} , $node$ is finalized, so ρ does not map any violation to $node$. If S_i occurs at line 143 or line 153, then $node$ was finalized by the preceding SCX, so ρ does not map any violation to $node$ in configuration C_{i-1} .

Now, suppose S_i occurs at line 129. Then, it immediately follows an invocation of NOFIXNEEDED that returns TRUE at line 221, line 225 or line 236. If it returns TRUE at line 221, then $node$ is a leaf, so there cannot be any violation at $node$ (so ρ cannot map any violation to $node$ in C_{i-1}). If it returns TRUE at line 225, then $node$ is finalized before C_{i-1} (so ρ cannot map any violation to $node$ in C_{i-1}). If it returns TRUE at line 236, then NOFIXNEEDED performs LLX($node$) and obtains a snapshot of the children of $node$, and sums the (immutable) degree fields of these children, then determines that there was no degree or slack violation at $node$ in the configuration C_j ($j < i - 1$) just before the LLX was performed. Since there was no degree or slack violation at $node$ in C_j , ρ does not map any degree or slack violation to $node$ in C_j . By part (3) of the invariant, prior to S_i , a process can take responsibility for violations *only* by performing a successful SCX, and cannot cause other processes to become responsible for violations. Consequently, since ρ does not map any degree or slack violation to $node$ in configuration C_j , and there is no successful SCX by P after C_j and before S_i (by inspection of the code), ρ cannot map any degree or slack violation to $\langle P, node \rangle$ in configuration C_{i-1} (although ρ *might* map a degree or slack violation to $\langle Q, node \rangle$, where $Q \neq P$).

Case 3 S_i is a successful SCX. Let $parent$ be the node whose pointer is changed by S_i . Recall that x is a violation of type $type$ at loc in configuration C_i (just after S_i). We consider four subcases depending on the value of $x = \langle type, loc \rangle$.

Subcase 1: $type$ is weight or degree and loc was in the tree in configuration C_{i-1} . In this case, we let $\rho(C_i, x) = \rho(C_{i-1}, x)$ (preserving part (3) of the invariant). Since tag bits are immutable and the number of pointers in nodes do not change, no transformation in Figure 9.1 can create a new weight or degree violation at a node that was already in the data structure. Thus, x was a violation (that occurred at loc) in configuration C_{i-1} , so $\rho(C_{i-1}, x)$ is well-defined. Since $\rho(C_i, x) = \rho(C_{i-1}, x)$ and loc is not changed or removed by S_i , parts (1) and (2) of the invariant are preserved.

Subcase 2: $type$ is slack and $loc \neq parent$ was in the tree in configuration C_{i-1} . In this case, we let $\rho(C_i, x) = \rho(C_{i-1}, x)$ (preserving part (3) of the invariant). As Figure 10.4 shows, only DELETE, OVERFLOW, SPLIT and COMPRESS can create a slack violation at $parent$. Furthermore, S_i cannot create a new slack violation at any other node $loc \neq parent$ that was in the tree in configuration C_{i-1} . Thus, x was a violation (that occurred at loc) in configuration C_{i-1} , so $\rho(C_{i-1}, x)$ is well-defined. Since $\rho(C_i, x) = \rho(C_{i-1}, x)$ and loc is not changed or removed by S_i , parts (1) and (2) of the invariant are preserved.

Subcase 3: $type$ is slack and $loc = parent$ was in the tree in configuration C_{i-1} . In this case, we let $\rho(C_i, x) = \langle P, parent \rangle$ (satisfying part (3) of the invariant). By Observation 10.3 and the semantics of SCX, S_i does not remove $parent$ from the tree, so $parent$ is in the tree just after S_i . Since x occurs at loc in configuration C_i and $loc = parent$,

and $\rho(C_i, x) = \langle P, \text{parent} \rangle$, part (1) of the invariant holds.

Recall that the call chain for each process starts with a GET, INSERT or DELETE. If P 's call chain starts with GET, then P cannot perform SCX, so P must be performing INSERT or DELETE. We argue that P performs an invocation I of $\text{FIXSLACK}(\text{parent})$ before its INSERT or DELETE terminates. By Figure 10.4, in order for there to be a slack violation at parent after S_i , S_i must perform Compress (but not Root-Replace), Delete, Overflow or Split. If S_i performs Compress (but not Root-Replace) at line 201, then $\text{doRootReplace} = \text{FALSE}$, so DOCOMPRESS returns $gp = \text{parent}$. Consequently, P performs I at line 152. If S_i performs Delete (at line 73), then P performs I at line 76. If S_i performs Overflow (at line 38), then P performs I at line 43. If S_i performs Split (at line 122), then P performs I at line 126. Thus, part (2) of the invariant holds.

Subcase 4: loc was not in the tree in configuration C_{i-1} . In this case, $\rho(C_i, x) = \langle P, loc \rangle$ (satisfying parts (1) and (3) of the invariant). Recall that the call chain for each process starts with a GET, INSERT or DELETE. If P 's call chain starts with GET, then P cannot perform SCX, so P must be performing INSERT or DELETE. It remains to argue that, if x is a weight (resp., degree or slack) violation, then P performs an invocation I of $\text{FIXWEIGHT}(loc)$ (resp., $\text{FIXSLACK}(loc)$) before its INSERT or DELETE terminates. Since loc is not in the tree before S_i , S_i must have performed an update that created and inserted loc into the tree. Consequently, in the terminology of Figure 10.4, loc is n or a child of n . (We know $loc \neq \text{parent}$, because parent is not newly created by the update.) Figure 10.4 shows precisely where x could be after each type of update.

For example, if S_i performs Split, then x is either a weight violation at $loc = n$ or a slack violation at $loc \in \{n, n.p_1, n.p_2\}$. (We know $loc \neq \text{parent}$, because parent is not a newly created node.) In this case, S_i occurs at line 122. P will invoke FIXSLACK for n and its two children at line 126 before its invocation of INSERT or DELETE terminates. If $gp = \text{entry}$, then n is the new root of the relaxed B-slack tree, and its weight was set to one at line 121, so there is no weight violation at n . Otherwise, P will also invoke $\text{FIXWEIGHT}(n)$ at line 125 before its invocation of INSERT or DELETE terminates. Consequently, part (2) of the invariant holds in this case. The argument for the other updates is very similar. ■

Corollary 10.13 *The lock-free relaxed B-slack tree is a B-slack tree whenever no process is executing INSERT or DELETE.*

Proof: If no process is executing INSERT or DELETE, then Lemma 10.12.2 implies that there are no violations in the tree. A relaxed B-slack tree with no violations is a B-slack tree. ■

Curiously, the proof of Case 2 in the preceding lemma was significantly more complex than the corresponding cases, in the corresponding lemmas, for the Chromatic tree, AVL tree and relaxed (a, b) -tree. We briefly explain why. In the other trees, violations can only be created at newly inserted nodes, because whether a node has a violation depends only on the node's immutable fields. However, in a relaxed B-slack tree, whether a node has a violation depends not only on the node's immutable fields, but also on the immutable fields of its *children*. Since the children of a node can change, it is possible for a node to be in the tree and have no violation in one configuration, and, in a subsequent configuration, be in the tree and *have* a violation. This possibility makes the proof of Case 2 quite subtle.

10.5 Modifications for amortized constant rebalancing

We briefly explain how to modify our pseudocode to obtain amortized constant rebalancing, as described in Section 9.4. This requires two changes. First, the definition of a slack violation is modified slightly, so that a slack

violation occurs at a node u whenever the total slack shared amongst the children of u is less than $b + u.d$ (instead of b). Second, the Compress update changes so that the total degree c shared amongst the children before the update are evenly distributed amongst $\lceil c/(b-1) \rceil$ new nodes after the update (instead of $\lceil c/b \rceil$ new nodes). These changes do not affect the proof.

The specific changes to the pseudocode follow. At line 196 of FIXSLACK, the condition $slack < b$ becomes $slack < b + p.d$. Similarly, at line 235 of NOFIXNEEDED, the condition $slack \geq b$ becomes $slack \geq b + node.d$. Finally, at line 250 of CREATECOMPRESSEDNODES, $numNewChildren$ is set to $\lceil pGrandDegree/(b-1) \rceil$.

10.5.1 Adding a range query operation

A RANGEQUERY operation takes, as its arguments, two keys low and $high$, and returns all key-value pairs present in the dictionary whose keys are in $[low, high]$. These operations are commonly used in databases. Consequently, a RANGEQUERY operation is a common addition to the (ordered) dictionary ADT.

We give a simple implementation of a lock-free RANGEQUERY operation for the relaxed B-slack tree that uses LLX and VLX to obtain a snapshot of the desired range. Recall that a VLX(V) by process p returns TRUE if no node $u \in V$ has changed since p last performed LLX(u), and FALSE otherwise.

The implementation of RANGEQUERY($low, high$) appears in Figure 10.10. It first performs a breadth-first traversal of the tree, pruning any subtrees that it determines cannot intersect $[low, high]$. The traversal visits the children of a node from left to right, and performs LLX on each node it visits. If any LLX returns FAIL or FINALIZED, then a node visited by the traversal has changed, so the RANGEQUERY restarts its traversal. So, suppose all invocations of LLX return snapshots. Then, the RANGEQUERY invoked VLX(V), where V is the sequence of nodes visited by the traversal. If the VLX returns FALSE, then a node visited by the traversal has changed (or replaced), so the RANGEQUERY restarts its traversal. Otherwise, the nodes visited by the traversal form a snapshot, and the RANGEQUERY returns $K \cap [low, high]$, where K is the set of keys in the leaves in V .

Each RANGEQUERY that performs an invocation of VLX(V) which returns TRUE is linearized at this invocation of VLX. We briefly explain why this algorithm is correct. Consider a RANGEQUERY R that returns $K \cap [low, high]$. Since (1) all child pointers followed by the traversal are obtained from snapshots returned by LLX, (2) the VLX succeeds only if none of these nodes have changed, and (3) a node in the tree is excluded from V only if its subtree cannot intersect $[low, high]$, K must contain all keys that are in $[low, high]$ when the VLX occurs. Thus, $K \cap [low, high]$ contains precisely the keys in the tree that are in $[low, high]$ when R is linearized.

Now, we briefly explain why the algorithm satisfies lock-freedom. With this algorithm, individual RANGEQUERY operations can be susceptible to starvation if updates are frequent, but they are guaranteed to succeed in a finite number of steps if there are no updates. Lock-freedom is satisfied as long as updates continue to succeed, and once updates stop succeeding, any ongoing RANGEQUERIES will eventually succeed (satisfying lock-freedom), unless the processes performing them crash.

10.6 Experiments

We implemented the lock-free relaxed B-slack tree in C++, using a fast memory reclamation scheme called DEBRA that is described in Chapter 11. This implementation includes the optimization for amortized constant rebalancing that is described in Section 10.5. We then performed a series of experiments to compare the lock-free relaxed B-slack tree

```

267 RANGEQUERY(low,high)
268 retry:
269   result := empty sequence
270   V := empty sequence
271   q := empty queue

273   ▷ Depth first traversal (of relevant subtrees)
274   q.enqueue(entry)
275   while not q.isEmpty()
276     node := q.dequeue()
277     if LLX(node) ∈ {FAIL,FINALIZED} then goto retry
278     Let  $c_1, c_2, \dots, c_n$  be the child pointers returned by LLX(node)
279     nkeys :=  $n - 1$ 

281   ▷ Visit node
282   Add node to V
283   if node.leaf                                ▷ node is a leaf, so we record its relevant keys
284     Add the keys in node ∩ [low,high] to result
285   else                                         ▷ node is internal, so we explore its children
286     ▷ Find right-most sub-tree that could contain a key in [low,high)
287     r := nkeys
288     while r > 0 and high < node.kr-1      ▷ Subtree rooted at cr contains only keys greater than high
289       r := r - 1                                  ▷ Skip child cr of node
290     ▷ Find left-most sub-tree that could contain a key in [low,high)
291     l := 0                                         ▷ Index of smallest key in node
292     while l < nkeys and low ≥ kl          ▷ Subtree rooted at cl contains only keys less than low
293       l := l + 1                                 ▷ Skip child cl of node

295   ▷ Enqueue relevant children to continue the BFS
296   for i = l..r do q.enqueue(ci)

298   ▷ Validation
299   if not VLX(V) then goto retry
300   return result

```

Figure 10.10: Pseudocode for RANGEQUERY.

with the unbalanced BST and relaxed (a,b) -tree implementations described in Section 8.6. (These implementations also reclaim memory using DEBRA.)

For the relaxed (a,b) -tree, we set $a = 6$ and $b = 16$ (for the same reason that we detailed in Section 8.6). Similarly, for the relaxed B-slack tree, we set the maximum degree b to 16. A single node size of 224 bytes was used for the relaxed B-slack tree and the relaxed (a,b) -tree (just large enough to hold 16 4-byte keys and 8-byte pointers, as well as the meta-data for the tree algorithm and for LLX and SCX). A single SCX-record size of 320 bytes was used for the relaxed B-slack tree and the relaxed (a,b) -tree. In the BST, each node occupies 40 bytes (enough to hold a 4-byte key, 4-byte value, two 8-byte child pointers, and meta-data for LLX and SCX), and each SCX-record occupies 120 bytes. Nodes and SCX-records were not padded.

Experimental system We performed experiments on a 4-socket AMD Opteron 6272 with 16 cores per socket, for a total of 64 threads. Each core has a private 64KB L1 data cache, and 2MB L2 cache that is shared with one other core. All cores on a socket share a 16MB L3 cache. The size of a cacheline is 64 bytes. This system has a non-uniform memory architecture (NUMA) in which threads have significantly different access costs to different parts of memory depending on which processor they are currently executing on. The machine has 128GB of RAM, and runs Ubuntu 14.04 LTS.

All code was compiled with the GNU C++ compiler (G++) 4.8.4 with build target x86_64-linux-gnu and compilation options `-std=c++0x -mcpu=core2 -O3`. Thread support was provided by the POSIX Threads library. We used the scalable allocator jemalloc 4.2.1 [54], which greatly improved performance for all algorithms. We used the Performance Application Programming Interface (PAPI) library [32] to collect statistics from hardware performance counters to determine cache miss rates, stall times, instructions retired, and so on. We pin threads to cores on the different sockets in a round-robin fashion (such that at most one thread is pinned to each core).

10.6.1 Steady-state performance

In this section, we describe a simple randomized benchmark for different workloads consisting of basic dictionary operations (GET, INSERT and DELETE) on uniformly random keys drawn from fixed key ranges. The goal is to study the performance of these basic dictionary operations for each of the three tree algorithms in the steady state (wherein the tree contains approximately half of the keys in the key range, and is neither growing nor shrinking). Since the three tree algorithms are all leaf-oriented, updates always perform modifications tree close to a leaf. Consequently, the internal structure of the tree is somewhat calcified in the steady state (especially near the top of the tree). Since relaxed B-slack trees are intended to be used in workloads with many GETs and few updates, we study workloads containing up to 5% INSERT and 5% DELETE operations.

Our microbenchmark was implemented as follows. For each algorithm $A \in \{\text{Relaxed B-slack tree, Relaxed } (a,b)\text{-tree, BST}\}$ workload $W \in \{0i\text{-}0d, 1i\text{-}1d, 5i\text{-}5d\}$ (where $xi\text{-}yd$ represents $x\%$ INSERTS, $y\%$ DELETES and $(100 - x - y)\%$ GETs) and key range size $S \in \{10^5, 10^6, 10^7\}$, we ran five timed *trials* for several thread counts n . Each trial proceeded in two phases: *prefilling* and *measuring*. In the prefilling phase, n concurrent threads performed 50% *Insert* and 50% *Delete* operations on keys drawn uniformly randomly from $[0, S]$ until the size of the tree converged to a steady state. Note that a steady state is achieved when the tree contains $S/2$ keys, since an insertion or deletion of a key k drawn uniformly from $[0, S]$ is then equally likely to succeed or fail. Next, the trial entered the measuring phase, during which threads began counting how many operations they performed. In this phase, each thread performed random operations

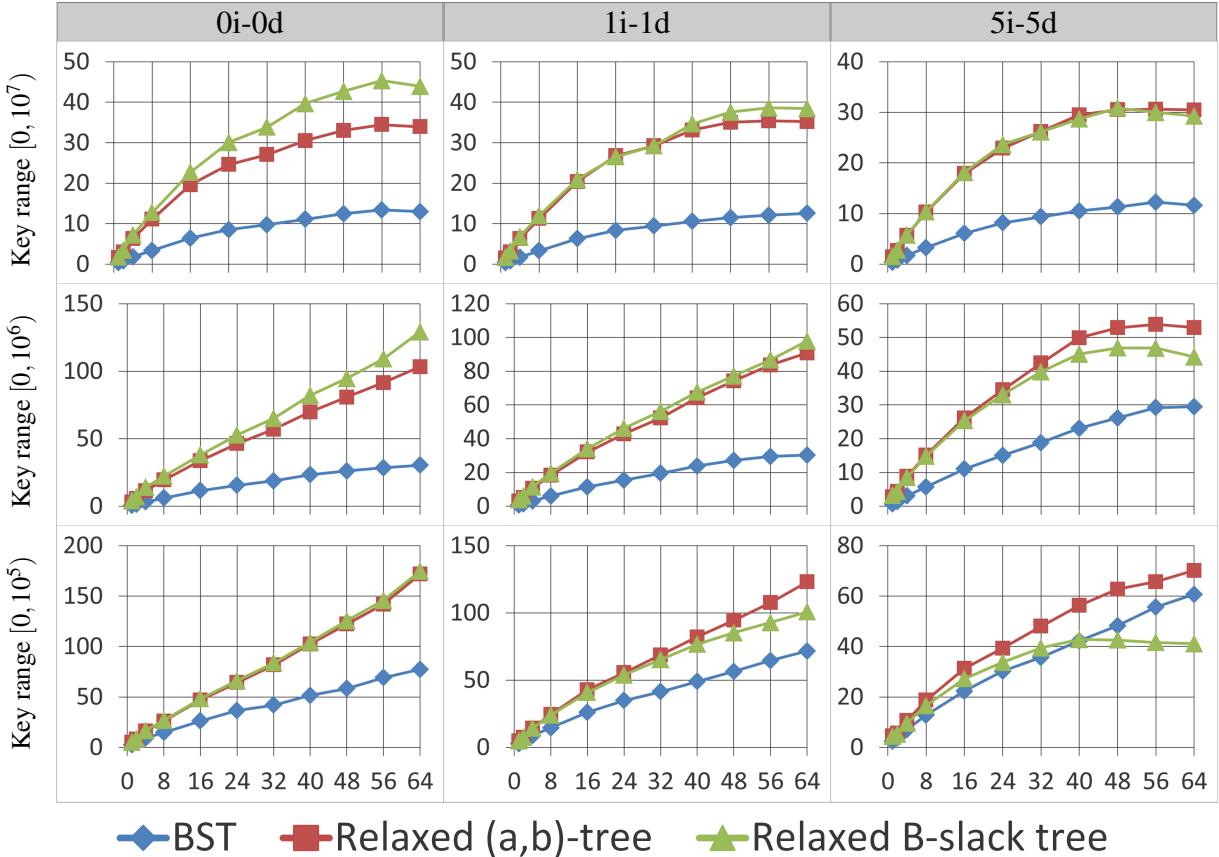


Figure 10.11: Experimental results. The x-axis represents the number of concurrent threads. The y-axis represents operations per microsecond.

according to the workload W on keys drawn uniformly from $[0, S]$ for three seconds.

As a way of validating correctness in each trial, each thread maintains a *checksum*. Each time a thread inserts (resp., deletes) a key, it adds the key to (resp., subtracts from) its checksum. At the end of the trial, the sum of all thread checksums must be equal to the sum of keys in the tree.

Results Results appear in Figure 10.11. As the first column shows, searches are much faster in the trees with large nodes than in the BST. (Note that the BST has been found to perform approximately as well as the Chromatic tree in these types of workloads, since insertions and deletions of uniform random keys yield approximately balanced trees.)

For the two larger key ranges, the relaxed B-slack tree outperforms the relaxed (a,b) -tree. We argue that this is due to the slightly smaller height and better cache utilization of relaxed B-slack tree nodes, which results from their higher average node degree. (Naturally, cache utilization only matters once the tree can no longer entirely fit in cache, which is only the case in the key ranges $[0, 10^6]$ and $[0, 10^7]$.) As an example, with 64 threads and key range $[0, 10^6]$, the relaxed B-slack tree has average node degree 14.54 and height 4, and the relaxed (a,b) -tree has average node degree 9.82 and height 5. Moreover, we used PAPI to measure the number of cycles during which a processor is stalled while waiting for a resource (e.g., a load from main memory), and found that operations experienced approximately 78% more stalled cycles in the relaxed (a,b) -tree. In our analysis, we prefer to use stalled cycles as a metric to explain

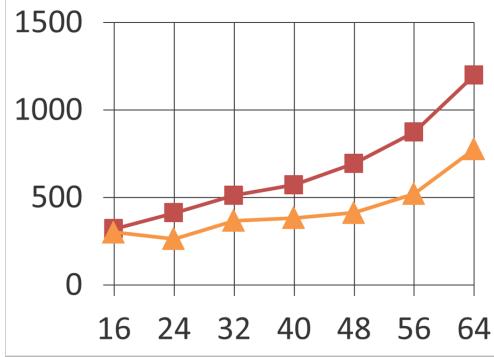


Figure 10.12: Stalled cycle measurements for 0i-0d with key range $[0, 10^7]$. Only the relaxed B-slack tree and the relaxed (a, b) -tree are depicted. The y-axis shows the average number of stalled cycles per operation. The x-axis shows the number of concurrent threads.

performance, rather than cache misses, because cache misses are not all equally costly. Whereas measuring the number of cache misses only captures *how many times* bad things happen, counting stalled cycles captures *how much impact* they have on the execution (in the aggregate).

Note that, for key range $[0, 10^7]$, the failure to scale at high thread counts is due to cache effects, rather than contention. Figure 10.12 shows the average number of stalled cycles per operation versus the number of concurrent threads in case 0i-0d with key range $[0, 10^7]$, for the relaxed B-slack tree and the relaxed (a, b) -tree. (The BST is omitted, since it experiences more than ten times as many stalled cycles than the other trees due to its poor cache utilization, so it would dominate the graph.) Since the L3 cache is shared between all threads on a socket, as the number of threads grows, the effective per-thread L3 cache size decreases. Consequently, threads experience more cache misses, and, hence, more stalled cycles per operation. The relaxed B-slack tree utilizes the cache more effectively, because of its higher node degree, so it experiences fewer stalled cycles.

As the number of update operations increases, the performance of the relaxed B-slack tree decreases relative to the other two algorithms. We believe this is due to the overhead of the relaxed B-slack tree's stricter rebalancing. However, the relaxed B-slack tree still manages to match or outperform the other algorithms in half of the workloads that contain updates. These results suggest that the relaxed B-slack tree is a good alternative to a relaxed (a, b) -tree when the tree is expected to be very large (which is the primary use case for trees with large node degree) and few updates occur.

10.6.2 Tree building performance

In this section, we study the performance of the three tree algorithms in the *prefilling phase* (as they work to reach a steady state). Whereas the internal structure of the tree undergoes relatively few changes in the steady state, the tree structure is completely built from nothing in the prefilling phase. Thus, by measuring the time needed to prefill each data structure, we can get some indication of the cost of building (and balancing) the internal structure of the tree.

We performed *trials* to measure the time needed to prefill each data structure using 64 threads, for range sizes: 2^{17} , 2^{18} , 2^{19} , ..., 2^{27} (from 131,072 to approximately 134 million). Each trial was implemented as a sequence of *prefilling intervals*. At the start of each interval, the main thread creates 64 new threads, each of which waits on a barrier until all threads have been created. Then, all threads perform 50% insertion and 50% deletion operations on keys drawn

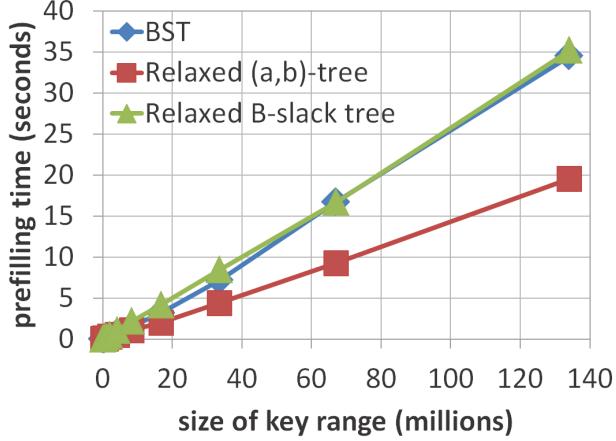


Figure 10.13: Experiment showing the time needed to prefill a data structure with 64 threads, for a variety of key ranges.

uniformly from the key range for 100 milliseconds. After 100 milliseconds, the main thread destroys these 64 threads, and computes the size of the tree. If the size is within 3% of the expected size of the tree in the steady state, then prefilling terminates, and the main thread outputs the number of prefilling intervals that have been performed. This allows us to approximate the time needed to prefill the tree (with a granularity of 100 milliseconds).

The results appear in Figure 10.13. The relaxed (a,b) -tree is a clear winner. Its prefilling takes slightly more than *half* of the time needed for the BST and relaxed B-slack tree. Interestingly, the relaxed (a,b) -tree shows approximately the same performance advantage over the BST and relaxed B-slack tree, although it outperforms them for very different reasons. The relaxed (a,b) -tree outperforms the BST because of the latter's poor cache utilization. However, it outperforms the relaxed B-slack tree (*in spite of* the relaxed B-slack tree's superior cache utilization) because of the high overhead of rebalancing in the relaxed B-slack tree.

Note that prefilling represents a workload in which all operations are updates, which is not an intended use case for the relaxed B-slack tree. In fact, this workload serves as a worst-case for the relaxed B-slack tree.

10.6.3 Memory usage of the final trees

In this section, we study the amount of memory used by a static relaxed B-slack tree that is constructed by many concurrent threads. For each key range, and algorithm, we ran a trial in which 64 threads prefilled the data structure, and then the trial terminated. We then computed the total amount of memory used by nodes in the final data structure.

The results appear in Figure 10.14. There, the memory usage for each data structure is expressed as a percentage value, relative to the memory usage of the BST (which is always 100%). Each data point for the BST is further annotated with its absolute memory usage in megabytes. Consistently, the relaxed (a,b) -tree uses approximately a third of the memory used by the BST, and the relaxed B-slack tree uses approximately one fifth. The relaxed B-slack tree achieves significant space savings over the relaxed (a,b) -tree, which uses between 52% and 60% more memory on average.

A more sophisticated analysis of memory consumption would go beyond the static tree at the end of each trial and study the usage of memory by threads in the process of building the trees. We leave this for future work.

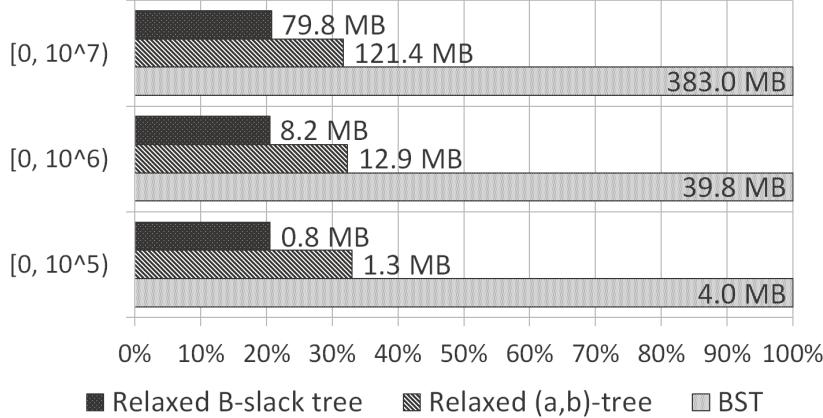


Figure 10.14: Experiment showing the amount of memory occupied by nodes in a static data structure constructed by 64 concurrent threads, for three different key ranges.

10.6.4 Workloads with range queries

In this section, we study workloads containing RANGEQUERY operations. Recall that a RANGEQUERY takes, as its arguments, two keys *low* and *high*, and returns all key-value pairs present in the dictionary whose keys are in $[low, high]$. RANGEQUERY operations can be significantly more efficient in data structures where nodes contain many keys, since they do not need to visit as many nodes to cover the range $[low, high]$.

This experiment was performed the same way as the experiment described in Section 10.6.1, but with different workloads: $W \in \{0i\text{-}0d\text{-}10rq, 1i\text{-}1d\text{-}10rq, 5i\text{-}5d\text{-}10rq\}$, where $xi\text{-}yd\text{-}zrq$ represents $x\%$ INSERT, $y\%$ DELETE, $z\%$ RANGEQUERY and $(100 - x - y - z)\%$ GET operations. For each operation $\text{RANGEQUERY}(low, high)$, we set *low* to be a key drawn uniformly randomly from the key range, and we set $high = low + 1000$. (Thus, in the steady state, when a tree is expected to contain half of the keys in a fixed key range, a RANGEQUERY is expected to return 500 keys.)

Results appear in Figure 10.15. Broadly, the BST performs extremely poorly, and the relaxed B-slack tree performs significantly better than the relaxed (a, b) -tree, except in the smallest key range.

In the BST, RANGEQUERY operations are inefficient, since they must visit many more nodes than in the other trees. Additionally, since the BST uses so much more memory than the other data structures, only a small part of the BST can fit in cache in any workload shown. Thus, the cache performance for the BST is quite poor.

In the relaxed B-slack tree, RANGEQUERY operations visit slightly fewer nodes than in the relaxed (a, b) -tree (because of the relaxed B-slack tree's higher average node degree). Moreover, when compared to the other trees, a larger proportion of the relaxed B-slack tree fits in the cache.

We now discuss the performance differences between the relaxed B-slack tree and the relaxed (a, b) -tree.

In the smallest key range, $[0, 10^5)$, both trees easily fit in cache (occupying less than one tenth of it). In workload $0i\text{-}0d\text{-}10rq$, the relaxed B-slack tree performs slightly better, because its RANGEQUERIES visit approximately 32% fewer nodes, on average. In workloads $1i\text{-}1d\text{-}10rq$ and $5i\text{-}5d\text{-}10rq$, the performance of the relaxed B-slack tree decreases relative to the relaxed (a, b) -tree, because of the higher cost of updates in the relaxed B-slack tree. These updates are more costly primarily because more rebalancing is needed to maintain the stricter balance property. Consequently, if one performs the same sequence of INSERT and DELETE operations in a relaxed B-slack tree and a relaxed (a, b) -tree,

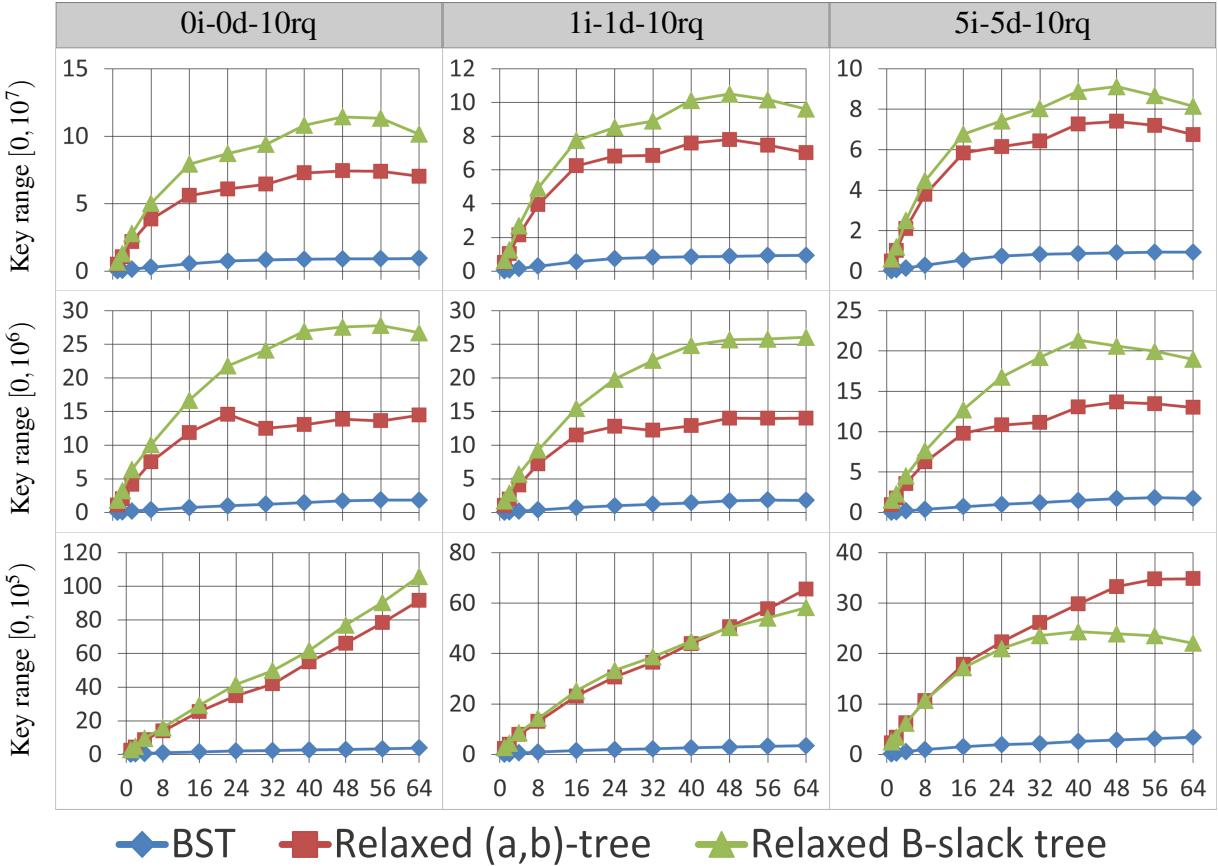


Figure 10.15: Experimental results for workloads including range queries. The x-axis represents the number of concurrent threads. The y-axis represents operations per microsecond.

many more nodes will be modified (or replaced) on average in the relaxed B-slack tree. Each time a node is modified (or replaced) by a process on one socket, last level cache invalidations occur on all other sockets. Hence, the next time a process on another socket accesses the node, it will incur an expensive last level cache miss. Note that this effect is *amplified* in the smallest key range, because the static tree easily fits in cache, so these cross-socket cache invalidations are the primary source of cache misses.

In the middle key range, $[0, 10^6]$, both static trees would fit in entirely the cache (with the relaxed B-slack tree occupying approximately half of the last level cache, and the relaxed (a, b) -tree occupying nearly the entire last level cache), if there were nothing else stored in the cache. However, this was not the case. Consider Figure 10.16(b), which shows that the number of stalled cycles per operation in the relaxed (a, b) -tree increases significantly after 24 threads. This coincides with the flattening of the relaxed (a, b) -tree after 24 threads in the 0i-0d-10rq graph in Figure 10.15. Compare Figure 10.16(b) with Figure 10.16(a), which shows the graph for key range $[0, 10^5]$, wherein both trees are small enough that they easily fit in the last level cache. If the trees in key range $[0, 10^6]$ were actually contained in the last level cache during our trials, then we would expect these two graphs to look similar. The number of stalled cycles per operation also increases with the number of threads for the relaxed B-slack tree (in Figure 10.16(b)), but it increases more slowly, and is much smaller at all thread counts (since the relaxed B-slack tree uses less space).

We briefly consider what else might occupy the last level cache (so that the trees cannot fit). Because of the

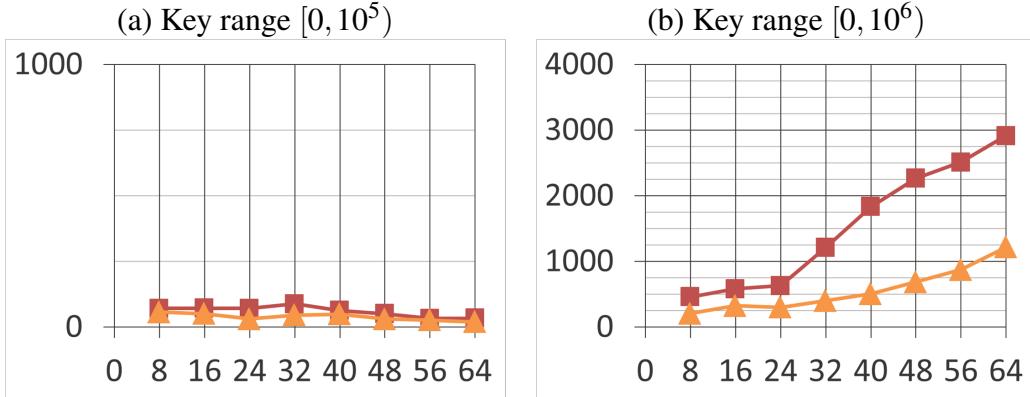


Figure 10.16: Stalled cycle measurements for workload 0i-0d-10rq. Only the relaxed B-slack tree and the relaxed (a, b) -tree are depicted. The y-axis shows the average number of stalled cycles per operation. The x-axis shows the number of concurrent threads.

implementation of LLX and SCX, nodes in the tree point to SCX-records created by the invocations of SCX that last modified them. These SCX-records are accessed by invocations of LLX, which are performed by INSERT, DELETE and RANGEQUERY, so they were also part of certain threads' working sets (and, hence, sometimes appeared in the cache). Threads also use a certain amount of stack space as they perform operations, and the amount used varies throughout the execution (and can grow relatively large, since we implemented rebalancing recursively in both trees). Recent accesses by a thread to its stack memory are cached. Moreover, cache space devoted to these memory accesses grows with the number of threads.

The graphs for the largest key range, $[0, 10^7]$, look similar to the graphs for key range $[0, 10^6]$, except that the performance advantage of the relaxed B-slack tree over the relaxed (a, b) -tree is smaller in $[0, 10^7]$. As we discussed above, part of the performance advantage of the relaxed B-slack tree over the relaxed (a, b) -tree in $[0, 10^6]$ was due to the fact that a significantly larger portion of the relaxed B-slack tree fit in the last level cache. However, in $[0, 10^7]$, only a small part of either tree fits in the cache, so caching effects have less impact on performance. Thus, in these graphs, the performance advantage of the relaxed B-slack tree is primarily due to its higher average node degree, and correspondingly smaller height.

Chapter 11

Reclaiming memory

In this chapter, we study the problem of performing *safe memory reclamation* for lock-free data structures. Consider a linked data structure that contains *records*, which point to one another. Figure 11.1 shows the life-cycle of a record. Initially all records are *unallocated*. A process can *allocate* a record, after which we say the record is *uninitialized*. The process then initializes and inserts the record into the data structure. Eventually, a process may remove the record from the data structure, after which we say the record is *retired*. The goal of safe memory reclamation is to determine when it is safe to *free* a retired record, returning it to an unallocated state. Once it is safe to free a record, one can either free it, or immediately reuse it. In either case, we say that the record has been *reclaimed*.

It is conceptually useful to divide the work of reclaiming an object into two parts: determining when a record is retired, and determining when it is safe to free (because no process can reach it by following pointers). We call a memory reclamation scheme *automatic* if it performs all of the work of reclaiming a record, and *non-automatic* if it requires a data structure operation to invoke a procedure whenever it removes a record from the data structure. Non-automatic memory reclamation schemes may also require a data structure operation to invoke procedures when other events occur, for instance, when the operation begins, or when it accesses a new record. A lock-free algorithm can invoke these procedures only if they are lock-free or wait-free. (Otherwise, they will break the lock-free progress property.)

To specify progress for a memory reclamation scheme, it is also necessary to provide some guarantee that records will eventually be reclaimed. Ideally, one would guarantee that *every* retired record is eventually reclaimed. However, it is impossible to guarantee such a strong property when processes can crash. Since it is not safe to free a record while a process has a pointer to it, and one cannot distinguish between a crashed process and a very slow one, a crashed process can prevent some records from ever being reclaimed. We call a memory reclamation scheme *fault-tolerant* if crashed processes can only prevent a bounded number of records from being reclaimed.

Automatic memory reclamation offers a simple programming environment, but can be inefficient, and is typically not fault-tolerant. Non-automatic memory reclamation schemes are particularly useful when developing data structures for software libraries, since optimizations can pay large dividends when code is reused. Additionally, non-automatic techniques can be used to build data structures that serve as building blocks to construct new automated techniques.

Garbage collectors comprise the largest class of automatic memory reclamation schemes. The literature on garbage collection is vast, and lies outside the scope of this thesis. Garbage collection schemes are surveyed in [75, 111]. Reference counting is another memory reclamation scheme that can be automatic. Limited forms of reference counting are also used to construct non-automatic memory reclamation schemes.

Non-automatic techniques can broadly be grouped into five categories: *unsafe reclamation*, *reference counting*, *hazard pointers*, *epoch based reclamation* and *transactional memory assisted reclamation*. Unsafe reclamation algo-

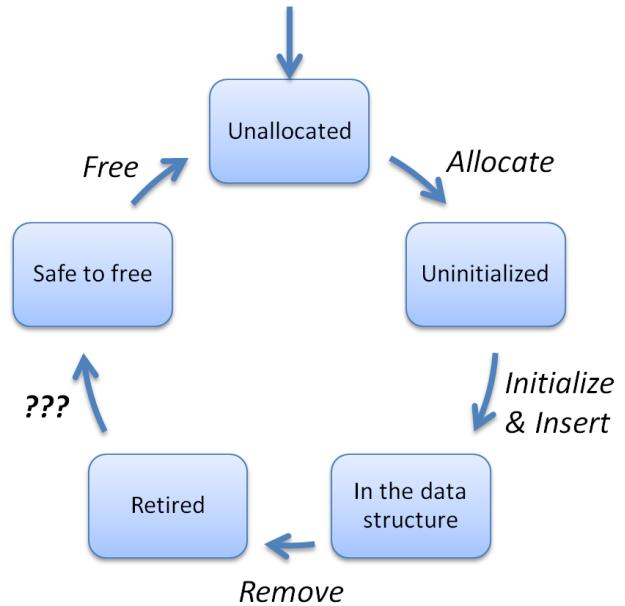


Figure 11.1: The lifecycle of a record.

rithms do not implement safe memory reclamation. Instead, they immediately reclaim records without waiting until they can safely be freed, which can cause numerous problems. For example, suppose a process p reads a field f of a record r and sees A, then performs a CAS instruction to change f from A to B. If r is reclaimed before it can safely be freed, then p can still have a pointer to r after r has been reclaimed. Thus, p 's CAS can be performed after r has been reclaimed. And, if f happens to contain A when p 's CAS is performed, then the CAS will erroneously succeed, effectively changing a different record than the one p intended to change. Unsafe reclamation algorithms must ensure that such problems do not occur. Algorithms in the other categories implement safe memory reclamation, and cannot experience such problems. As will be discussed in Section 11.1, existing techniques either do not work for, or are inefficient for, many natural lock-free data structures.

In epoch based reclamation (EBR), the execution is divided into epochs. Each record removed from the data structure in epoch e is placed into a shared *limbo bag* for epoch e . Limbo bags are maintained for the last three epochs. Each time a process starts an operation, it reads and announces the current epoch, and checks the announcements of other processes. If all processes have announced the current epoch, then a new epoch begins, and the contents of the oldest limbo bag can be reclaimed. If a process sleeps or crashes during an operation, then no memory can be reclaimed, so EBR is not fault tolerant.

The first contribution of this work is DEBRA, a distributed variant of EBR with numerous advantages over classical EBR. DEBRA supports *partial fault tolerance* by allowing reclamation to continue after a process crashes, as long as that process was not performing an operation on the data structure. DEBRA also significantly improves the performance of EBR by: amortizing the cost of checking processes' announced epochs over many operations, eliminating the shared limbo bags in favour of private limbo bags for each process, and optimizing for good cache performance (even on NUMA systems). DEBRA performs $O(1)$ steps at the beginning and end of each data structure operation and $O(1)$ steps each time a record is removed from the data structure.

Our main contribution is DEBRA+, the first fault tolerant epoch based reclamation scheme. The primary challenge was to find a way to allow a process to advance the current epoch without waiting for another process, which may have crashed. In order to advance the epoch, we must ensure that such a process will not access any retired record if it takes another step. The technique we introduce for adding fault tolerance to DEBRA uses signals, an interprocess communication mechanism supported by many operating systems (e.g., Linux and UNIX). With DEBRA+, at most $O(mn^2)$ records are in the limbo bags, waiting to be freed, where n is the number of processes and m is the largest number of records removed from the data structure by one operation.

A major problem arises when auditioning non-automatic techniques to see which performs best for a given lock-free data structure. The set of operations exposed to the programmer by non-automatic techniques varies widely. This is undesirable for several reasons. First, from a theoretical standpoint, there is no reason that a data structure should have to be aware of how its records are allocated and reclaimed. Second, it is difficult to interchange memory reclamation schemes to determine which performs best in a given situation. Presently, doing this entails writing many different versions of the data structure, each version tailored to a specific memory reclamation scheme. Third, as new advancements in memory reclamation appear, existing data structures have to be reimplemented (and their correctness painstakingly re-examined) before they can reap the benefits. Updating lock-free data structures in this way is non-trivial. Fourth, moving code to a machine with a different memory consistency model can require changes to the memory reclamation scheme, which currently requires changing the code for the lock-free data structure as well. Fifth, if several instances of a data structure are used for very different purposes (e.g., many small trees with strict memory footprint requirements and one large tree with no such requirement), then it may be appropriate to use

different memory reclamation schemes for the different instances. Currently, this requires writing and maintaining code for different versions of the data structure.

These issues were all considered in the context of sequential data structures when the C++ standard template libraries were implemented. Their solution was to introduce an *Allocator* abstraction, which allows a data structure to perform *allocate* and *deallocate* operations. With this abstraction, the memory allocation scheme for a data structure can easily be changed without modifying the data structure code at all. Unfortunately, the Allocator abstraction cannot be applied directly to lock-free programming, since its operations do not align well with the operations of lock-free memory reclamation schemes. (For example, it requires the data structure to know when it is safe to free a record.) The main challenge in developing an appropriate generalization of the Allocator abstraction is to find the right set of operations to expose to the data structure implementation, so that the result is simultaneously highly efficient, versatile and easy to program. In Section 11.4, we present our third contribution, the first generalization of the Allocator abstraction for lock-free programming.

Experiments on C++ implementations of DEBRA and DEBRA+ show that overhead is very low. Compared with performing *no reclamation* at all, DEBRA is on average 4% slower, and at worst 21% slower, over a wide variety of thread counts and workloads. In some experiments, DEBRA actually *improves* performance by as much as 20%. Although it seems impossible to achieve better performance when computation is spent reclaiming records, DEBRA reduces the memory footprint of the data structure, which improves memory locality and cache performance. Adding fault tolerance to DEBRA adds 2.5% overhead, on average. Section 11.5 presents extensive experiments comparing DEBRA and DEBRA+ with other leading memory reclamation schemes. For example, DEBRA+ also outperforms a highly efficient implementation of hazard pointers by an average of 75%.

11.1 Related work

There are many existing techniques for reclaiming memory in lock-free data structures. We give a detailed survey of the literature, and identify significant problems with some of the most widely used memory reclamation algorithms, and some of the most recent ones. These problems are poorly understood, and make these reclamation algorithms unsuitable for use with a large class of lock-free algorithms.

Reference Counting (RC). RC augments each record o with a counter that records the number of pointers that processes, entry points and records have to r . A record can safely be freed once its reference count becomes zero. Reference counts are updated every time a pointer to a record is created or destroyed. Naturally, a process must first read a pointer to reach a record before it can increment the record’s reference counter. This window between when a record is reached and when its reference counter is updated reveals the main challenge in designing a RC scheme: the reference count of a freed record must not be accessed. However, the reference count of a retired record can be accessed.

Detlefs et al. [44] introduced lock-free reference counting (LFRC), which is applicable to arbitrary lock-free data structures. LFRC uses the double compare-and-swap (DCAS) synchronization primitive, which atomically modifies two arbitrary words in memory, to change the reference count of a record only if a certain pointer still points to it. DCAS is not natively available in modern hardware, but it can be implemented from CAS [62]. Herlihy et al. [66] subsequently improved LFRC to single-word lock-free reference counting (SLFRC), which uses single-word CAS instead of DCAS. To prevent the reference count of a freed record from being accessed, the implementation of SLFRC

uses a variant of Hazard Pointers (described below) to prevent records from being freed until any pending accesses have finished. Lee [89] developed a distributed reference counting scheme from fetch-and-increment and swap. Each node contains several limited reference counts, and the true reference count for a node is distributed between itself and its parent. The scheme was developed for in-trees (in which each node only has a pointer to its parent), but it may be generalizable.

RC requires extra memory for *each record* and it cannot reclaim records whose pointers form a cycle (since their reference counts will never drop to zero). Manually breaking cycles to allow reclamation requires knowledge of the data structure and adds more overhead. RC has high overhead, since following a pointer involves incrementing, and later decrementing, its reference count, which is expensive. Experiments confirm that RC is less efficient than other techniques [63].

Hazard Pointers (HPs). Michael introduced HPs [97], and provided a wait-free implementation from atomic read/write registers. (Herlihy et al. [66] independently developed another version of HPs called Pass-the-Buck (PTB), providing a lock-free implementation from CAS, and a wait-free implementation from double-wide CAS.) HPs track which records might be accessed by each process. Before a process can access a field of a record r , or use a pointer to r as the expected value for a CAS, it must first acquire a hazard pointer to r . To correctly use HPs, one must satisfy the following constraint. Suppose a record r is retired at time t_r and later accessed by a process p at time t_a . Then, one of p 's HPs must continuously point to r from before t_r until after t_a . This constraint implies that, after a record r is retired and is not pointed to by any HP, no process can acquire a HP to r until it is freed, allocated again, and inserted back into the data structure. Therefore, a process can safely free a retired record after scanning all HPs and seeing that none of them point to r .

To acquire a HP to r , a process first announces the HP by writing a pointer to r in a shared memory location only it can write to. This announces to all processes that r might be accessed and cannot safely be freed. Then, the process verifies that r is in the data structure. If r is not in the data structure, then the process can behave as if its operation had failed due to contention (typically by restarting its operation), without threatening the progress guarantees of the data structure. As we will discuss below, for many data structures, a process cannot easily tell with certainty whether a record is in the data structure. Such data structures are modified in ad-hoc ways so that operations restart whenever they cannot tell whether a record is in the data structure. This requires reproving the data structures' progress guarantees (a subtlety that has been missed by many).

On modern Intel and AMD systems, a *memory barrier* must be issued immediately after a HP is announced to ensure that the announcement is immediately flushed to main memory, where it can be seen by other processes. Otherwise, a HP announcement might be delayed so that a process performing reclamation will miss it, and erroneously free the record it protects. Memory barriers are costly, and this introduces significant overhead.

Many lock-free algorithms require only a small constant number k of HPs per process, since a HP can be released once a process will no longer access the record to which it points during an operation. Scanning the HPs of all processes takes $\Theta(nk)$ steps, so doing this each time a record is retired would be costly. However, with a small modification, the expected amortized cost to retire an object is $O(1)$. Each process maintains a local collection of records that it has removed from the data structure. (Note that each record is only removed by one process.) When a collection contains $nk + \Omega(nk)$ objects, the process creates a hash table T containing every HP, and, for each object o in its collection, checks whether o is in T . If not, o is freed. Since there are at most nk records in T , $\Omega(nk)$ records can be freed. Thus, the number of records waiting to be freed is $O(kn^2)$. To obtain good performance, one typically chooses a fairly large

constant for the $\Omega(kn^2)$ term.

Aghazadeh et al. [6] introduced an improved version of HPs with a worst case constant time procedure for scanning HPs each time a record is retired. Their algorithm maintains two queues of length nk for each process. These queues are used to incrementally scan HPs as records are retired. The algorithm adds a limited type of reference count to each record that tracks the number of incoming references from one of the queues. Note that $\Theta(\log(nk))$ bits are reserved in each record for the reference count.

Recall that, with traditional HPs, a process must issue a costly memory barrier immediately after announcing a HP. Dice et al. [46] recently introduced three techniques for implementing HPs more efficiently by eliminating these frequent barriers. The first technique harnesses the *memory protection* mechanism of modern operating systems. Whenever a process is about to scan all hazard pointers to reclaim memory, it first *write-protects* the memory pages that contain all hazard pointers. Enabling write-protection on a page acts like a global memory barrier that causes all processes to flush any pending writes to the page before it is write-protected. Thus, from the perspective of the process performing memory reclamation, it is as if all processes had been issuing memory barriers immediately after their HP announcements. Unfortunately, this technique is *not* lock-free, since a process that crashes during reclamation will cause all processes to block. Nevertheless, it could be useful for lock-based algorithms that perform searches without acquiring locks. The second technique exploits an idiosyncrasy of certain x86 architectures to achieve a non-blocking variant of the first technique. However, the authors stress that this is not a portable solution. The third technique is a hardware-assisted mechanism that relies on an extension to current processor architectures suggested by the authors.

Problems with Hazard Pointers. A major problem with HPs is that they cannot be used with many lock-free data structures. Recall that in order to acquire a HP to a record r , one must first announce a HP to r , then verify that r is not retired. If one can determine that r is *definitely not retired*, then a HP to r has successfully been acquired. On the other hand, if one can determine that r is *definitely retired*, then the operation can simply behave as if it failed due to contention (typically restarting). However, in many data structures, it is not clear how an operation can determine whether a node is *definitely* retired or not retired. And, if the operation cannot tell for sure which is the case, then behaving as if it failed due to contention can cause the data structure to lose its progress guarantee, as in the following example.

Many data structures with lock-free operations use *marking* to prevent processes from erroneously performing modifications to records just before, or after, they are removed from the data structure. Specifically, before a record is removed from the data structure, it is marked, and no process is allowed to change a marked node. Search operations can often traverse marked nodes, and even leave the data structure to traverse some retired nodes, and still succeed (see, e.g., [10, 30, 31, 49, 52, 51, 65, 70, 101, 108, 113]).

As an example, consider a lock-free singly-linked list in which nodes are marked before they are retired, and operations can traverse retired nodes. Suppose that, while searching this list, a process p has acquired a HP to node u and needs to acquire a HP to the next node, u' . To acquire the HP to u' , p reads and announces the pointer to u' , and must then *verify that u' is in the list*.

One might initially think of checking whether u' is marked to determine whether it is in the list. Since nodes are marked before they are retired, a node that is not marked is definitely not retired. Unfortunately, there are two problems with this approach. First, in order to check whether u' is marked, p would already need to have a HP to u' . Second, if p were to see that u' is marked, then it would learn nothing about whether u' is actually in the list. (Since nodes are marked *before* they are retired, a process may have marked u' but not yet retired it.)

We can get around the first problem by having p check whether the *previous* node u is marked, instead of checking whether u' is marked. (Recall that p already has a HP to u .) If u points to u' and is not marked, then u and u' are definitely both in the list. However, this does not resolve the second problem: if u is marked, we learn nothing about whether u is in the list (and, hence, we learn nothing about whether u' is in the list). All we have done is reduced the problem of determining whether u' is in the list to the problem of determining whether u is in the list.

To resolve the second problem, one might think of having p continue to move backwards in the list until it reaches a node that is not marked. (Since nodes do not change after they are marked, if p were to find a node that is not marked, and points to the chain of marked nodes p traversed, then p would know that all of these nodes are in the list.) However, unless p holds HPs to *every* node it traverses while moving backwards in the list, it will run into the first problem again. In the worst case, p will have to hold HPs to every node it visits during its search. Thus, the number of HPs needed by each process can be *arbitrarily large*. This can introduce significant space and time overhead, and also enables a crashed process to prevent an arbitrarily large number of nodes from being reclaimed. Additionally, since the list is singly-linked, p will have to remember all of the previous nodes that it visited, which may require changes to the algorithm.

Another possible way to verify that u' is in the list is to recursively start a new search looking for u' . Of course, this new search can encounter the exact same problem at a different node on the way to u' . Additionally, p would require additional HPs to perform this search (without releasing the HPs it currently holds on u). Clearly, this approach would be extremely complex and inefficient. Moreover, for some data structures, the amortized cost of operations depends on the number of marked nodes they traverse. Searching again from an entry point can provably lead to an asymptotic increase in the amortized cost of operations [51].

Given that lots of data structures use HPs in practice, one might ask how they deal with these problems. In practice, it is common for data structures that use HPs to simply restart an operation whenever a marked node is encountered. This is usually done without any concern for whether restarting operations will preserve the data structure's progress guarantees. We argue that this will not always preserve lock-freedom. (And we suspect that this will *almost always violate* lock-freedom.) In our example, suppose p sees that u is marked, and restarts its search without being certain that u is actually retired. If the process that marked u crashed before actually retiring u , then after p restarts the search, it will simply encounter u again, see that it is marked, and restart again. Thus, p will restart its search forever, never making progress. Other processes can get into similar infinite loops just as easily, so this can cause all processes to block, violating lock-freedom. It is straightforward to see that restarting in this way will violate lock-freedom for search procedures in many data structures (including [10, 30, 31, 49, 52, 51, 65, 70, 101, 108, 113]).

Additionally, HPs introduce complications in data structures with *helping*. In many lock-free data structures, whenever a process p is prevented from making progress by another operation O (perhaps because O has marked a node that p would like to modify), p *helps* O by performing the steps necessary to complete O (removing any marked nodes from the data structure), on behalf of the process that started O . This involves accessing several nodes, and modifying some of them. Of course, before p can access these nodes, it must acquire HPs to them. As we explained above, in order for p to acquire HPs to these nodes, it must be able to determine that they are definitely not retired. Consequently, one cannot use helping to resolve the problems described above.

Note that, more generally, these problems occur whenever a process can traverse a pointer from a retired node to another retired node (regardless of whether nodes are marked). It appears this scenario can occur in any data structure where retired records can point to records that are still in the data structure (which can later be retired), and searches do not help other operations (which means they cannot restart without violating lock-free progress).

Considering the exposition above, it seems likely that precisely determining whether nodes are retired after announcing HPs is approximately as difficult as maintaining a reference count in each node that counts the number of incoming pointers from other nodes in the data structure.

Beware & Cleanup (B&C). B&C was introduced by Gidenstam et al. [57] to allow processes to acquire hazard pointers to records that have been retired but not reclaimed. A limited form of RC is added to HPs to ensure that a retired record is freed only when no other record points to it. The reference count of a record counts incoming pointers from other records, but does not count incoming pointers from processes' local memories. Before reclaiming a record, a process must verify that its reference count is zero, and that no HP points to it. Consequently, after announcing a HP to a record r , to determine whether r is retired, it suffices to check whether its reference count is nonzero.

Unfortunately, if a data structure allows retired records to point to other retired records, then a retired record's reference count may never be zero. To address this issue, the authors make the following assumption: "each [pointer] in a [retired record] that references another [retired record] can be replaced with a reference to [a record that is not retired], with retained semantics for all of the involved threads." To use B&C, one must implement a procedure that takes a retired record and changes it so that it no longer points to any retired record. Designing such a procedure and proving that it "retains semantics for all of the involved threads" is non-trivial. Additionally, B&C's algorithm for retiring records is extremely complicated, and the technique has significantly higher overhead than HPs.

The authors did not mention the problems with HPs described above. They stated that operations using HPs would need to restart whenever they encountered a retired record, but did not consider how such operations would actually determine whether a record is retired. The goal of the work was simply to *improve performance* for operations that frequently restart. Nevertheless, their technique does solve the problems with HPs described above. Regrettably, it does not appear to be practical.

ThreadScan (TS). TS is a variant of hazard pointers that avoids making an expensive announcement for each record accessed by treating the private memory of each process as if every pointer it contains is an announcement [8]. TS was developed independently, at the same time as DEBRA+. Like DEBRA+, TS uses operating system signals to enable threads obtain a progress guarantee.

Each process maintains a single-writer multi-reader *delete buffer* in shared memory that contains the records it has removed from the data structure, but has not yet freed. When process p 's buffer becomes sufficiently large, p starts *reclamation* by acquiring a global lock (which prevents other processes from starting reclamation). It then collects the records in the buffers of all processes, and sends a signal to every other process. Whenever a process q receives a signal, it scans through its own private memory (stack and registers) and *marks* each record that it has a pointer to. Then, q sends an acknowledgment back to p , indicating that it has finished marking records. Process p waits until it receives acknowledgments from all processes, and then frees any records in its buffer that are not marked.

The authors claim that TS offers strong progress guarantees under the assumption that: "the operating system does not starve threads." However, this assumption is extremely strong. It implies that processes cannot fail, which means that TS cannot be used by a lock-free algorithm. There are two reasons why TS needs this assumption. First, TS uses a global lock to ensure that only one process is performing reclamation at a time. Second, a process performing reclamation must wait to receive acknowledgments from all other processes.

Although TS is not fault-tolerant, it is a very attractive option in practice, since it is easy to use, and it appears to be quite efficient. In order to use TS, one simply invokes a procedure whenever a process has just removed a record

from the data structure. However, TS can only be applied to algorithms that satisfy a set of assumptions. One of these assumptions is particularly problematic: “nodes in [a process’] delete buffer have already been removed [from the data structure], and cannot be accessed through shared references, or through references present on the heap.” Restated using our terminology, this assumption says that records in a process’ delete buffer are retired, and cannot be accessed by following pointers from other records (even other retired records). The authors claim that this assumption “follows from the definition of concurrent memory reclamation [97, 66].” However, it is not clear that this assumption follows from the definitions in those papers.

Consider the relevant part of the definition in [97]: A hazard pointer to a record must be acquired before the record is retired. This admits the possibility that a process can hold hazard pointers to (and, hence, safely access) potentially many retired records, as long as the necessary hazard pointers were acquired before these records were retired. Thus, it is theoretically possible for a process to follow pointers from retired records to other retired records (although there are some problems in practice with traversing pointers from retired records, as we discussed above). Similarly, in the pass the buck algorithm of Herlihy et al. [66], it is safe to access any record that is *in jail*, which means it has been allocated since it was last freed. Since records are retired before being freed, any record that is retired but has not yet been freed is still *in jail*. Thus, it is perfectly fine to follow pointers in shared memory from retired records to other retired records. Therefore, it would appear that TS’s assumption above is strictly stronger than each of these.

Applicability of TS. TS’s assumption prevents it from being used with algorithms where a process can traverse a pointer in shared memory from a retired record to another retired record. This includes all of the algorithms listed where we discussed the problem with HPs. We briefly consider what happens if such an algorithm is used with TS.

Suppose a process p has a pointer in its private memory to a node u that another process q has removed from the data structure, and is about to read a pointer in u that points to another node u' which q has also been removed from the data structure. Then, suppose that, before p reads the pointer in u that it would follow to reach u' , q begins reclamation, and signals all processes, including p . This causes all processes to stop what they are doing and help q perform reclamation, by scanning their private memories for any pointers to nodes in q ’s buffer (including u and u'). If no process has a pointer to u' in its memory, then no process will mark node u' , and process q will free u' . This can occur because p has a pointer to u in its private memory, but does have a direct pointer to u' . However, since u points to u' , after p finishes helping q perform its reclamation, and resumes its regular algorithm, it will follow the pointer from u to u' , performing an illegal access to a freed node.

Dynamic Collect Dragojević et al. [50] explored how hardware transactional memory (HTM) can be used to easily produce several implementations of a *dynamic collect* object. A dynamic collect object has four operations: *Register*, *DeRegister*, *Update* and *Collect*. Intuitively, one can think of a dynamic collect object as a collection of hazard pointers that can increase and decrease in size. Register adds a new HP to the dynamic collect object, and DeRegister removes one. Update sets a HP. An invocation of Collect returns the set of HPs that were set before the Collect began (and were not removed during the Collect), and possibly some others. Similarly to HPs, it is not clear how one could use dynamic collect to reclaim memory for a lock-free data structure wherein processes can traverse pointers from retired records to other retired records.

One interesting observation made by the paper is that, in RC schemes, if a process traverses several records inside a transaction, then some increments and decrements of reference counts can be elided. We explain with an example. Let r_1 , r_2 and r_3 be records in a data structure. Suppose a process executing in a transaction follows a pointer from

r_1 to r_2 , and a pointer from r_2 to r_3 . Observe that it is not necessary to increment or decrement r_2 's reference count, since it would be incremented when the process follows the pointer from r_1 to r_2 , and decremented when the process follows the pointer from r_2 to r_3 (and the atomicity of transactions guarantees that neither change will be visible on its own). In general, if a process follows a chain of pointers inside a transaction, only the reference counts of the first and last records must be updated. The authors described how to efficiently traverse large linked lists by splitting a traversal into many small transactions, and using this technique to eliminate most of the overhead of reference counting.

StackTrack (ST). Alistarh et al. [7] introduced an algorithm called StackTrack, which relies on the cache coherence protocol of the processor to detect conflicts between a process that frees a record and a process that accesses the record. The key idea is to execute each operation of the lock-free data structure in a transaction, and use the implementation of HTM to automatically monitor all pointers stored in the private memory of processes without having to explicitly announce pointers to them before they are accessed. If a transaction accesses a record which is freed during the transaction, then the HTM system will simply abort the transaction, instead of causing the failure of the entire system.

To decrease the probability of aborting transactions, each operation is split into many small transactions called *segments*. This takes advantage of the fact that lock-free algorithms do *not* depend on transactions for atomicity. Segments are executed in order, with each segment being repeatedly attempted until it either commits or has aborted k times. The size of segments is adjusted dynamically based on how frequently segments commit or abort. If segments frequently abort, then they are made smaller (which increases overhead, but makes them more likely to commit), and vice versa. If a segment aborts k times, then a non-transactional fallback code path for the segment is executed instead. This fallback path uses HPs. A process p executing a segment on the fallback path may need to access some records that it obtained pointers to while executing its previous segment. In order to access these records, p must know that they have not been freed. Thus, p must already have HPs to these records when it begins executing on the fallback path. Consequently, at the end of each segment executed *in a transaction* by a process p , all pointers in p 's private memory are announced as HPs.

Any time a process wants to free a record r , it must first verify that no HP points to r . Each time a process removes a record r from the data structure, it places r in a local list. If the size of the list exceeds a predefined threshold, then the process invokes a procedure called *ScanAndFree*, which iterates over each record r in the list, and frees r if it is not announced. ST requires a programmer to insert code before and after each operation, whenever a record is retired, and after every few lines of lock-free data structure code.

A Problem with StackTrack. Although no such constraint is stated in the paper, ST cannot be applied to any data structure in which an operation traverses a pointer from a retired record to another retired record [95]. This includes all of the data structures mentioned above where we discussed the problems with HPs.

We briefly explain what happens when ST is used to reclaim memory for such a data structure. Consider a simple list consisting three nodes, A , B and C , where A points to B and B points to C . For simplicity, we assume the list supports an operation that atomically deletes two consecutive nodes. Suppose process p is searching the list for $C.key$, and process q is concurrently removing B and C . Process p starts a transaction, obtains a pointer to B by reading $A.next$, announces B , and commits the transaction. Then, process q starts a transaction, changes $A.next$ to NULL (to remove B and C from the list), commits the transaction, and adds B and C to its list of removed records. Next, q invokes *ScanAndFree* and sees that p has announced B , so q does not free B . However, p has not announced C , so q frees C . Then, p starts a new transaction and reads $B.next$, obtaining a pointer to C , and then follows that pointer, which causes

the transaction to abort. This transaction will abort every time it is retried. Consequently, the data structure operation cannot make progress on the fast path. Furthermore, as we described above, the HP fallback path cannot accommodate data structures where operations traverse pointers from a retired record to another retired record.

Epochs. A process is in a *quiescent state* whenever it does not have a pointer to any record in the data structure. A grace period is any time interval during which every process has a point when it is in a quiescent state. Fraser [55] described epoch based reclamation (EBR), which assumes that a process is in a quiescent state between its successive data structure operations. More specifically, EBR can be applied only if processes cannot save pointers read during an operation and access them during a later operation. We expand on the brief description of EBR given at the beginning of the chapter. EBR uses a single global counter, which records the current *epoch*, and an announce array. Each data structure operation first reads and announces the current epoch ε , and then checks whether all processes have announced the current epoch. If so, it increments the current epoch using CAS. The key observation is that the period of time starting from when the epoch was changed from $\varepsilon - 2$ to $\varepsilon - 1$ until it was changed from $\varepsilon - 1$ to ε is a grace period (since each process announced a new value, and, hence, started a new operation). So, any records retired in epoch $\varepsilon - 2$ can safely be freed in epoch ε . Whenever a record is retired in epoch ε , it is appended to a limbo bag for that epoch. It is sufficient to maintain three limbo bags (for epochs ε , $\varepsilon - 1$ and $\varepsilon - 2$, respectively). Whenever the epoch changes, every record in the oldest limbo bag is freed, and that limbo bag becomes the limbo bag for the current epoch.

Since EBR only introduces a small amount of overhead at the beginning of each *operation*, it is significantly more efficient than HPs, which requires costly synchronization *each time a new record is accessed*. The penalty for writing to memory only at the beginning of each operation, rather than each time a new record is accessed, is that processes have little information about which records might be accessed by a process that is suspected to have crashed. Consequently, EBR is not fault tolerant.

Quiescent state-based reclamation (QSBR) [96] is a generalization of EBR that can be used with data structures where processes can save pointers read during an operation and access them during a later operation. However, to use QSBR, one must manually identify times when individual processes are quiescent.

Drop the Anchor (DTA). Braginsky, Kogan and Petrank [23] introduced DTA, a specialized technique for singly-linked lists, which explores a middle ground between HPs and EBR. Instead of acquiring a HP each time a pointer to a node is read, a HP is acquired only once for every c pointers read. When a HP to a node u is acquired, it prevents other processes from reclaiming u and the next $c - 1$ nodes currently in the list. (It also prevents other processes from reclaiming any nodes that are inserted amongst these nodes.) Suppose a process q performs s operations without seeing any progress by another process p . Then, q will cut all nodes that p might access out of the list, replacing them with new copies. It will also mark the old nodes so that p can tell what has happened. If p has crashed, then the nodes that q cuts out of the list can never be freed. However, if p has not crashed, then it will eventually see what has happened, and attempt to free these marked nodes. Observe that memory reclamation can continue in the list regardless of whether p has crashed. Consequently, DTA is fault tolerant.

DTA has been shown to be efficient [7, 23]. However, it is not clear how it could be extended to work for other data structures. Additionally, DTA needs to be integrated with the mechanism for synchronizing updates to the linked list, because a sequence of nodes can be cut out of the list concurrently with other updates.

In the worst case, the number of retired nodes that cannot be freed is $\Omega(scn^2)$. To see why, consider the following.

Let p be a process with a HP at a node u_1 that allows it to access nodes u_1, \dots, u_c . Suppose each process other inserts s nodes between u_1 and u_c before p performs another step. Observe that p 's next step might access any of the $\Omega(scn)$ nodes starting at u_1 and ending at u_c . Thus, a process q that suspects p has crashed will cut all of these nodes out of the list. None of these nodes can be freed if p crashes. If this is repeated for each process $p \neq q$, then there will be $\Omega(scn^2)$ nodes that cannot be reclaimed.

QSense (QS). Recently, Balmau et al. [16] introduced QS, another algorithm that combines HPs and EBR. Like DTA, QS adds fault-tolerance to EBR by using HPs. Like the accelerated implementations of HPs in [46], the performance benefit of QS over HPs comes from a reduction in the overhead of memory barriers issued to ensure that HP announcements are visible to all threads. At a high level, QS uses two execution paths: an EBR-based fast path and a HP-based slow path. As long as processes continue to make progress, they continue to use the fast path. If a process has not made progress for a sufficiently long time, then all processes switch to the slow path. To guarantee that nodes are not erroneously freed when the algorithm switches from the fast path to the slow path, the fast path and slow path both acquire HPs. On the fast path, EBR is effectively used to reduce the cost of reclamation by eliminating the need to scan HPs to determine whether nodes can be freed.

In order to reduce the overhead of issuing memory barriers for the HPs acquired in QS, the authors make the following observation: In a modern operating system, running on an x86/64 architecture, whenever a process experiences a context switch, the kernel code for performing a context switch issues at least one memory barrier. Suppose a process p announces a HP at time t , and does *not* perform a memory barrier after announcing the HP. Additionally, suppose another process q begins scanning HPs (to perform reclamation) at time $t' > t$. If p experiences a context switch between t and t' , then q will see p 's HP announcement, as if p had issued a memory barrier. (This is somewhat similar to the HP schemes of Dice et al. [46] discussed above, which also harness operating system and/or hardware primitives to eliminate memory barriers.)

The authors introduce *rooster processes* to trigger context switches for all processes at regular intervals. Each processor has a rooster process pinned to it that sleeps for some fixed length of time T , then wakes up (forcing a context switch), then immediately sleeps again. Whenever a process wants to reclaim a node u that was retired at time t , it waits until time $t + T + \varepsilon$ (for some small $\varepsilon > 0$), by which point a rooster process should have woken up, forcing a context switch and guaranteeing that the reclaiming process can see any HPs announced before the node was removed from the data structure. The ε term above is necessary because, in real systems, when a process requests to sleep for T time units, it may sleep longer. If a rooster process sleeps for more than $t + T + \varepsilon$ time, then its failure to trigger a timely context switch might cause a reclaiming process to miss a HP and erroneously free a node that is still in use. Thus, *QS only works under the following assumption*: a bound on ε must be known and rooster processes never fail. Consequently, QS does not work in a fully asynchronous system. (In comparison, the TS algorithm also makes timing assumptions, but it only loses its *progress* guarantee in a fully asynchronous system.)

QS switches from the slow path to the fast path only when every process has completed an operation since it last switched to the slow path. Consequently, if a process crashes either before or while processes are executing on the slow path, then all processes remain on the slow path forever (where they cannot use EBR to reclaim memory). This is not true for DEBRA+, which allows EBR to continue, even if a process has crashed. Furthermore, since QS uses HPs, it cannot be used with data structures where operations traverse pointers from a retired record to another retired record (as we described above).

Optimistic Access (OA). Cohen and Petrank [36] introduced an approach where algorithms optimistically access

parts of memory that might have already been reclaimed, and check after each read whether this was the case. (Their approach was developed independently, at the same time as DEBRA+.) Crucially, OA relies on processes being able to access reclaimed memory without causing the system to crash. Consequently, an algorithm that uses OA must either (a) never release memory to the operating system, or (b) trap segmentation fault and bus fault signals and ignore them. If option (a) is used, then OA has the same downsides as OPs. On the other hand, if option (b) is used, then one must cope with the downsides of trapping segmentation and bus faults. In particular, if an algorithm contains bugs that cause segmentation or bus faults, then option (b) makes it significantly more difficult to identify these bugs, because one cannot easily distinguish between faults caused by a program bug and faults caused by accesses to reclaimed memory. Such algorithms cannot be used in software that already traps segmentation or bus faults, including many common debugging tools. (Incidentally, like DEBRA+, option (b) is lock-free only if the operating system's signaling mechanism is lock-free.)

OA can be used only with algorithms that appear in a *normalized* form. At a high level, an operation in a normalized algorithm proceeds in three phases: CAS generation, CAS execution and wrap-up. In the CAS generation phase, the operation reads shared memory and constructs a sequence of CAS steps. These CAS steps are performed in the CAS execution phase. Any additional processing is performed in the wrap-up phase. In the CAS execution and wrap-up phases, the operation can be helped by other processes. As we will see, the class of algorithms in lock-free normalized form is quite similar to the class of algorithms that can use DEBRA+ in a straightforward way.

At a high-level, OA works as follows. Each process p has a *warning* bit that is cleared whenever p starts a new operation, and is set whenever any process begins reclaiming memory. After p performs a read from shared memory, it checks whether its warning bit is set, and, if so, jumps back to a safe checkpoint earlier in the operation (intuitively restarting the operation). Unlike reads, CAS operations cannot be optimistically performed on reclaimed memory, since this could cause data corruption. Thus, *before* performing CAS on any record, a process first announces a HP to ensure that another process does not reclaim the record (and possibly does the same for the old value and new value for the CAS), and then checks whether its warning bit is set. If so, it releases its HP and jumps back to a safe checkpoint. Whenever p retires a record, it places it in a local buffer. If this local buffer becomes sufficiently large, then p performs reclamation. To perform reclamation, p simply sets the warning bits of all processes, and then frees every record in its local buffer that is not pointed to by a HP.

Before one can use OA to reclaim memory for a lock-free algorithm, one must first transform the algorithm into normalized lock-free form, then replace each shared memory read or CAS with a small procedure. Each shared memory read is replaced with: two reads, a branch, a possible write, and a possible jump. Each shared memory write or CAS is replaced with: four to seven writes, one read, one branch, a CAS, a memory fence, three possible bit-masking operations, and a possible jump.

Conceptually, OA is somewhat simpler than DEBRA+. However, in practice, OA would likely be less efficient, and more time consuming to apply to complex data structures, since it requires code modifications for each read, write and CAS on shared memory, and for each record retired. It is also quite likely to be less efficient, for this reason. In a subsequent paper, Cohen and Petrank presented an extended version of OA called automatic optimistic access (AOA). AOA uses garbage collection techniques to eliminate the need for a procedure to be invoked whenever a record is retired [35].

Applying each technique. The effort required to apply these techniques varies widely. See Figure 11.2 for a summary.

Necessary code modifications	RC	HP	B&C	TS	ST	EBR	DTA	QS	OA	DEBRA	DEBRA+
per accessed record	✓	✓	✓	✓ ✓ ✓ b, c	✓	✓	✓	✓	✓	✓	✓
per operation		✓	✓		✓	✓	✓	✓	✓	✓	✓
per retired record			✓		✓	✓	✓	✓	✓	✓	✓
other	a	b	a		d	b	e				f
Special timing assumptions				For progress				For correctness			
Fault tolerant	✓	✓	✓		✓		✓	✓	✓		✓
Termination of memory reclamation procedures	L	W	L	Blocking	L	L/W	L	$L_{rooster}$	L	W	W_{sig}
Can traverse pointer from retired record to retired record	✓		✓			✓	✓		✓	✓	✓

Figure 11.2: Summary of reclamation schemes. **Other code modifications:** (a) break cycles in pointers; (b) write recovery code for when a process fails to acquire a HP; (c) insert transaction *checkpoints* after every few lines of code; (d) integrate crash recovery with the lock-free data structure’s synchronization mechanisms (and only works for lists); (e) transform lock-free algorithm into normalized form, and then instrument every read, write and CAS on shared memory; (f) write crash recovery code (which is trivial for many data structures). **Termination of memory reclamation procedures:** (L) lock-free; ($L_{rooster}$) lock-free if rooster processes cannot crash; (W) wait-free; (W_{sig}) wait-free if the operating system’s signaling mechanism is wait-free

11.2 DEBRA: Distributed Epoch Based Reclamation

In this section, we present DEBRA, a distributed version of EBR with numerous advantages over classical EBR.

First, DEBRA supports a sort of *partial fault tolerance*. Consider an execution of a data structure that uses EBR. Observe that a process that sleeps for a long time will delay reclamation for all other processes, *even if it is not currently executing an operation on the data structure*. In DEBRA, a process can prevent other operations from reclaiming memory **only** if it is currently executing an operation on the data structure. Thus, if a process sleeps for a long time or crashes while it is not executing an operation on the data structure, other processes will continue to reclaim memory as usual. This can have a significant impact in real applications, where operations on a data structure may represent a small part of the execution. It also makes it possible to terminate some of the processes operating on a data structure, or reassign them to different tasks, without permanently halting reclamation for all other processes.

Second, recall that in EBR, each time a process begins a new operation, it reads the epoch announcements of all processes. This can be expensive, especially on NUMA systems, where reading the epoch announcements of processes on other sockets is likely to incur extremely expensive last-level cache misses. In DEBRA, we read the epoch announcements of all processes *incrementally* over many operations. This can slightly delay the reclamation of some records, but it dramatically reduces the overall cost of reading epoch announcements.

Third, instead of having all processes synchronize on shared epoch bags, each process has its own local epoch bags, and reclamation proceeds independently for each process. Additionally, these bags are carefully optimized for good cache performance and extremely low overhead.

Using DEBRA DEBRA provides four operations: *leaveQstate()*, *enterQstate()*, *retire(r)* and *isQuiescent()*, where r is a record. Each of these operations takes $O(1)$ steps in the worst-case. Let T be a lock-free data structure. To use DEBRA, T simply invokes *leaveQstate* at the beginning of each operation, *enterQstate* at the end of each operation, and *retire(r)* each time a record r is retired (i.e., removed from T). Like EBR, DEBRA assumes that a process does not hold a pointer to any record between successive operations on T . Each process alternates invocations of *leaveQstate* and *enterQstate*, beginning with an invocation of *leaveQstate*. Each process is said to be *quiescent* initially and after

(Applying DEBRA)

```

1 Value search(Key key) {
2 +   leaveQstate();
3
4   Node *node = root;
5
6   while (!node.isLeaf()) {
7
8     if (key < node->key) {
9       node = node->left;
10    } else {
11      node = node->right;
12    }
13
14
15
16
17
18
19   }
20   if (key == node->key) {
21     Value result = node->value;
22 +   enterQstate();
23     return result;
24   }
25 +   enterQstate();
26   return NO_VALUE;
27 }
```

(Applying hazard pointers)

```

1 Value search(Key key) {
2 +   announce_root
3 +   if (root is retired) restart search
4   Node *node = root;
5   Node *prev = NULL;
6   while (!node.isLeaf()) {
7 +     prev = node;
8     if (key < node->key) {
9       node = node->left;
10    } else {
11      node = node->right;
12    }
13 +     announce_node
14 +     if (node is retired) {
15 +       release hazard pointer to prev
16 +       restart search
17 +     }
18 +     release hazard pointer to prev
19   }
20   if (key == node->key) {
21     Value result = node->value;
22 +     release hazard pointer to node
23     return result;
24   }
25 +     release hazard pointer to node
26   return NO_VALUE;
27 }
```

Figure 11.3: Applying DEBRA and HPs to a search in a binary search tree. (+) denotes a new line.

invoking *enterQstate*, and is said to be *non-quiescent* after invoking *leaveQstate*. An invocation of *isQuiescent* by a process returns true if it is quiescent, and false otherwise. Figure 11.3 is an example of DEBRA applied to code for searching a lock-free binary search tree. For comparison, it also shows how HPs could be applied *if* one could determine whether a node is retired. Note that determining whether node is *retired* (and not just *marked*) can be extremely difficult, as discussed in Section 11.1 (where we discussed problems with HPs).

Implementation C++ style pseudocode for DEBRA appears in Figure 11.4. Each process *p* has three limbo bags, denoted *bag₀*, *bag₁* and *bag₂*, which contain records that it removed from the data structure. At any point, one of these bags is designated as *p*'s limbo bag for the current epoch, and is pointed to by a local variable *currentBag*. Whenever *p* removes a record from the data structure, it simply adds it to *currentBag*. Each process has a *quiescent bit*, which indicates whether the process is currently quiescent. The only thing *p* does when it enters a quiescent state is set its quiescent bit. Whenever *p* leaves a quiescent state, it reads the current epoch *e* and announces it in *announce_p*. If this changes the value of *announce_p*, then the contents of the oldest limbo bag can be reused or freed. In this case, *p* changes *currentBag* to point to the oldest limbo bag, and then moves the contents of *currentBag* to an object pool. Next, *p* attempts to determine whether the epoch can be advanced, which is the case if each process is either quiescent or has announced *e*. To do this efficiently, *p* checks the announcements and quiescent bits of all processes *incrementally*, reading one announcement and one quiescent bit in each *leaveQstate* operation. Process *p* repeatedly checks the announcement and quiescent bit of the same process *q* in each of its *leaveQstate* operations, until *q* either announces the current epoch or becomes quiescent, or until the epoch changes. A local variable *checkNext* keeps track of the next process whose announcement should be checked. Once *checkNext* is *n*, *p* performs a CAS to increment the current epoch.

```

1 process local variables:
2     long pid;                                // process id
3     long checkNext;                          // the next process whose announcement should be checked
4     blockbag *bags[0..2];                    // limbo bags for the last three epochs
5     blockbag *currentBag;                  // pointer to the limbo bag for the current epoch
6     long index;                            // index of currentBag in bags[0..2]
7     long opsSinceCheck;                // # ops performed since checking another process' announcement
8 shared variables:
9     long epoch;                                // current epoch
10    long announce[n];                      // per-process announced epoch and quiescent bit
11    objectpool *pool;                     // pointer to object pool
12
13 bool getQuiescentBit(long otherPid) { return announce[otherPid] & 1; }
14 void setQuiescentBitTrue(long otherPid) { announce[otherPid] = announce[otherPid] | 1; }
15 void setQuiescentBitFalse(long otherPid) { announce[otherPid] = announce[otherPid] & ~1; }
16 bool isEqual(long readEpoch, long announcement) {
17     return readEpoch == (announcement & ~1); // compare read epoch to epoch-bits from announcement
18 }
19
20 void retire(record *p) { currentBag->add(p); }
21 bool isQuiescent() { return getQuiescentBit(pid); }
22 void enterQstate() { setQuiescentBitTrue(pid); }
23 bool leaveQstate() {
24     bool result = false;
25     long readEpoch = epoch;
26     if (!isEqual(readEpoch, announce[pid])) { // our announcement differs from the current epoch
27         opsSinceCheck = checkNext = 0;        // we are now scanning announcements for a new epoch
28         rotateAndReclaim();
29         result = true;                      // result indicates that we changed our announcement
30     }
31     // incrementally scan all announcements
32     if (++opsSinceCheck >= CHECK_THRESH) {
33         opsSinceCheck = 0;
34         long other = checkNext % n;
35         if (isEqual(readEpoch, announce[other]) || getQuiescentBit(other)) {
36             long c = ++checkNext;
37             if (c >= n && c >= INCR_THRESH) { // if we scanned every announcement
38                 CAS(&epoch, readEpoch, readEpoch+2);
39             }
40             announce[pid] = readEpoch;          // announce new epoch with quiescent bit = false
41         }
42     }
43     void rotateAndReclaim() { // rotate limbo bags and reclaim records retired two epochs ago
44         index = (index+1) % 3;           // compute index of oldest limbo bag
45         currentBag = bags[index];      // reuse the oldest libmo bag as the new currentBag
46         pool->moveFullBlocks(currentBag); // move all full blocks to the pool
47     }
}

```

Figure 11.4: C++ style pseudocode for DEBRA, where n is the number of processes, INCR_THRESH is the minimum number of times a process must invoke *leaveQstate* before it can increment the epoch, and CHECK_THRESH is the number of times a process must invoke *leaveQstate* before it will check the epoch announcement of another process.

Correctness DEBRA reclaims a record only when no process has a pointer to it: Suppose p places a record r in limbo bag b at time t_1 , and moves r from b to the pool at time t_2 . Assume, to obtain a contradiction, that a process q has a pointer to r at time t_2 . At time t_1 , b is p 's current limbo bag, and just before time t_2 , b is changed from being p 's oldest limbo bag to being p 's current limbo bag, again. Thus, $currentBag$ must be changed at least three times between t_1 and t_2 . Since p changes $currentBag$ only in an invocation of $leaveQstate$ that changes $announce_p$, p must perform at least three such invocations between t_1 and t_2 . The current epoch must change between any pair of invocations of $leaveQstate$ that change $announce_p$, so the current epoch must change at least twice between t_1 and t_2 . Consider two consecutive changes of the current epoch, from e to e' and from e' to e'' . At some point between these two changes, q must either be quiescent or have announced e' . Process q must announce e' after reading the current epoch and seeing e' , before the current epoch changes from e' to e'' . Thus, q must announce e' after the current epoch changes from e to e' , and before it changes from e' to e'' . Since q can only announce an epoch when it is in a quiescent state, q must therefore be quiescent at some point between t_1 and t_2 . This means q must obtain its pointer to r by following pointers from an entry point after it was quiescent, which is after t_1 . However, r is removed from the data structure before t_1 , and, hence, it is no longer reachable by following pointers from an entry point. This is a contradiction.

Object pool The object pool shared by all processes is implemented as a collection of n *pool bags*, one per process, and one shared bag. Whenever a process moves a record to the pool, it places the record in its pool bag. If its pool bag is too large, it moves some records to the shared bag. Whenever a process wants to allocate a record, it first attempts to remove one from its pool bag. If its pool bag is empty, it attempts to take some records from the shared bag. If the process fails to take any record from the shared bag, then it will allocate some new records and place them in its pool bag.

Block bags For efficiency, each pool bag and limbo bag is implemented as a *blockbag*, which is a singly-linked list of *blocks*. Each block contains a *next* pointer and up to B records. (In our experiments, $B = 256$.) The head block in a blockbag always contains fewer than B records, and every subsequent block contains exactly B records. With this invariant, it is straightforward to design constant time operations to add and remove records in a blockbag, and to move all full blocks from one blockbag to another. This allows a process to move all full blocks of records in its oldest limbo bag to the pool highly efficiently after announcing a new epoch. However, if the process only moves *full* blocks to the pool, then this limbo bag may be left with some records in its head block. These records *could* be moved to the pool immediately, but it is more efficient to leave them in the bag, and simply move them to the pool later, once the block that contains them is full. One consequence of not moving these records to the pool is that each limbo bag can contain at most $B - 1$ records that were retired two or more epochs ago. This does not affect correctness. The shared bag is implemented as a lock-free singly-linked list of blocks with operations to add and remove a full block. Moving entire blocks to and from the shared bag greatly reduces synchronization costs.

Operating on blocks instead of individual records significantly reduces overhead. However, it also requires a process to allocate and deallocate blocks. To reduce the number of blocks that are allocated and deallocated during an execution, each process has a bounded *block pool* that is used by all of its local blockbags. Instead of deallocating a block, a process returns the block to its block pool. If the block pool is already full, then the block is freed. Experiments show that allowing each process to keep up to 16 blocks in its block pool reduces the number of blocks allocated by more than 99.9%. No blocks are allocated for the shared bag, since blocks are simply moved between pool bags and the shared bag.

Minor optimizations We make two additional optimizations in our implementation. First, the least significant bit of $announce_p$ is used as p 's quiescent bit. This allows both values to be read and written atomically, which reduces the number of reads and writes to shared memory. Second, a process attempts to increment the current epoch only after invoking $leaveQstate$ at least INCR_THRESH times, where INCR_THRESH is a constant (100 in our experiments). This is especially helpful when the number of processes is small. For example, in a single process system, without this optimization, the process will advance the epoch and try to move records to the pool at the beginning of every single operation, introducing unnecessary overhead.

Optimizing for NUMA systems Memory layout and access pattern can have a significant impact on the performance of an algorithm on a NUMA system. Recall that each invocation of $leaveQstate$ by a process q reads one announcement $announce_p$ (which is periodically changed by process p). If p is on the same socket as q (so p and q share the last-level cache), then this read will usually be quite fast, since a write by p to $announce_p$ will not invalidate q 's cached copy of $announce_p$. However, if they are on different sockets, then a write by p to $announce_p$ will invalidate q 's cached copy, and will cause a subsequent read of $announce_p$ by q to incur a last-level cache miss. The cost of these last-level cache misses is not noticeable in all workloads, but we *did* notice it in experimental workloads where all processes performed short read-only operations on small data structures that fit in the last-level cache (so the only last-level cache misses were caused by these reads). To reduce the impact of these cache misses, we suggest reading an announcement only once every CHECK_THRESH invocations of $leaveQstate$ (where CHECK_THRESH is a small constant). Of course, checking announcements less frequently can delay reclamation for some records.

11.3 Adding fault tolerance

The primary disadvantage of DEBRA is that a crashed or slow process can stay in a non-quiescent state for an arbitrarily long time. This prevents any other processes from freeing memory. Although we did not observe this pathological behaviour in our experiments, many applications require a bound on the number of records waiting to be freed.

Lock-free data structures are ideal for building fault tolerant systems, because they are designed to be provably fault tolerant. If a process crashes while it is in the middle of any lock-free operation, and it leaves the data structure in an inconsistent state, other processes can always repair that state. The onus is on a process that wants to access part of a data structure to restore that part of the data structure to a consistent state before using it. Consequently, a lock-free data structure always provides procedures to repair and access parts of the data structure that are damaged (by a process crash) or undergoing changes. (Furthermore, these procedures are necessarily designed so that processes can crash while executing them, and other processes can still repair the data structure and continue to make progress.) We use these procedures to design a mechanism that allows a process to *neutralize* another process that is preventing it from advancing the epoch.

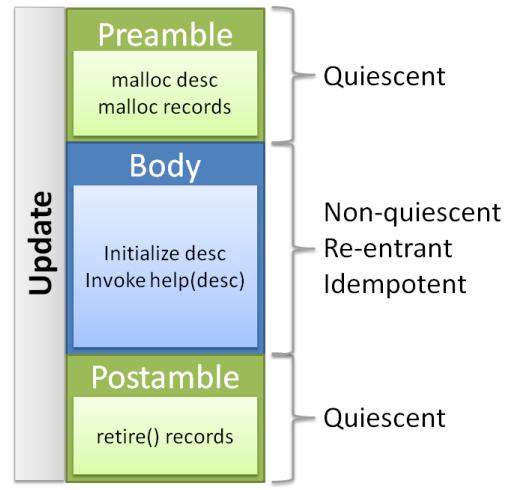
A novel aspect of DEBRA+ is our use of two features offered by Unix, Linux and other POSIX-compliant operating systems. The first is *signaling*, an inter-process communication mechanism. Signals can be sent to a process by the operating system, and by other processes. When a process receives a signal, the code it was executing is interrupted, and the process begins executing a *signal handler*, instead. When the process returns from the signal handler, it resumes executing from where it was interrupted. A process can specify what action it will take when it receives a particular signal by registering a function as its signal handler.

The second feature is *non-local goto*, which allows a process to begin executing from a different instruction, *outside* of the current function, by using two procedures: *sigsetjmp* and *siglongjmp*. A process first invokes *sigsetjmp*, which saves its local state and returns false. Later, the process can invoke *siglongjmp*, which restores the state saved by *sigsetjmp* immediately prior to its return, but causes it to return true instead of false. The standard way to use these primitives is with the idiom: “if (*sigsetjmp*(...)) alternate(); else usual();”. A process that executes this code will save its state and execute *usual()*. Then, if the process later invokes *siglongjmp*, it will restore the state saved by *sigsetjmp* and immediately begin executing *alternate()*.

At the beginning of each operation by a process q , q invokes *sigsetjmp*, following the idiom described above, and then proceeds with its operation. Another process p can interrupt q by sending a signal to q . We design q 's signal handler so that, if q was interrupted while it was in a quiescent state, then q will simply return from the signal handler and resume its regular execution from wherever it was interrupted. However, if q was interrupted in a non-quiescent state, then it is neutralized: it will enter a quiescent state and perform a *siglongjmp*. Then, q will execute special *recovery code*, which allows it to clean up any mess it left because it was neutralized. Since q 's signal handler performs *siglongjmp* only if q was interrupted in a non-quiescent state, q will not perform *siglongjmp* while it is executing recovery code. Hence, if q receives a signal while it is executing recovery code, it will simply return from the signal handler and resume executing wherever it left off.

Our technique requires the operating system to guarantee that, after a process p sends q a signal, the next time process q takes a step, it will execute its signal handler. (This requirement is satisfied by the Linux kernel [76]. It can also be weakened with small modifications to DEBRA+, which are discussed at the end of this section.) With this guarantee, after p has sent a signal to q , it knows that q will not access any retired record until q has executed its recovery code and subsequently executed *leaveQstate*. Thus, as soon as p has sent q a signal, p can immediately proceed as if q is quiescent.

Operations for which recovery is simple The main difficulty is using DEBRA+ is designing recovery code. Although recovery must be tailored to the data structure, it is straightforward for lock-free operations of the following form. Each operation is divided into three parts: a quiescent bookkeeping *preamble*, a non-quiescent *body*, and a quiescent bookkeeping *postamble*. Processes can be neutralized while executing in the body, but cannot be neutralized while executing in the preamble or postamble (because a process will not call *siglongjmp* while it is quiescent). Consequently, processes should not do anything in the body that will corrupt the data structure if they are neutralized part way through. Allocation, deallocation, manipulation of process-local data structures that persist between operations, and other non-reentrant actions should occur only in the preamble and postamble.



Applying DEBRA+ Figure 11.5 shows how to apply DEBRA+ to the type of operations described above. The remainder of this section explains the steps shown there. Consider an operation O . In the quiescent preamble, O allocates a special record called a *descriptor*. A pointer to this descriptor is stored in a process local variable called

```

1 || process local variables:
2 ||   descriptor *desc;
3 || void signalHandler(args):
4 ||   if (isQuiescent()) then
5 ||     enterQstate();
6 ||     siglongjmp(...);           // jump to recovery code
7 || int doOperationXYZ(args):
8 ||   ...                         // quiescent preamble
9 ||   while (!done)
10 ||     if (sigsetjmp(...))        // begin recovery code
11 ||       if (isRProtected(desc)) done = help(desc);
12 ||       RUnprotectAll();
13 ||     else                      // end recovery code
14 ||       leaveQstate();          // begin body
15 ||       do search phase
16 ||       initialize *desc
17 ||       RProtect each record that will be accessed, or used as the old value of a CAS, by help(desc)
18 ||       RProtect(desc);
19 ||       done = help(desc);
20 ||       enterQstate();          // end body
21 ||       RUnprotectAll();
22 ||     ...                         // quiescent postamble
23 ||   perform retire() calls

```

Figure 11.5: Applying DEBRA+ to a typical lock-free operation. Lines 14 and 20 are necessary for both DEBRA and DEBRA+. Gray lines are necessary for fault tolerance (DEBRA+).

desc. Any other records that might be needed by the body are also allocated in the preamble.

The body first reads some records, and then initializes the descriptor *desc*. Intuitively, this descriptor contains a description of all the steps the operation *O* will perform. The body then executes a *help* procedure, which uses the information in the descriptor to perform the operation. We assume that the descriptor includes all pointers that the *help* procedure will follow (or use as the expected value of a CAS). The *help* procedure can also be used by other processes to help the operation complete. In a system where processes may crash, a process whose progress is blocked by another operation cannot simply wait for the operation to complete, so helping is necessary. The *help* procedure for any lock-free algorithm is typically reentrant and idempotent, because, at any time, one process can pause, and another process can begin helping. The end of the body is marked with an invocation of *enterQstate* (which is, itself, reentrant and idempotent). The quiescent postamble then invokes *retire* for each record that was removed from the data structure by the operation.

Recovery We now describe recovery for an operation *O* performed by a process *p*. Suppose *p* receives a signal, enters a quiescent state, and then performs *siglongjmp* to begin executing its recovery code. Although there are many places where *p* might have been executing when it was neutralized, it is fairly simple to determine what action it should take next. The main challenge is determining whether another process already performed *O* on *p*'s behalf. To do this, *p* checks whether it announced a descriptor for *O* before it was neutralized. If it did, then some other process might have seen this descriptor and started helping *O*. So, *p* invokes *help* (which is safe even if another process already helped *O*, since *help* is idempotent). Otherwise, *p* can simply restart the body of *O*.

DEBRA allows a non-quiescent process executing an operation to safely access any record that it reached by following pointers from an entry point during the operation. However, DEBRA does *not* allow quiescent processes to safely access any records. In DEBRA+, once a process *p* has been sent a signal, other processes treat *p* as if were quiescent. Furthermore, *p* enters a quiescent state before executing recovery code, and it remains in a quiescent state throughout the recovery code. However, the help procedure in *p*'s recovery code must access the descriptor record, and possibly some of the records to which it points. Thus, we need an additional mechanism in DEBRA+ to allow *p*

to access this limited set of records even though it is quiescent. We use HPs in a very limited way to prevent these records from being freed by other processes before p has finished its recovery code. This lets p safely run its recovery code in a quiescent state, so that other processes can continue to advance the current epoch and reclaim memory.

We now describe how HPs are used. Let S be the set of records that will be accessed, or used as the old value of a CAS, by $help(desc)$. In the body of an operation by p , p announces HPs to all records in S by invoking $RProtect(r)$ for each $r \in S$. Then, p invokes $RProtect(desc)$ to announce a HP to the descriptor, and invokes $help(desc)$. After performing $help(desc)$ and invoking $enterQstate$, p invokes $RUnprotectAll$ to release all of its HPs. Note that, since $RProtect$ is performed in the body, while p is non-quiescent, p might be neutralized while executing $RProtect$. Hence, $RProtect$ must be reentrant and idempotent.

When executing recovery code, p first invokes $isRProtected(desc)$ (which returns true if some HP points to $desc$ and false otherwise) to determine whether it announced a HP to $desc$. Suppose it did. Since p announces a HP to the descriptor d *after* announcing HPs to all records in S , when p performs recovery, if it sees that it announced a HP to d , then it knows it already announced HPs to all records in S . Thus, its HPs will prevent everything it will access during its recovery from being reclaimed until it has finished using them. So, p can safely execute $help(desc)$. Now, suppose p did not announce a HP to $desc$. Since p announces a HP to $desc$ *before* invoking $help(desc)$, this means p has not yet invoked $help(desc)$, so no other process is aware of p 's operation. Therefore, p can simply terminate its recovery code and restart its operation. At the end of the recovery code, p invokes $RUnprotectAll$ to release all of its HPs.

Recall that, in DEBRA, whenever a process p announces a new epoch, it can immediately reclaim all records in its oldest limbo bag and move them to its pool. In DEBRA+, some of the records in p 's oldest limbo bag might be pointed to by HPs, so p cannot simply move all of these to its pool. Before p can move a record r to the pool, it must first verify that no HP points to it. We discuss how this can be done efficiently, below.

Complexity Thanks to our new *neutralizing* mechanism, we can bound the number of records waiting to be freed. Each time a process p performing $leaveQstate$ encounters a process q that is not quiescent and has not announced epoch e , p checks whether the size of its own current limbo bag exceeds some constant c . If so, p neutralizes q . After p 's current limbo bag contains at least c elements, and p performs n more data structure operations, it will have performed $leaveQstate$ n times, and each non-quiescent process will either have announced the current epoch or been neutralized by p . Consequently, p will advance the current epoch, and, the next time it performs $leaveQstate$, it will announce the new epoch and reclaim records. It follows that p 's current limbo bag can contain at most $c + O(nm)$ elements, where m is the largest number of records that can be removed from the data structure by a high-level operation. Therefore, the total number of records waiting to be freed is $O(n(nm + c))$.

In DEBRA, all full blocks in a limbo bag are moved to the pool in constant time. In DEBRA+, records can be moved to the pool only if no HP points to them, so this is no longer possible. One way to move records from a limbo bag b to the pool is to iterate over each record r in b , and check if a HP points to r . To make this more efficient, we move records from b to the pool only when b contains $nk + \Omega(nk) + B$ records, where k is the number of HPs needed per process, and B is the maximum number of records in a block. Before we begin iterating over records in b , we create a hash table containing every HP. Then, we can check whether a HP points to r in $O(1)$ expected time. We can further optimize by rearranging records in b so that we can still move full blocks to the pool, instead of individual records. To do this, we iterate over the records in b , and move the ones pointed to by HPs to the beginning of the blockbag. All full blocks that do not contain a record pointed to by a HP are then moved in $O(1)$ time. Since there are at most nk HPs, and we scan only when b contains at least $nk + B + \Omega(nk)$ records, we will be able to move at least

$\max\{B, \Omega(nk)\}$ records to the pool. Thus, the expected amortized cost to move a record to the pool (or free it) is $O(1)$.

Implementation C++ style pseudocode for DEBRA+ appears in Figure 11.6. There, only the procedures that are different from DEBRA are shown. There are three main differences from DEBRA. First, in an invocation of *leaveQstate* by process p , if p encounters a process q that has not announced the current epoch, and is not quiescent, p invokes a procedure called *suspectNeutralized*. This procedure checks whether p 's current limbo bag contains more than a certain number of records, and, if so, neutralizes q . Recall that p neutralizes q by sending a signal to q . The signal handler that is executed by q when it receives a signal is called *signalhandler*. Second, a limited version of HPs is provided by procedures *isRProtected*, *RProtect* and *RUnprotectAll*. Third, the procedure *rotateAndReclaim* implements the algorithm described above efficiently moving records from a limbo bag to a pool (only if the records are not pointed to by HPs) in expected amortized $O(1)$ steps per record.

One problem with DEBRA+ is that it seems difficult to apply to lock-based data structures, because it is dangerous to interrupt a process and force it to restart while it holds a lock. Of course, there is little reason to use DEBRA+ for a lock-based data structure, since locks can cause deadlock if processes crash. For lock-based data structures, DEBRA can be used, instead.

Alternative implementation options Above, we specified the guarantee that the operating system signaling mechanism must provide: after a process p sends process q a signal, the next time q takes a step, it will execute its signal handler. It is possible to modify DEBRA+ to work with a weaker guarantee. Suppose the operating system instead guaranteed that, after a process p sends process q a signal, q is guaranteed to begin executing its signal handler when it next experiences a context switch. Then, after p sends a signal to q , it can either wait or defer reclamation until q is next context switched or is not running. For most operating system schedulers, every process is context switched out after a bounded length of time (a scheduling quantum). Many operating systems also provide mechanisms to determine whether a process is currently running, how many context switches it has undergone, how much time remains until it will be context switched, and so on. For example, Linux provides access to this information through a virtual file system (rooted at "/proc").

11.4 A lock-free memory management abstraction

There are many compelling reasons to separate memory allocation and reclamation from data structure code. Although the steps that a programmer must take to apply a non-automatic technique to a data structure vary widely, it is possible to find a small, natural set of operations that allows a programmer to write data structure code once, and easily plug in many popular memory reclamation schemes. In this section, we describe a record management abstraction, called a Record Manager, that is easy to use, and provides sufficient flexibility to support: HPs (all versions), B&C, TS, EBR, DTA, QS, DEBRA and DEBRA+. (ST is not supported because it requires a programmer to insert transactions throughout the code, and to annotate the beginning and end of each stack frame. OA is not supported because it requires each read and CAS to be instrumented.)

A Record Manager has three components: an Allocator, a Reclaimer and a Pool. The Allocator determines how records will be allocated (e.g., by individual calls to *malloc* or by handing out records from a large range of memory) and freed. The Reclaimer is given records after they are removed from the data structure, and determines when they

```

1 process local variables:
2     hashtable scanning;           // hash table used to collect all RProtected records
3 shared variables:
4     arraystack RProtected[n];    // array of RProtected record* for each process
5
6 bool isRProtected(record *r) { return RProtected[pid].contains(r); }
7 bool RProtect(record *r) { RProtected[pid].add(r); } // O(1) time
8 void RUnprotectAll() { RProtected[pid].clear(); } // O(1) time
9 bool leaveQstate() {
10     bool result = false;
11     long readEpoch = epoch;
12     if (!isEqual(readEpoch, announce[other])) { // our announcement differs from the current epoch
13         checkNext = 0;                         // we are now scanning announcements for a new epoch
14         rotateAndReclaim();
15         result = true;                      // result indicates that we changed our announcement
16     }
17     // incrementally scan all announcements
18     if (++opsSinceCheck >= CHECK_THRESH) {
19         opsSinceCheck = 0;
20         long other = checkNext % n;
21         if (isEqual(readEpoch, announce[other]) || isQuiescent(other) || suspectNeutralized(other)) {
22             long c = ++checkNext;
23             if (c >= n && c >= INCR_THRESH) { // if we have scanned every announcement
24                 CAS(&epoch, readEpoch, readEpoch+1);
25             }
26             announce[pid] = readEpoch;          // announce new epoch with quiescent bit = false
27             return result;
28         }
29     void rotateAndReclaim() {
30         index = (current+1) % 3;           // compute index of oldest limbo bag
31         currentBag = bags[index];        // reuse the oldest limbo bag as the new currentBag
32         // if currentBag contains sufficiently many records to get amortized O(1) time per record
33         if (currentBag->getBlockSize() >= scanThreshold) {
34             // hash all announcements
35             scanning.clear();
36             for (int other=0; other < n; ++other) {
37                 int sz = RProtected[other].size();
38                 for (int i=0; i<sz; ++i) {
39                     record *hp = RProtected[other].get(i);
40                     if (hp != NULL) {
41                         scanning.insert(hp);
42                     }
43                 }
44                 // if any records in currentBag are RProtected, swap them to the front
45                 blockbag_iterator it1 = currentBag->begin();
46                 blockbag_iterator it2 = currentBag->begin();
47                 while (it1 != currentBag->end()) {
48                     if (scanning.contains(*it1)) { // record pointed to by it1 is RProtected
49                         swap(it1, it2);           // swap records pointed to by it1 and it2
50                         it2++;                  // advance iterator it2
51                     }
52                     it1++;                  // advance iterator it1
53                 }
54                 // now, every record after it2 can be freed, so we reclaim all full blocks after it2
55                 pool->moveFullBlocks(it2); // O(1) time
56             }
57             bool suspectNeutralized(long other) {
58                 return (currentBag->getBlockSize() >= SUSPECT_THRESHOLD_IN_BLOCKS)
59                 && (!pthread_kill(getPthreadID(other), SIGQUIT)); // successfully send signal to other
}
}

1 void signalhandler(int signum, siginfo_t *info, void *uctx) {
2     // if the process is not in a quiescent state, it jumps to a different instruction
3     // and cleans up after itself, instead of continuing its current operation.
4     if (!isQuiescent()) {
5         enterQstate();
6         siglongjmp(...);
7     } // otherwise, the process simply continues its operation as if nothing had happened.

```

Figure 11.6: C++ style pseudocode for data and procedures **added to DEBRA** to obtain DEBRA+, where n is the number of processes, INCR_THRESH is the minimum number of times a process must invoke *leaveQstate* before it can increment the epoch, and CHECK_THRESH is the number of times a process must invoke *leaveQstate* before it will check the epoch announcement of another process.

can be safely handed off to the Pool. The Pool determines when records are handed to the Allocator to be freed, and whether a process actually uses the Allocator to allocate a new record.

We implement data structures, Allocators, Reclaimers and Pools in a modular way, so that they can be combined easily. This clean separation into interchangeable components allows, e.g., the same Pool implementation to be used with both a HP Reclaimer and a DEBRA Reclaimer. Modularity is typically achieved with inheritance, but inheritance introduces significant runtime overhead. For example, when the precompiled data structure invokes *retire*, it does not know which of the precompiled versions of *retire* it should use, so it must perform work at runtime to choose the correct implementation. In C++, this can be done more efficiently with *template parameters*, which allow a compiler to reach into code and replace placeholder calls with calls to the correct implementations. Unlike inheritance, templates introduce no overhead, since the correct implementation is compiled into the code. Furthermore, if the correct implementation is a small function, the compiler can simply insert its code directly into the calling function (eliminating the function call altogether).

A programmer interacts with the Record Manager, which exposes the operations of the Pool and Reclaimer. A Pool provides *allocate* and *deallocate* operations. A Reclaimer provides operations for the basic events that memory reclamation schemes are interested in: starting and finishing data structure operations (*leaveQstate* and *enterQstate*), reaching a new pointer and disposing of it (*protect* and *unprotect*), and retiring a record (*retire*). It also provides operations to check whether a process is quiescent (*isQuiescent*) and whether a pointer can be followed safely (*isProtected*). Finally, it provides operations for making information available to recovery code (*RProtect*, *RUnprotectAll*, *isRProtected*).

Most of these are described in Section 11.2 and Section 11.3. We describe the rest here. *protect*, which must be invoked on a record *r* before accessing any field of *r*, returns true if the process successfully protects *r* (and, hence, is permitted to access its fields), and returns false otherwise. Once a process has successfully protected *r*, it remains protected until the process invokes *unprotect(r)* or becomes quiescent. *isProtected(r)* returns true if *r* is currently protected by the process.

Reclaimers for DEBRA and DEBRA+ are effectively described in Section 11.2 and Section 11.3. For these techniques, *unprotect* does nothing, and *protect* and *isProtected* simply return true. (Consequently, these calls are optimized out of the code by the compiler.) For HPs, *leaveQstate*, *RProtect* and *RUnprotectAll* all do nothing, and *isQuiescent* and *isRProtected* simply return false. *unprotect(r)* releases a HP to *r*, and *enterQstate* clears all announced HPs. *protect* announces a HP to a record and executes a function, which determines whether that record is in the data structure. *retire(r)* places *r* in a bag, and, if the bag contains sufficiently many records, it constructs a hash table *T* containing all HPs, and moves all records not in *T* to the Pool (as described in Section 11.1).

A predicate called *supportsCrashRecovery* is added to Reclaimers to allow a programmer to add crash recovery to a data structure without imposing overhead for Reclaimers that do not support crash recovery. For example, a programmer can check whether *supportsCrashRecovery* is true before invoking *RProtect*. The code statement “if

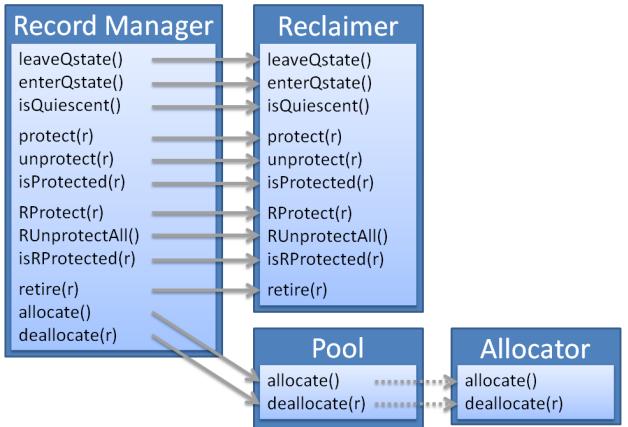


Figure 11.7: Operations provided by a Record Manager, Reclaimer, Pool and Allocator. Solid arrows (resp., dashed arrows) indicate that an operation on one object invokes (resp., may invoke) an operation on another object.

(*supportsCrashRecovery*)” statically evaluates to “if (true)” or “if (false)” at compile time, once the Reclaimer template has been filled in. Consequently, these if-statements are completely eliminated by the compiler. In our experiments, this predicate is used to invoke *sigsetjmp* only for DEBRA+ (eliminating overhead for the other techniques).

11.5 Experiments

Our primary experimental system was an Intel i7 4770 machine with 4 cores, 8 hardware threads and 16GB of memory, running Ubuntu 14.04 LTS. All code was compiled with GCC 4.9.1-3 and the highest optimization level (-O3). Google’s high performance Thread Caching malloc (tcmalloc-2.4) was used.

We ran experiments to compare the performance of various Reclaimers: DEBRA, DEBRA+, HP, ST and no reclamation (None). We used the Record Manager abstraction to perform allocation and reclamation for a lock-free balanced binary search tree (BST) [30]. Searches in this BST can traverse pointers from retired nodes to other retired nodes, so ST cannot be used, and we must confront the problems described in Section 11.1 to apply HP. Properly dealing with HP’s problems would be highly complex and inefficient, so we simply restart any operation that suspects a node is retired. Consequently, applying HP causes the BST to lose its lock-free progress guarantee. To determine whether this significantly affects the performance of HP, we added the same restarting behaviour to DEBRA, and observed that its impact on performance was small. (See Figure 11.13.) Note that the HP scheme was tuned for high performance (instead of space efficiency) by allowing processes to accumulate large buffers of retired nodes before attempting to reclaim memory.

Code for ST was graciously provided by its authors. They used a lock-based skip list to compare None, HP and ST. We modified their code to use a Record Manager for allocating and pooling nodes, and used it to compare None, DEBRA, HP and ST. The actual reclamation code for HP and ST is due to the authors of ST. Since the skip list uses locks, it cannot use DEBRA+.

Experiment 1 Our first experiment compared the overhead of performing reclamation for the various Reclaimers. In this experiment, each Reclaimer performed all the work necessary to reclaim nodes, but nodes were not actually reclaimed (and, hence, were not reused). The Record Manager used a *Bump Allocator*: each process requests a large region of memory from the operating system at the beginning of an execution, and then divides that region into nodes, which it allocates in sequence. Since nodes were not actually reclaimed, we eliminated the Pool component of the Record Manager. In this experiment, a data structure suffers the overhead of reclamation, but does *not* enjoy its benefits (namely, a smaller memory footprint and fewer cache misses).

For the balanced BST, we ran eight *trials* for each combination of Reclaimers in {None, DEBRA, DEBRA+, HP}, thread counts (in {1, 2, ..., 16}), operation mixes in {25i-25d, 50i-50d} (where xi-yd means x% insertions, y% deletions and (100 – x – y)% searches) and key ranges in {[0, 10000), [0, 1000000)}. For the skip list, the thread counts and operation mixes were the same, but ST was used instead of DEBRA+, and there was only one key range, [0, 200000). In each trial, the data structure was first prefilled to half of the key range, then the appropriate number of threads performed random operations (according to the operation mix) on uniformly random keys from the key range for two seconds. The average of each set of eight trials became a data point in a graph. (Unfortunately, the system quickly runs out of memory when nodes are not reclaimed, so it is not possible to run all trials for longer than two seconds. However, we ran long trials for many cases to verify that the results do not change.)

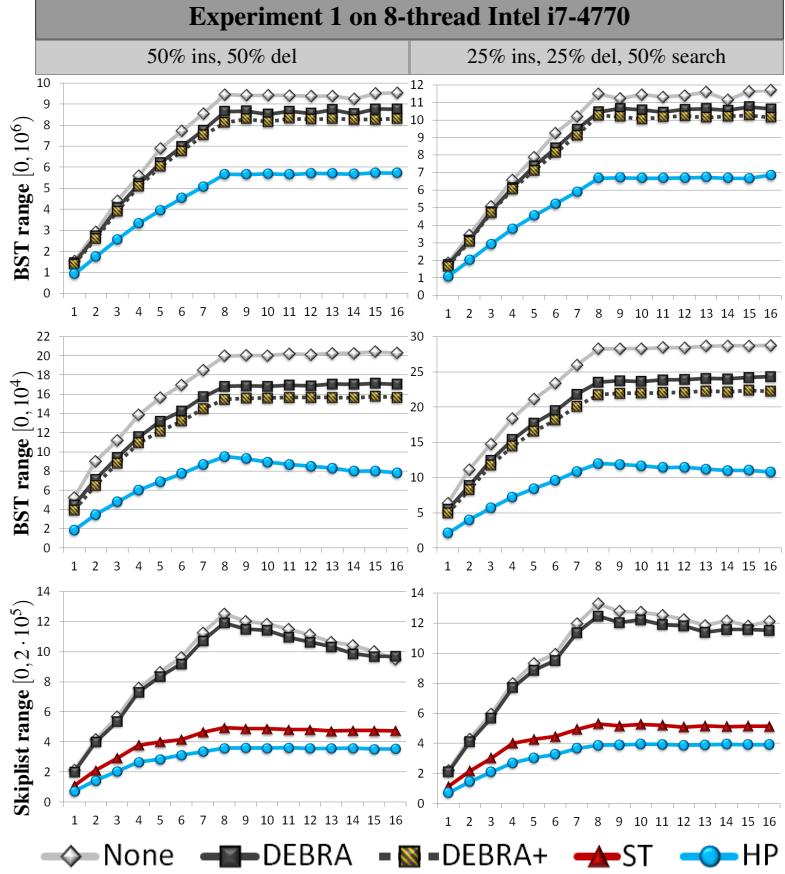


Figure 11.8: Results for Experiment 1 (Overhead of reclamation). The x-axis shows the number of processes. The y-axis shows throughput, in millions of operations per second.

The results in Figure 11.8 show that DEBRA and DEBRA+ have extremely low overhead. In the BST, DEBRA has between 5% and 22% overhead (averaging 12%), and DEBRA+ has between 7% and 28% overhead (averaging 17%). Compared to HP, on average, DEBRA performs 94% more operations and DEBRA+ performs 83% more. This is largely because DEBRA and DEBRA+ synchronize once *per operation*, whereas HP synchronizes each time a process reaches a new node. In the skip list, DEBRA has up to 6% overhead (averaging 4%), outperforms HP by an average of 200%, and also outperforms ST by between 93% and 168% (averaging 133%). ST has significant overhead. For instance, on average, it starts almost four transactions per operation (each of which announces one or more pointers), and runtime checks are frequently performed to determine if a new transaction should be started.

Experiment 2 In our second experiment, nodes were actually reclaimed. The Reclaimers were each paired with the same Pool as DEBRA. The only exception was None, which does not use a Pool. (Thus, None is the same as in the first experiment.)

The results appear in Figure 11.9. In the BST, DEBRA is only 8% slower than None on average, and, for some data points, DEBRA actually improves performance by up to 12%. This is possible because DEBRA reduces the memory footprint of the data structure, which allows a larger fraction of the allocated nodes to fit in cache and, hence, reduces

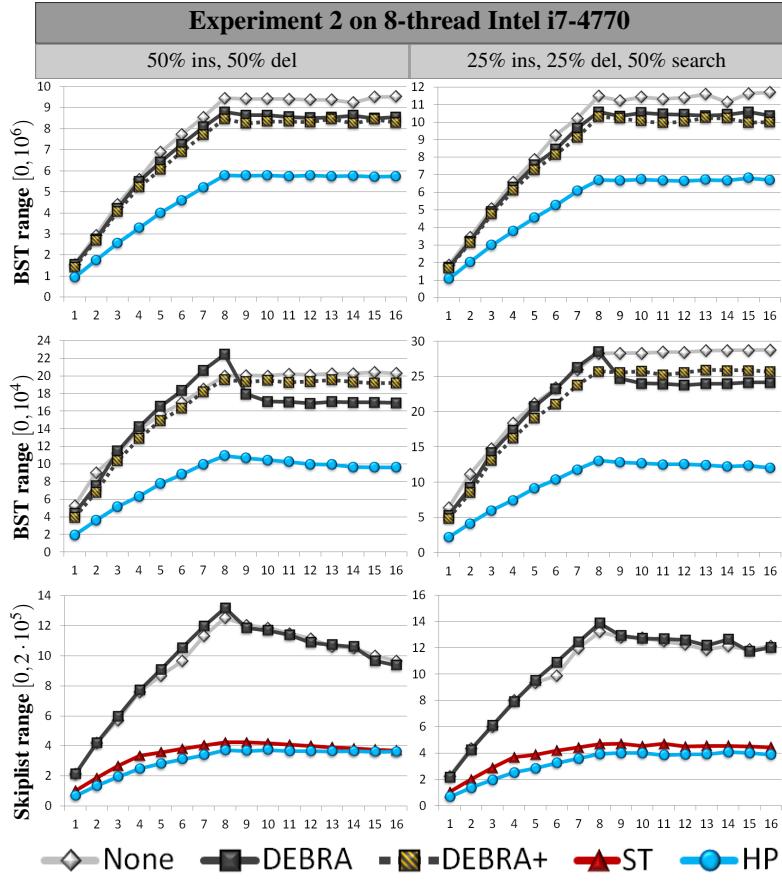


Figure 11.9: Experiment 2 (Using a Bump Allocator and a Pool).

the number of cache misses. DEBRA+ is between 2% and 25% slower than None, averaging 10%. Compared to HP, DEBRA is between 48% and 145% faster, averaging 80%, and DEBRA+ is between 43% and 123% faster, averaging 76%. In the skip list, DEBRA performs *as well as None*. DEBRA also *outperforms* ST by between 108% and 211%, averaging 160%.

To measure the benefit of neutralizing slow processes, we tracked the total amount of memory allocated for records in each trial. Since we used bump allocation, this simply required determining how far each bump allocator's pointer had moved during the execution. Thus, we were able to compute the total amount of memory allocated *after* each trial had finished (without having any impact on the trial while it was executing). Figure 11.11 shows the total amount of memory allocated for records in the second experiment in the BST with key range 10,000 and workload 50i-50d. (The other cases were similar.) DEBRA, DEBRA+ and HP all perform similarly up to eight processes. However, for more than eight processes, some processes are always context switched out, and they often prevent DEBRA from advancing the epoch in a timely manner. DEBRA+ fixes this issue. With 16 processes, DEBRA+ neutralizes processes an average of 935 times per trial, reducing memory usage by an average of 94% over DEBRA.

We also ran the second experiment on a NUMA Oracle T4-1 system with 8 cores and 64 hardware contexts. Figure 11.10 shows a representative sample of the results. Note that ST could not be run on this machine, since it does not support HTM.

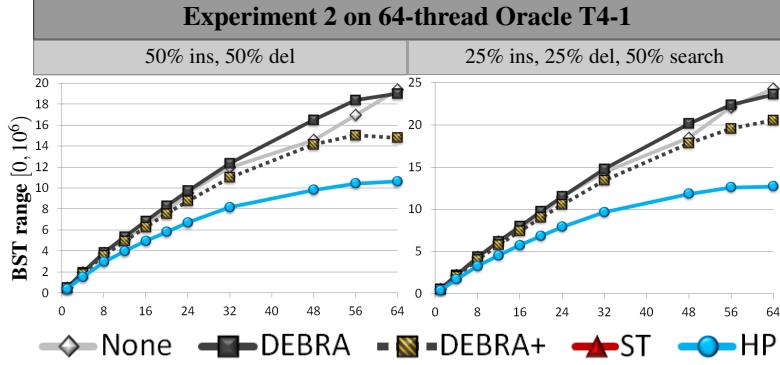


Figure 11.10: Extra results for Experiment 2 on Oracle T4-1. (ST could not be run, since the Oracle T4-1 does not support hardware transactional memory.)

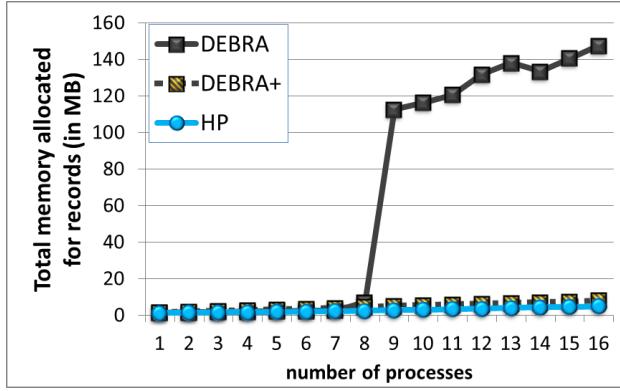


Figure 11.11: Memory allocated for records in Experiment 2 in the BST with keyrange 10,000 and workload 50i-50d.

Experiment 3 Our third experiment is like the second, except we used a different Allocator, which does not preallocate memory. The Allocator’s *allocate* operation simply invokes *malloc* to request memory from the operation system (and its *deallocate* operation invokes *free*).

The results (which appear in Figure 11.12) are similar to the results for the second experiment. However, the absolute throughput is significantly smaller than in the previous experiments, because of the overhead of invoking *malloc*. Although HP and ST are negatively affected, proportionally, they slow down less than None, DEBRA and DEBRA+. This illustrates an important experimental principle: *overhead should be minimized, because uniformly adding overhead to an experiment disproportionately impacts low-overhead algorithms, and obscures their advantage*.

11.6 Summary

In this work, we presented a distributed variant of EBR, called DEBRA. Compared to EBR, DEBRA significantly reduces synchronization overhead and offers high performance even with many more processes than physical cores. Our experiments show that, compared with performing no reclamation at all, DEBRA is 4% slower on average, 21% slower at worst, and up to 20% *faster* in some cases. Moreover, DEBRA outperforms StackTrack by an average of 138%. DEBRA is easy to use, and only adds O(1) steps per data structure operation and O(1) steps per retired record.

We also presented DEBRA+, the first epoch based reclamation scheme that allows processes to continue reclaiming

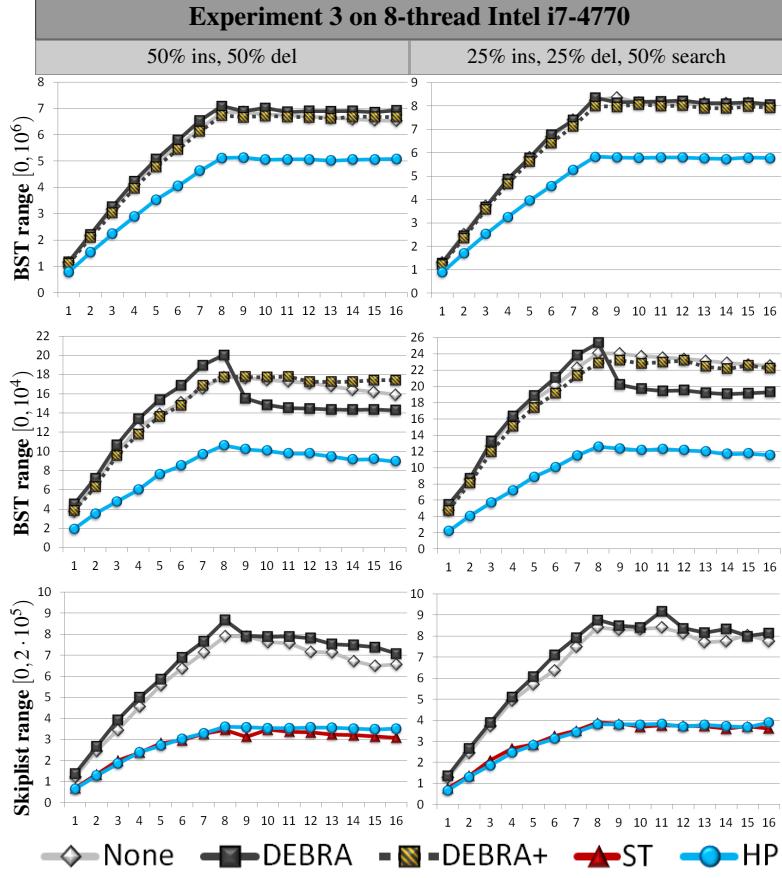


Figure 11.12: Experiment 3 (Using malloc and a Pool).

memory after a process has crashed. In an n process system, the number of objects waiting to be freed is $O(mn^2)$, where m is the largest number of objects retired by one data structure operation. The cost to reuse or free a record is $O(1)$ expected amortized time. In our experiments, DEBRA+ reduced memory consumption over DEBRA by 94%. Compared with performing no reclamation, DEBRA+ is only 10% slower on average. DEBRA+ also outperforms a highly efficient implementation of hazard pointers by an average of 70%.

We introduced the *Record Manager*, the first generalization of the C++ *Allocator* abstraction that is suitable for lock-free programming. A Record Manager separates memory reclamation code from lock-free data structure code, which allows a dynamic data structure to be implemented without knowing how its records will be allocated, reclaimed and freed. This abstraction adds virtually no overhead. It is highly flexible, allowing a programmer to interchange techniques for reclamation, object pooling, allocation and deallocation by changing one line of code.

Besides DEBRA and DEBRA+, the neutralizing technique introduced in this work is of independent interest. It would be useful to find different ways to neutralize processes, so, for example, the neutralizing technique could be used with different operating systems. There may also be opportunities to apply neutralizing in other contexts, such as garbage collection. Finally, it would also be interesting to understand whether these ideas can be extended to lock-based algorithms (even for a restricted class, such as reentrant and idempotent algorithms).



Figure 11.13: Overhead introduced by restarting BST operations instead of helping whenever a marked node is reached.

Chapter 12

Reusing descriptors

In simple lock-free data structures (e.g., [121, 61, 98, 101, 88]), a process can determine how to help an operation that blocks it by inspecting a small part of the data structure. In more complex lock-free data structures (such as [52, 70, 113, 30], and those implemented with *LLX* and *SCX*), processes publish *descriptors* for their operations, and helpers look at these descriptors to determine how to help. A descriptor typically encodes a sequence of steps that a process should follow in order to complete the operation that created it.

Since lock-free algorithms cannot use mutual exclusion, many helpers can simultaneously help an operation, potentially long after the operation has terminated. Thus, to avoid situations where helpers read inconsistent data in a descriptor and corrupt the data structure, or try to access a descriptor that has been freed to the operating system and crash, each descriptor must remain consistent and accessible until it can be determined that no helper will ever access it again. This leads to *wasteful algorithms* which allocate a new descriptor for each operation.

In this chapter, we introduce two simple abstract data types (ADTs) that capture the way descriptors are used by wasteful algorithms. The *immutable descriptor* ADT appears in Section 12.1.1. It provides two operations, *CreateNew* and *ReadField*, which respectively create and initialize a new descriptor, and read one of its fields. The *mutable descriptor* ADT, which appears in Section 12.1.2, extends the immutable descriptor ADT by adding two operations: *WriteField* and *CASField*. These allow a helper to modify fields of the descriptor (e.g., to indicate that the operation has been partially or fully completed). We also give examples of wasteful algorithms whose usage of descriptors is captured by these ADTs.

The natural way to implement the immutable and mutable descriptor ADTs is to have *CreateNew* allocate memory and initialize it, and to have *ReadField*, *WriteField* and *CASField* perform a read, write and CAS, respectively. Every implementation of one of these ADTs must eventually reclaim the descriptors it allocates. Otherwise, the algorithm would eventually exhaust memory, and either cause the system to crash, or block while waiting for more memory to become available, violating lock-free progress. Usually, a lock-free memory reclamation algorithm is used for the reclamation of descriptors. We briefly explain why reclaiming descriptors this way is expensive.

It is non-trivial to determine when a descriptor is safe to free. In order to safely free a descriptor, a process must know that the descriptor is no longer *reachable*. This means no other process can reach the descriptor by following pointers in shared memory *or* in its private memory. State of the art lock-free memory reclamation algorithms (e.g., DEBRA and Hazard Pointers) can determine when no process has a pointer to an object in its *private* memory, but typically require the underlying algorithm to identify a time after which the object is no longer reachable from *shared* memory (and then invoke a *Retire* function).

Thus, for each high-level operation attempt O , an algorithm must identify a time t such that no operation attempt started after t can encounter a pointer in shared memory to O 's descriptor. In an algorithm where each operation attempt removes all pointers to its descriptor from shared memory before it terminates, t is when O completes. However, in some algorithms (such as the implementation of *LLX* and *SCX* in Chapter 3), pointers to descriptors are “lazily” cleaned up by subsequent operation attempts. In such an algorithm, t may be long after O completes (and, consequently, t may be quite difficult to identify). The overhead of reclaiming descriptors comes both from identifying t , and from actually running a lock-free memory reclamation algorithm.

Additionally, in some applications, such as embedded systems, it is important to have a small, predictable number of descriptors in the system. In such cases, one must use memory reclamation algorithms that prioritize having a small *descriptor footprint*, i.e., the largest number of descriptors in the system at one time. Such algorithms incur high overhead. For example, *hazard pointers* [99] can be used to achieve a small descriptor footprint, but it must perform costly memory fences *every* time a process tries to access a new descriptor (as we described in Chapter 11).

To circumvent the aforementioned problems, we introduce a *weak descriptor* ADT (in Section 12.2) that has slightly *weaker semantics* than the mutable descriptor ADT, but can be implemented *without memory reclamation*. The crucial difference is that each time a process invokes *CreateNew* to create a new descriptor, it *invalidates* all of its previous descriptors. An invocation of *ReadField* on an invalid descriptor *fails* and returns a special value \perp . Invocations of *WriteField* and *CASField* on invalid descriptors have no effect. We believe the weak descriptor ADT can be useful in designing new lock-free algorithms, since an invocation of *ReadField* that returns \perp can be used to inform a helper that it no longer needs to continue helping (making further accesses to the descriptor unnecessary).

We also identify a class of lock-free algorithms that use the descriptor ADT, and which can be *transformed* to use the weak descriptor ADT (in Section 12.2.2). At a high level, these are algorithms in which (1) each operation attempt creates a descriptor and invokes a *Help* function on it, and (2) *ReadField*, *WriteField* and *CASField* operations occur only inside invocations of *Help*. Intuitively, the fact that these operations occur only in *Help* makes it easy to determine how the transformed algorithm should proceed when it performs an invalid operation: the operation being helped must have already terminated, so it no longer needs help. We prove correctness for our transformation, and demonstrate its use by transforming a wasteful implementation of a double-compare-single-swap (DCSS) primitive [62].

We then present an extension to our weak descriptor ADT, and show how an even larger class of lock-free algorithms can be transformed to use this extension (in Section 12.3). In particular, the algorithms in this class can also perform *ReadField* operations *outside* of *Help*. We prove correctness and progress for the transformation, and demonstrate its use by transforming a wasteful implementation of a k -compare-and-swap (k -CAS) primitive [62], as well as the *LLX* and *SCX* implementation in Chapter 3.

We used mostly known techniques to produce an efficient, provably correct implementation of our extended weak descriptor ADT in Section 12.4. With this implementation, the transformed algorithms for k -CAS, and LLX and SCX, have some desirable properties. In the original k -CAS algorithm, *each operation attempt* allocates at least $k + 1$ new descriptors. In contrast, the transformed algorithm allocates only two descriptors *per process, once, at the beginning of the execution*, and these descriptors are reused. Similarly, in the original algorithm for LLX and SCX, each SCX operation creates a new descriptor, but the transformed algorithm allocates only one descriptor per process, at the beginning of the execution. This entirely eliminates dynamic allocation *and* memory reclamation for descriptors (significantly reducing overhead), and results in an extremely small descriptor footprint.

We present extensive experiments on a 64-thread AMD system and a 48-thread Intel system (in Section 12.5). These experiments use a variety of workloads to compare our transformed implementations with wasteful implementations that use state of the art memory reclamation algorithms. Our results show that our transformed implementations always perform at least as well as their wasteful counterparts, and *significantly* outperform them in some workloads. In a k -CAS microbenchmark, our implementation outperformed wasteful implementations using fast distributed epoch-based reclamation [28], hazard pointers [99] and read-copy-update (RCU) [43] by up to 2.3x, 3.3x and 5.0x, respectively. In a microbenchmark using a binary search tree (BST) implemented with LLX and SCX, our transformed implementation is up to 57% faster than the next best wasteful implementation.

The crucial observation in this work is that, in algorithms where descriptors are used only to facilitate helping, a descriptor is no longer needed once the operation that created it has terminated. This allows a process to reuse a descriptor as soon as its operation attempt finishes, instead of allocating a new descriptor for each operation attempt, and waiting considerably longer (and incurring much higher overhead) to reclaim it using standard memory reclamation techniques. The challenge in this work is to characterize the set of algorithms that can benefit from this observation, and to design and prove the correctness of a transformation that takes such algorithms and produces new

algorithms that simply reuse a small number of descriptors. As a result of developing this transformation, we also produce significantly faster implementations of k -CAS, and *LLX* and *SCX*.

12.1 Wasteful Algorithms

In this section, we describe increasingly complex classes of lock-free wasteful algorithms, and progressively build up a descriptor ADT to capture their behaviour. First, we consider algorithms in which descriptors are not changed after they are initialized. Such descriptors are called *immutable*. We then discuss algorithms in which descriptors are modified by helpers. We call such descriptors *mutable*.

For the sake of illustration, we start by describing one common way that lock-free wasteful algorithms are implemented. Consider a lock-free algorithm that implements a set of *high-level* operations. Each high-level operation consists of one or more *attempts*, which either succeed, or fail due to contention. Each high-level operation attempt accesses a set of objects (e.g., individual memory locations or nodes of a tree). Conceptually, a high-level operation attempt locks a subset of these objects and then possibly modifies some of them. These locks are special: instead of providing exclusive access to a *process*, they provide exclusive access to a *high-level operation attempt*. Whenever a high-level operation attempt by a process p is unable to lock an object because it is already locked by another high-level operation attempt O , p first *helps* O to complete, before continuing its own attempt or starting a new one. By helping O complete, p effectively removes the locks that prevent it from making progress. Note that p is able to access objects locked for a different high-level operation attempt (which is not possible in traditional lock-based algorithms), but only for the purpose of helping the other high-level operation attempt complete.

We now discuss how helping is implemented. Each high-level operation or operation attempt allocates a new *descriptor* object, and fills it with information that describes any modifications it will perform. This information will be used by any processes that help the high-level operation attempt. For example, if the lock-free algorithm performs its modifications with a sequence of CAS steps, then the descriptor might contain the addresses, expected values and new values for the CAS steps.

A high-level operation attempt locks each object it would like to access by publishing pointers to its descriptor, typically using CAS. Each pointer may be published in a dedicated field for descriptor pointers, or in a memory location that is also used to store application values. For example, in the BST of Ellen et al., nodes have a separate field for descriptor pointers [52], but in Harris' implementation of multi-word CAS from single-word CAS, high-level operations temporarily replace application values with pointers to descriptors [62].

When a process encounters a pointer ptr to a descriptor (for a high-level operation attempt that is not its own), it may decide to help the other high-level operation attempt by invoking a function $Help(ptr)$. Typically, $Help(ptr)$ is also invoked by the process that started the high-level operation. That is, the mechanism used to help is the same one used by a process to perform its own high-level operation attempt.

Wasteful algorithms typically assume that, whenever an operation attempt allocates a new descriptor, it uses fresh memory that has never previously been allocated. If this assumption is violated, then an *ABA problem* may occur. Suppose a process p reads an address x and sees A , then performs a CAS to change x from A to C , and interprets the success of the CAS to mean that x contained A at all times between the read and CAS. If another process changes x from A to B and back to A between p 's read and CAS, then p 's interpretation is invalid, and an ABA problem has occurred. Note that safe memory reclamation algorithms will reclaim a descriptor only if no process has, or can obtain, a pointer to it. Thus, no process can tell whether a descriptor is allocated fresh or reclaimed memory. So, safe memory

reclamation will not introduce ABA problems.

12.1.1 Immutable descriptors

We give a trivial *immutable descriptor* ADT that captures the way that descriptors are used by the class of wasteful algorithms we just described. A *descriptor* has a set of fields, and each field contains a value. The ADT offers two operations: *CreateNew* and *ReadField*. *CreateNew* takes, as its arguments, a descriptor type and a sequence of values, one for each field of the descriptor. It returns a unique descriptor pointer *des* that has never previously been returned by *CreateNew*. Every descriptor pointer returned by *CreateNew* represents a new immutable descriptor object. *ReadField* takes, as its arguments, a descriptor pointer *des* and a field *f*, and returns the value of *f* in *des*.

In wasteful algorithms, whenever a process wants to create a new descriptor, it simply invokes *CreateNew*. Whenever a helper wants to access a descriptor, it invokes *ReadField*.

Progress If the immutable descriptor ADT is implemented so that *CreateNew* allocates and initializes a new descriptor, and *ReadField* reads and returns a field of a descriptor, then its operations will be *wait-free* (i.e., each operation will terminate after a finite number of its own steps). However, wait-free descriptor operations are not necessary to guarantee lock-freedom for high-level operations that use descriptors. Instead, we simply require descriptor operations to be lock-free. We now explain why this is sufficient to implement lock-free data structures.

Consider a lock-free algorithm that uses a wait-free implementation of the immutable descriptor ADT. Suppose we transform this algorithm by replacing the wait-free implementation of the descriptor ADT with a *lock-free* implementation. We argue that the transformed algorithm remains lock-free. In other words, we show that, if processes take infinitely many steps in the transformed algorithm, then infinitely many high-level operations complete. In the original algorithm, if processes take infinitely many steps, then infinitely many high-level operations will complete. The only steps we change to obtain the transformed algorithm are invocations of *CreateNew* and *ReadField*, some of which might no longer terminate. Therefore, the only way the transformed algorithm can *fail* to satisfy lock-freedom is if, eventually, all processes take steps only in non-terminating invocations of *CreateNew* and *ReadField*. (Otherwise, processes take infinitely many steps of the original algorithm, so infinitely many high-level operations will succeed.) In this case, only finitely many invocations of *CreateNew* and *ReadField* will terminate. However, since *CreateNew* and *ReadField* are lock-free, infinitely many invocations of *CreateNew* and/or *ReadField* must terminate. Thus, a lock-free implementation of the immutable descriptor ADT is sufficient to implement lock-free algorithms.

Example Algorithm: DCSS We use the double-compare single-swap (*DCSS*) algorithm of Harris et al. [62] as an example of a lock-free algorithm that fits the preceding description. Its usage of descriptors is easily captured by the immutable descriptor ADT. A *DCSS*(a_1, e_1, a_2, e_2, n_2) operation does the following *atomically*. It checks whether the values in addresses a_1 and a_2 are equal to a pair of expected values, e_1 and e_2 . If so, it stores the value n_2 in a_2 and returns e_2 . Otherwise it returns the current value of a_2 .

Pseudocode for the *DCSS* algorithm appears in Figure 12.1. At a high level, *DCSS* creates a descriptor, and then attempts to lock a_2 by using CAS to replace the value in a_2 with a pointer to its descriptor. Since the *DCSS* algorithm replaces values with descriptor pointers, it needs a way to distinguish between values and descriptor pointers (in order to determine when helping is needed). So, it steals a bit from each memory location and uses this bit to *flag* descriptor pointers.

```

1  type DCSSdes : {ADDR1,EXP1,ADDR2,EXP2,NEW2}
2
3  ▷ DCSS ADT operations
4  DCSS(a1,e1,a2,e2,n2) :
5    des := CreateNew(DCSSdes,a1,e1,a2,e2,n2)
6    fdes := flag(des)
7    loop
8      r := CAS(a2,e2,fdes)
9      if r is flagged then DCSSHHelp(r)
10     else exit loop
11     if r = e2 then DCSSHHelp(fdes)
12     return r
13
14 DCSSRead(addr) :
15   loop
16     r := *addr
17     if r is flagged then DCSSHHelp(r)
18     else exit loop
19     return r
20
21 ▷ Private procedures
22 DCSSHHelp(fdes) :
23   des := unflag(fdes)
24   addr1 := ReadField(des,ADDR1)
25   addr2 := ReadField(des,ADDR2)
26   exp1 := ReadField(des,EXP1)
27   if *addr1 = exp1 then
28     new2 := ReadField(des,NEW2)
29     CAS(addr2,fdes,new2)
30   else
31     exp2 := ReadField(des,EXP2)
32     CAS(addr2,fdes,exp2)

```

Figure 12.1: Code for the *DCSS* algorithm of Harris et al. [62] using the *immutable descriptor* ADT.

We now give a more detailed description. *DCSS* starts by creating and initializing a new descriptor *des* at line 5. It then flags *des* at line 6. We call the result *fdes* a *flagged pointer*. *DCSS* then attempts to lock *a*₂ in the loop at lines 7-10. In each iteration, it tries to store its flagged pointer in *a*₂ using CAS. If the CAS is successful, then the operation attempt invokes *DCSSHHelp* to complete the operation (at line 11). Now, suppose the CAS fails. Then, the *DCSS* checks whether its CAS failed because *a*₂ contained another *DCSS* operation's flagged pointer (at line 9). If so, it invokes *DCSSHHelp* to help the other *DCSS* complete, and then retries its CAS. *DCSS* repeatedly performs its CAS (and helping) until the *DCSS* either succeeds, or fails because *a*₂ did not contain *e*₂.

DCSSHHelp takes a flagged pointer *fdes* as its argument, and begins by unflagging *fdes* (to obtain the actual descriptor pointer for the operation). Then, it reads *a*₁ and checks whether it contains *e*₁ (at line 27). If so, it uses CAS to change *a*₂ from *fdes* to *n*₂, completing the *DCSS* (at line 29). Otherwise, it uses CAS to change *a*₂ from *fdes* to *e*₂, effectively aborting the *DCSS* (at line 32). Note that this code is executed by the process that created the descriptor, and also possibly by several helpers. Some of these helpers may perform a CAS at line 27 and some may perform a CAS at line 29, but only the first of these CAS steps can succeed.

When a program uses DCSS, some addresses can contain either values or descriptor pointers. So, each read of such an address must be replaced with an invocation of a function called *DCSSRead*. *DCSSRead* takes an address *addr* as its argument, and begins by reading *addr* (at line 16). It then checks whether it read a descriptor pointer (at line 17) and, if so, invokes *DCSSHHelp* to help that *DCSS* complete. *DCSSRead* repeatedly reads and performs helping until it sees a value, which it returns (at line 19).

12.1.2 Mutable descriptors

In some more advanced lock-free algorithms, each descriptor also contains information about the *status* of its high-level operation attempt, and this status information is used to coordinate helping efforts between processes. Intuitively, the status information gives helpers some idea of what work has already been done, and what work remains to be done. Helpers use this information to direct their efforts, and update it as they make progress. As a trivial example, the state information might simply be a bit that is set (by the process that started the high-level operation, or a helper) once the high-level operation succeeds.

As another example, in an algorithm where high-level operation attempts proceed in several phases, the descriptor might store the current phase, which would be updated by helpers as they successfully complete phases. Observe that, since lock-free algorithms cannot use mutual exclusion, helpers often use CAS to avoid making conflicting changes to status information, which is quite expensive. Updating status information may introduce contention. Even when there is no contention, it adds overhead. Lock-free algorithms typically try to minimize updates to status information. Moreover, status information is usually simplistic, and is encoded using a small number of bits.

Status information might be represented as a single field in a descriptor, or it might be distributed across several fields. Any fields of a descriptor that contain status information are said to be *mutable*. All other fields are called *immutable*, because they do not change during an operation.

Mutable descriptor ADT We now extend the immutable descriptor ADT to provide operations for changing (mutable) fields of descriptors. The *mutable descriptor* ADT offers four operations: *CreateNew*, *WriteField*, *CASField* and *ReadField*. The semantics for *CreateNew* and *ReadField* are the same as in the immutable descriptor ADT. *WriteField* takes, as its arguments, a descriptor pointer *des*, a field *f* and a value *v*. It stores *v* in field *f* of *des*. *CASField* takes, as its arguments, a descriptor pointer *des*, a field *f*, an expected value *exp* and a new value *v*. Let *v_f* be the value of *f* in *des* just before the *CASField*. If *v_f = exp*, then *CASField* stores *v* in *f*. *CASField* returns *v_f*.

As in the immutable descriptor ADT, we require the operations of the mutable descriptor ADT to be lock-free.

Example Algorithm: *k*-CAS A *k*-CAS($a_1, \dots, a_k, e_1, \dots, e_k, n_1, \dots, n_k$) operation atomically does the following. First, it checks if each address a_i contains its expected value e_i . If so, it writes a new value n_i to a_i for all i and returns true. Otherwise it returns false.

The *k*-CAS algorithm of Harris et al. [62] is an example of a lock-free algorithm that has descriptors with mutable fields. At a high level, a *k*-CAS operation *O* in this algorithm starts by creating a descriptor that contains its arguments. It then tries to lock each location a_i for the operation *O* by changing the contents of a_i from e_i to *des*, where *des* is a pointer to *O*'s descriptor. If it successfully locks each location a_i , then it changes each a_i from *des* to n_i , and returns true. If it fails because a_i is locked for another operation, then it helps the other operation to complete (and unlock its addresses), and then tries again. If it fails because a_i contains an application value different from e_i , then the *k*-CAS fails, and unlocks each location a_j that it locked by changing it from *des* back to e_j , and returns false. (The same thing happens if *O* fails to lock a_i because the operation has already terminated.)

We now give a more detailed description of the algorithm. Pseudocode appears in Figure 12.2. A *k*-CAS operation creates its descriptor at line 5, and then invokes a function *k-CASHelp* to complete the operation. In addition to the arguments to its *k*-CAS operation, a *k*-CAS descriptor contains a 2-bit *state* field that initially contains *Undecided* and is changed to *Succeeded* or *Failed* depending on how the operation progresses. This *state* field is used to coordinate helpers.

```

1 type k-CASdes : {STATE, ADDR1, EXP1, NEW1, ADDR2, EXP2, NEW2, ..., ADDRk, EXPk, NEWk}
2
3 ▷ k-CAS ADT operations
4 k-CAS(a1, e1, n1, a2, e2, n2, ..., ak, ek, nk) :
5   des := CreateNew(k-CASdes, Undecided, a1, e1, n1, ...)
6   fdes := flagged version of des
7   return k-CASHelp(fdes)
8
9 k-CASRead(addr) :
10  loop
11    r := DCSSRead(addr)
12    if r is flagged then k-CASHelp(r)
13    else exit loop
14  return r
15
16 ▷ Private procedures
17 k-CASHelp(fdes) :
18   des := remove the flag from fdes
19   ▷ Use DCSS to store fdes in each of a1, a2, ..., ak
20   ▷ only if des has STATE Undecided and ai = ei for all i
21   if ReadField(des, STATE) = Undecided then
22     state := Succeeded
23     for i = 1...k do
24       retry_entry:
25         a1 := ReadField(des, STATE)
26         a2 := ReadField(des, ADDRi)
27         e2 := ReadField(des, EXPi)
28         val := DCSS({des, STATE}, Undecided, a2, e2, fdes)
29         if val is flagged then
30           if val ≠ fdes then
31             k-CASHelp(val)
32             goto retry_entry
33         else
34           if val ≠ e2 then
35             state := Failed
36             break
37       CASField(des, STATE, Undecided, state)
38
39   ▷ Replace fdes in a1, ..., ak with n1, ..., nk or e1, ..., ek
40   state := ReadField(des, STATE)
41   for i = 1...k do
42     a = ReadField(des, ADDRi)
43     if state = Succeeded then
44       new := ReadField(des, NEWi)
45     else
46       new := ReadField(des, EXPi)
47     CAS(a, fdes, new)
48   return (state = Succeeded)

```

Figure 12.2: Code for the *k-CAS* algorithm of Harris et al. [62] using the *mutable descriptor* ADT.

Let *p* be a process performing (or helping) a *k-CAS* operation *O* that created a descriptor *d*. If *p* fails to lock some address *a_i* in *d*, then *p* attempts to change the *state* of *d* using CAS from *Undecided* to *Failed*. On the other hand, if *p* successfully locks each address in *d*, then *p* attempts to change the *state* of *d* using CAS from *Undecided* to *Succeeded*. Since the *state* field changes only from *Undecided* to either *Failed* or *Succeeded*, only the first CAS on the state field of *d* will succeed. The *k-CAS* implementation then uses a lock-free DCSS primitive (the one presented in Section 12.1.1) to ensure that *p* can lock addresses for *O* only while *d*'s *state* is *Undecided*. This prevents helpers from erroneously performing successful CAS steps after the *k-CAS* operation is already over.

Recall that the DCSS algorithm allocates a descriptor for each DCSS operation. A *k-CAS* operation performs

potentially *many* DCSS operations (at least k for a successful k -CAS), and also allocates its own k -CAS descriptor. The k -CAS algorithm need not be aware of DCSS descriptors (or of the bit reserved in each memory location by the DCSS algorithm to flag values as DCSS descriptor pointers), since it can simply use the *DCSSRead* procedure described above whenever it accesses a memory location that might contain a DCSS descriptor. However, the converse is *not true*, since the k -CAS algorithm performs DCSS on the *state* field of a k -CAS descriptor. Of course, the *state* field must be accessed using the k -CAS descriptor's *ReadField* operation. So, to allow DCSS to access the *state* field, we must modify DCSS slightly. First, instead of passing an address a_1 to DCSS, we pass a pointer to the k -CAS descriptor and the name of the *state* field (at line 28 of Figure 12.2). Second, we replace the read of $addr_1$ in DCSS (at line 27 of Figure 12.1) with an invocation of *ReadField*.

Since k -CAS descriptor pointers are temporarily stored in memory locations that normally contain application values, the k -CAS algorithm needs a way to determine whether a value in a memory location is an application value or a k -CAS descriptor pointer. In the DCSS algorithm, the solution was to reserve a bit in each memory location, and use this bit to *flag* the value contained in the location as a pointer to a DCSS descriptor. Similarly, the k -CAS algorithm reserves a bit in each memory location to flag a value as a k -CAS descriptor pointer. The k -CAS and DCSS algorithms need not be aware of each other's reserved bits, but they should not reserve the same bit (or else, e.g., a DCSS operation could encounter a k -CAS descriptor pointer, and interpret it as a DCSS descriptor pointer).

When the k -CAS algorithm is used, some memory addresses may contain either values or descriptor pointers, so reads of such addresses must be replaced by a *k-CASRead* operation. This operation reads an address, and checks whether it contains a k -CAS descriptor pointer. If so, it helps the k -CAS operation to complete, and tries again. Otherwise, it returns the value it read. For further details, refer to [62].

12.2 Weak descriptors

In this section we present a *weak descriptor* ADT that has weaker semantics than the mutable descriptor ADT, but can be implemented more efficiently (in particular, without requiring any memory reclamation for descriptors). We identify a class of algorithms that use the mutable descriptor ADT, and which can be transformed to use the weak descriptor ADT, instead.

We first discuss a restricted case where operation attempts only create a single descriptor, and we give an ADT, transformation and proof for that restricted case. (In the next section, we describe how the ADT and transformation can be modified slightly to support operation attempts that create multiple descriptors.)

12.2.1 Weak descriptor ADT

The weak descriptor ADT is a variant of the mutable descriptor ADT that allows some operations to *fail*. To ease the discussion, we introduce the concept of descriptor validity. Let des be a pointer returned by a *CreateNew* operation O by a process p , and d be the descriptor pointed to by des . In each configuration, d is either **valid** or **invalid**. Initially, d is valid. If p performs another *CreateNew* operation O' after O , then d becomes invalid immediately after O' (and will never be valid again).

We say that a *ReadField*(des, \dots), *WriteField*(des, \dots) or *CASField*(des, \dots) operation is performed **on a descriptor** d , where des is a pointer to d . An operation on a valid (resp., invalid) descriptor is said to be valid (resp., invalid). Invalid operations have no effect on any base object, and return a special value \perp (which is never contained in a field

of any descriptor) instead of their usual return value. We say that a $CreateNew(T, \dots)$ operation O is performed **on a descriptor** d if O returns a pointer to d . Observe that a $CreateNew$ operation is always valid.

The semantics for $CreateNew$ are the same as in the mutable descriptor ADT. The semantics for the other three operations are the same as in the mutable descriptor ADT, except that they can be invalid.

As in the other descriptor ADTs, we require the operations of the weak descriptor ADT to be lock-free.

12.2.2 Transforming a class of algorithms to use the weak descriptor ADT

We now formally define a class of lock-free algorithms that use the mutable descriptor ADT, and can easily be transformed so that they use the weak descriptor ADT, instead. We say that a process p **owns** a descriptor d if it performed a $CreateNew$ operation that returned a pointer des to d . Note that, in the following, we abuse notation slightly by referring interchangeably to a descriptor and a pointer to it.

We say that a step s of an execution is *nontrivial* if it changes the state of an object o in shared memory, and *trivial* otherwise. In particular, all invalid operations are trivial, and an unsuccessful CAS or a CAS whose expected and new values are the same are both trivial.

Definition 1 *Weak-compatible algorithms (WCA) are lock-free wasteful algorithms that use the mutable descriptor ADT, and have the following properties:*

1. *Each high-level operation attempt O by a process p may create (and initialize) a single descriptor d . Inside O , p may perform at most one invocation of a function $Help(d)$ (and p may not invoke $Help(d)$ outside of O).*
2. *A process may help any operation attempt O' by another process by invoking $Help(d')$ where d' is the descriptor that was created by O' .*
3. *If O terminates at time t , then any steps taken in an invocation of $Help(d)$ after time t are trivial (i.e., do not change the state of any shared object, incl. d).*
4. *While a process $q \neq p$ is performing $Help(d)$, q cannot change any variables in its private memory that are still defined once $Help(d)$ terminates (i.e., variables that are local to the process q , but are not local to $Help$).*
5. *All accesses (read, write or CAS) to a field of d occur inside either $Help(d)$ or O .*

At a high level, properties 1 and 2 of WCA describe how descriptors are created and helped. Property 4 intuitively states that, whenever a process q finishes helping another process perform its operation attempt, q knows only that it finished helping, and does not remember anything about what it did while helping the other process. In particular, this means that q cannot pay attention to the return value of $Help$. We explain why this behaviour makes sense. If q creates a descriptor d as part of a high-level operation attempt O and invokes $Help(d)$, then q might care about the return value of $Help$, since it needs to compute the response of O . However, if q is just helping another process p 's high-level operation attempt O , then it does not care about the response of $Help$, since it does not need to compute the response of O . The remaining properties, 3 and 5, allow us to argue that the contents of a descriptor are no longer needed once the operation that created it has terminated (and, hence, it makes sense for the descriptor to become invalid). In Section 12.3, we will study a larger class of algorithms with a weaker version of property 5.

```

1  DCSSHelp(fdes) :
2    des := Unflag(fdes)
3    addr1 := ReadField(des, ADDR1)
4    if addr1 = ⊥ then return
5    addr2 := ReadField(des, ADDR2)
6    if addr2 = ⊥ then return
7    exp1 := ReadField(des, EXP1)
8    if exp1 = ⊥ then return
9    if *addr1 = exp1 then
10      new2 := ReadField(des, NEW2)
11      if new2 = ⊥ then return
12      CAS(addr2, fdes, new2)
13    else
14      exp2 := ReadField(des, EXP2)
15      if exp2 = ⊥ then return
16      CAS(addr2, fdes, exp2)

```

Figure 12.3: Applying the transformation to DCSS.

The transformation Each algorithm in WCA can be transformed in a straightforward way into an algorithm that uses the weak descriptor ADT as follows. Consider any *ReadField* or *CASField* operation op performed by a high-level operation attempt O in an invocation of $Help(d)$, where d was created by a *different* high-level operation attempt O' . Note that op is performed while O is *helping* O' . After op , a check is added to determine whether op was invalid, in which case p returns from $Help$ immediately. (In this case, $Help$ does not need to continue, since op will be invalid only if O' has already been completed by the process that owns d or a helper.)

Example Algorithm: DCSS Figure 12.3 shows code for the *DCSS* algorithm in Figure 12.1 that has been *transformed* to use the weak descriptor ADT. There, we include only the *DCSSHelp* procedure, since it is the only one that differs from Figure 12.1. The transformation adds lines 4, 6, 8, 11 and 15 to *check* whether the preceding invocations of *ReadField* are invalid.

Correctness We argue that our transformation takes a linearizable algorithm $\mathcal{A} \in \text{WCA}$ that uses mutable descriptors and produces a linearizable algorithm \mathcal{A}' that uses weak descriptors. Consider any execution e' of the transformed algorithm \mathcal{A}' . We prove there exists an execution e of the original algorithm \mathcal{A} that performs the *same* high-level operations, in the same order, and with the same responses, as in e' . We explain how this helps. Since e is a correct execution of the original algorithm \mathcal{A} , the high-level operations performed in e must respect the sequential specification(s) of the object(s) implemented in \mathcal{A} . Furthermore, since e' performs the same high-level operations, in the same order, and with the same responses, the high-level operations in e' must also respect the sequential specification(s) of the same object(s). Therefore, the transformed algorithm \mathcal{A}' is correct.

We construct e as follows. By Property 5 of WCA, all *ReadField*, *WriteField* and *CASField* operations occur in *Help*. Whenever a check by a process p follows a *ReadField* or *CASField* in e' that returns \perp (because the operation attempt O being helped by p has already terminated), we replace that check by a consecutive sequence of steps in which p finishes its invocation of *Help*. All other checks immediately following *ReadField* or *CASField* are simply removed.

By Property 3 of WCA, none of the steps added to e change the state of any shared object. So, these steps will not change the behaviour of any other process. We also argue that none of these steps make any changes to p 's private memory that persist after p finishes its invocation of *Help*. (I.e., any changes these steps make to p 's private memory

```

1  DCSSHelp(fdes) :
2    des := Unflag(fdes)
3    values := ReadImmutables(des)
4    if values = ⊥ then return
5    ⟨addr1, exp1, addr2, exp2, new2⟩ := values
6
7    if *addr1 = exp1 then
8      CAS(addr2, fdes, new2)
9    else
10      CAS(addr2, fdes, exp2)

```

Figure 12.4: Using *ReadImmutables* to optimize and streamline the transformed DCSS algorithm.

are *reverted* by the time p finishes its invocation of *Help*, so p 's private memory is the same just after the invocation of *Help* as it was just before the invocation of *Help*.) So, these steps will not change the behaviour of p after it finishes its invocation of *Help*. Observe that, whenever a process performs a *ReadField* or *CASField* operation on a descriptor that it created, this operation will return a value different from \perp . This is due to Property 1 of WCA, and the definition of the weak descriptor ADT, which states that d becomes invalid only after O has terminated. Since p 's invocation of *ReadField* or *CASField* returns \perp , p must therefore be performing $\text{Help}(d)$ where d was created by a *different* process. Thus, Property 4 of WCA implies that, after p performs the sequence of steps to finish its invocation of $\text{Help}(d)$, its private memory has the same state as it did just before it invoked *Help*.

Reading immutable fields efficiently If an invocation of $\text{Help}(des)$ accesses many immutable fields of a descriptor, then we can optimize it by replacing many *ReadField* operations with something more efficient. To this end, we can add a new operation, *ReadImmutables*. This operation reads and returns *all* of a descriptor's immutable fields, unless the descriptor is invalid, in which case it returns \perp .

To use *ReadImmutables* in $\text{Help}(des)$, one can simply perform, at the beginning of *Help*, a *ReadImmutables* operation, followed by an *if*-statement that checks whether it the operation invalid, and, if so, returns immediately. Then, in the body of $\text{Help}(des)$, each invocation of $\text{ReadField}(des, f)$, where f is immutable, is replaced with a direct read from the set of values returned by *ReadImmutables*. We demonstrate this approach on the transformed pseudocode for DCSS in Figure 12.3. Figure 12.4 shows the result. Since all fields of a DCSS descriptor are immutable, *every* invocation of *ReadField* can be replaced with a direct read from the result of the *ReadImmutables* operation performed at line 3. (This will not be the case in an algorithm where the *Help* procedure reads mutable fields.) Since *ReadImmutables* replaces several invocations of *ReadField*, it has the added benefit of making code simpler and shorter.

12.3 Extended Weak Descriptors

In this section, we describe an extended version of the weak descriptor ADT, and an extended version of the transformation in Section 12.2.2. This extended transformation weakens property 5 of WCA so that *ReadField* operations on a descriptor d can also be performed *outside* of $\text{Help}(d)$.

Extended weak descriptor ADT This ADT is the same as the weak descriptor ADT, except that *ReadField* is extended to take, as an additional argument, a default value dv that is returned instead of \perp when the operation is invalid. Observe that the weak descriptor ADT is a special case of the extended weak descriptor ADT where each argument dv to an invocation of *ReadField* is \perp .

The extended transformation We now describe how the *weak transformation* in Section 12.2.2 is extended. The transformation is similar to the one in Section 12.2.2. For *CASField* and *WriteField* operations, the transformation is the same. However, an invocation of *ReadField(des, f)* is handled differently depending on whether it occurs inside an invocation of *Help(des)*. If it is, it is replaced with an invocation of *ReadField(des, f, ⊥)* followed by the check, as in the WCA transformation. If not, it is replaced with an invocation of *ReadField(des, f, dv)*, where the choice of *dv* is specific to the algorithm being transformed.

Let \mathcal{A} be any algorithm that uses mutable descriptors, and satisfies properties 1-4 of WCA algorithms (see Definition 1), as well as a weaker version of property 5 which states: every write or CAS to a field of a descriptor d must occur in an invocation of *Help(d)*. Consider an extended transformation of \mathcal{A} . Let e be an execution of \mathcal{A} and let e' be an execution that is the same as e , except that one (arbitrary) descriptor d becomes invalid at some point t after the high-level operation attempt O that created d terminates. (When we say that d becomes invalid at time t , we mean that after t , each invocation of *ReadField(d, f, dv)* that is performed outside of *Help(d)* returns its default value *dv*.)

Let O' be any high-level operation attempt in e' which, after t , performs *ReadField* on d outside of *Help(d)*. We say that the extended transformation is *correct for \mathcal{A}* if, for all choices of e, e', d, t , and O' , the exact same changes are performed by O' in e and e' to any variables that are still defined once O' terminates (i.e., variables that are local to the process performing O' , but are not local to O' , and variables in shared memory), and O' returns the same response in both executions. An algorithm \mathcal{A} is an *extended weak-compatible algorithm* (and is in the class *EWCA*) if there is an extended transformation that is correct for \mathcal{A} .

Correctness Consider any extended transformation which is correct for a linearizable algorithm \mathcal{A} that uses mutable descriptors. We prove the result of applying this transformation to \mathcal{A} is a linearizable algorithm \mathcal{A}' that uses extended weak descriptors. Specifically, let e' be any execution of \mathcal{A}' . We prove there is an execution e of \mathcal{A} that performs the same high-level operations, in the same order, with the same responses, as in e' .

First, we define an execution e_0 . Whenever a check in e' by a process p in *Help(d)* determines that the preceding *ReadField* or *CASField* on a descriptor d is *invalid* (which means that the operation attempt being helped by p has already terminated), we replace that check by a consecutive sequence of steps in which p finishes its invocation of *Help(d)*. By Property 3 of WCA, none of these added steps change the state of any shared variable. Moreover, by Property 4 of WCA, p does not change any variable that is still defined after its invocation of *Help*, so p has the same local state after *Help* in e_0 and e' . Whenever such a check determines that the preceding *ReadField* or *CASField* is *valid*, we simply remove this check. Observe that each invalid operation in e_0 is an invalid *ReadField* operation on some descriptor d performed outside of *Help(d)*.

Let d_1, d_2, \dots be the sequence of descriptors created in e_0 . We inductively construct a sequence e_1, e_2, \dots of executions such that e_i differs from e_{i-1} only in that descriptor d_i never becomes invalid in e_i . Specifically, for each high-level operation attempt O' that performs an invalid *ReadField* operation on descriptor d_i outside of *Help(d_i)*, consider the first such *ReadField* operation R . All of the steps of O' prior to R are the same in e_i as in e_{i-1} . After R , O' continues to take steps in e_i , but each *ReadField* operation that O' performs on a field f of d_i returns the contents of f (instead of a default value). This may result in O' executing completely different code paths in e_{i-1} and e_i . However, by the definition of an extended transformation that is correct for \mathcal{A} , O' returns the same response in e_i and e_{i-1} and performs the *exact same changes* to any variables that are still defined once O' terminates. Thus, for each variable v that is still defined once O' terminates, we can schedule the sequence of changes to v in the exact same way in e_i and e_{i-1} (which implies that any reads in e_{i-1} which see these changes can be scheduled appropriately in e_i).

Since the claim holds for all i , there is an execution e in which no descriptor becomes invalid (so e is an execution of \mathcal{A}), and the same high-level operation attempts are performed, in the same order, and with the same responses.

Multiple descriptors per operation attempt In some lock-free algorithms, a high-level operation attempt can create several different descriptors, and potentially invoke a different *Help* procedure for each descriptor. We describe how to adjust the definitions above to support these kinds of algorithms. For simplicity, we think of there being a single *Help* procedure that checks the type of the descriptor passed to it, and behaves differently for different types.

In order to allow a high-level operation attempt to create multiple descriptors without simply invalidating the ones it previously created, we update the definition of valid and invalid descriptors. Let des be a pointer to a descriptor d of type T returned by a *CreateNew* operation C performed by process p . Initially, d is valid. If p performs another *CreateNew* operation C' with the *same descriptor type* T after C , then d becomes invalid immediately after C' (and will never be valid again).

With this definition of valid and invalid descriptors, it might initially seem like an operation cannot create multiple descriptors of the same type T . However, this turns out not to be a problem. If an operation should create multiple descriptors of type T , we can simply imagine creating multiple *clone* types T_1, T_2, \dots that have the exact same fields as T . To create k descriptors of type T , one would then create k clone types, and have an operation invoke *CreateNew* once for each clone type. (However, we are unaware of any algorithms in which a high-level operation attempt creates multiple descriptors of the same type.)

We also slightly modify Property 1 of (extended) weak-compatible algorithms, as follows, to accommodate the use of multiple descriptors. Each high-level operation attempt O by a process p may create (and initialize) a sequence D of descriptors, each with a **unique type**. Inside O , p may perform at most one invocation of a function $Help(d)$ for each $d \in D$ (and p may not invoke $Help(d)$ outside of O). Note that the proof for the extended weak transformation goes through unchanged.

12.3.1 Example Algorithm: k-CAS

In this section, we explain how the extended transformation is applied to the k -CAS algorithm presented in Section 12.1.2. There is only one place in the algorithm where an invocation I of *ReadField* on a k -CAS descriptor des is performed *outside* of $Help(des)$ (the *Help* procedure for k -CAS). (Note that no invocations of *ReadField* on a DCSS descriptor des' are performed outside of $HelpDCSS(des')$.) Specifically, I reads the *state* field of a k -CAS descriptor inside the modified version of *HelpDCSS*. Recall that the k -CAS algorithm passes a k -CAS descriptor pointer and the name of the *state* field as the first argument to DCSS at line 28 of Figure 12.2, and the DCSS algorithm is modified to use *ReadField* at line 27 of Figure 12.1 to read this *state* field. We choose the default value $dv = Succeeded$ for this invocation of *ReadField*. We explain why this extended transformation of the k -CAS algorithm is correct.

When I is performed at line 27 of Figure 12.1, its response is compared with exp_1 , which contains *Undecided*. If I returns *Undecided*, then the CAS at line 29 is performed, and the process p performing I returns from *HelpDCSS*. Otherwise, the CAS at line 32 is performed, and p returns from *HelpDCSS*.

Suppose I is invalid. Then, we know the k -CAS operation attempt that created des has been completed. We use the following algorithm specific knowledge. After a k -CAS operation attempt has completed, its k -CAS descriptor has *state Succeeded* or *Failed* (and is never changed back to *Undecided*). (This can be determined by inspection of the code.) Thus, if I were valid, its response would *not* be *Undecided*, and p would perform the CAS at line 32 and return

from *HelpDCSS*. Since $dv = \text{Succeeded}$, p does exactly the same thing when I is invalid. (Note that the exact value of *state* is unimportant. It is only important that it is not *Undecided*.)

12.3.2 Example Algorithm: LLX and SCX

In this section, we explain how the extended transformation is applied to the *LLX* and *SCX* implementation in Chapter 3. Pseudocode for *LLX* and *SCX* using mutable descriptors is presented in Figure 12.5. Recall that each *SCX* creates a new descriptor called an *SCX-record*, which has two mutable fields: a 2-bit *state* field and an *allFrozen* bit. The *state* field contains one of three values: *InProgress*, *Committed* and *Aborted*.

The following properties of the *LLX* and *SCX* algorithm are relevant for our purposes.

- P1. Before the first invocation of *Help(des)* for an *SCX O* (performed by O or a helper) has been completed, the *SCX-record des* created by O has its *state* field set to *Committed* or *Aborted*, and, after this, the *state* field of *des* is never changed again.
- P2. A marked Data-record remains marked forever.
- P3. A marked Data-record cannot point to an *SCX-record* with *state* = *Aborted*.
- P4. Each time the *info* field of a Data-record changes, it changes to a new value that has never previously been stored there (to avoid the ABA problem).

There is only one place in the code where an invocation I of *ReadField(SCX-record, d, f, dv)* can occur outside of *Help(des)*: at line 4 of *LLX* in Figure 12.5. I reads the *state* field of d . We choose the default value $dv = \text{Committed}$ for I . We give a rigorous, but straightforward, proof that this extended transformation of *LLX* and *SCX* is correct.

Let e be an execution of the original *LLX* and *SCX* algorithm \mathcal{A} , and let e' be an execution that is the same as e , except that one arbitrary *SCX-record* d becomes invalid at some point t after the *SCX* operation attempt O that created d terminates. Let O' be any *LLX* in e which, after t , performs an invocation I of *ReadField* on d outside of *Help(d)*. We must prove that O' performs the exact same changes in e and e' to any variables that are still defined after O' terminates, and returns the same response in both executions.

Since I is invalid in e' , by definition, the *SCX O* that created d must have terminated before I . Thus, by P1, I must return *Committed* or *Aborted* in e . If I returns *Committed* in e , then I returns the same response in e and e' , so O' is exactly the same in both executions. Now, suppose I returns *Aborted* in e . We consider three cases, depending on where O' returns in e .

Case 1: O' returns at line 10 in e . If $\text{marked}_2 = \text{FALSE}$, then O' behaves exactly the same way in e and e' . So, suppose $\text{marked}_2 = \text{TRUE}$. Then, O' will enter the if-statement at line 6 in e , but not in e' . In this case, O' saw that the Data-record r pointed to an *SCX-record* with *state* = *Aborted* when it performed line 3, and that r was marked when it performed line 5. By P3, r cannot simultaneously be marked and point to an *SCX-record* with *state* = *Aborted*, so $r.\text{info}$ must change between these two lines. By P4, it must change to a value different from $r.\text{info}$, so the if-block at line 8 will not be executed in e . However, this contradicts our assumption that O' returns at line 10.

Case 2: O' returns FINALIZED at line 12 in e . Observe that O' does not execute line 9 in e (since it would then return at the following line). We first prove that O' does not execute line 9 in e' . Since O' sees $\text{marked}_1 = \text{TRUE}$ just before returning at line 12 in e , P2 implies that $\text{marked}_2 = \text{TRUE}$ (in both e and e'). Since I returns *Committed* in e' , O' will not enter the if-block at line 6 in e' . Thus, O' reaches line 11 in both e and e' .

Since I returns *Committed* in e' , and we have assumed I returns *Aborted* in e , O' will not invoke *Help* at line 11 in e or e' . Therefore, O' does not change any variable that is still defined after it terminates. So, it suffices to prove

```

type SCXdes
  ▷ Immutable descriptor fields
  NFREEZE           ▷ number of Data-records to be frozen
  NFINALIZE         ▷ number of Data-records to be finalized
  V1,V2,...      ▷ Data-records to be frozen
  R1,R2,...      ▷ Data-records to be finalized (must be a subsequence of (V1,V2,...))
  DES1,DES2,...  ▷ descriptor pointers read from V1.info,V2.info, ...
  FLD               ▷ pointer to a field of some Vi
  NEW               ▷ value to be written into the field FLD
  OLD               ▷ value previously read from the field FLD
  ▷ Mutable descriptor fields
  STATE              ▷ one of {InProgress, Committed, Aborted}
  ALLFROZEN         ▷ Boolean

1 LLX(r) by process p
2   marked1 := r.marked
3   r.info := r.info
4   state := ReadField(r.info, STATE)
5   marked2 := r.marked
6   if state = Aborted or (state = Committed and not marked2) then                                ▷ if r was not frozen at line 4
7     read r.m1,...,r.my and record the values in local variables m1,...,my
8     if r.info = r.info then                                                        ▷ if r.info contains the same
9       store ⟨r,r.info,(m1,...,my)⟩ in p's local table
10      return ⟨m1,...,my⟩
11
12      if state = InProgress then Help(r.info)
13      if marked1 then return FINALIZED
14      else return FAIL
15
16 SCX(V = <V1,V2,...,Vk>,R = <R1,R2,...,Rl>,fld,new) by process p
17   ▷ Preconditions: (1) for each r in V, p has performed an invocation Ir of LLX(r) linked to this SCX
18     (2) new is not the initial value of fld
19     (3) for each r in V, no SCX(V',R',fld,new) was linearized before Ir was linearized
20   Let des1,des2,...,desk be the descriptor pointers for V1,V2,...,Vk in p's local table of LLX results
21   Let old be the value for fld stored in p's local table of LLX results
22   des := CreateNew(SCXdes, {(NFREEZE,k),(NFINALIZE,l),(V1,V1),(V2,V2),...,(Vk,Vk),
23     (R1,R1),(R2,R2),...,(Rl,Rl),(DES1,des1),(DES2,des2),...,(DESk,desk),
24     (FLD,fld),(NEW,new),(OLD,old),(STATE,InProgress),(ALLFROZEN, FALSE)})}
25   return Help(des)

26 Help(des)
27   ▷ Freeze all Data-records in des. V to protect their mutable fields from being changed by other SCXs
28   ⟨nffreeze,nfinal,V1,V2,...,Vnfreeze,R1,R2,...,Rnfinal,des1,des2,...,desnfreeze,fld,new,old⟩ := ReadImmutables(des)
29   for i = 1..nffreeze do
30     if not CAS(Vi.info,desi,des) then                                              ▷ freezing CAS
31       if Vi.info ≠ des then
32         ▷ Could not freeze Vi because it is frozen for another SCX
33         if ReadField(des,ALLFROZEN) = TRUE then                                         ▷ frozen check step
34           ▷ the SCX has already completed successfully
35           return TRUE
36         else
37           ▷ Atomically unfreeze all Data-records frozen for this SCX
38           WriteField(des,STATE,Aborted)                                                 ▷ abort step
39           return FALSE
40
41   ▷ Finished freezing Data-records (Assert: state ∈ {InProgress,Committed})                               ▷ frozen step
42   WriteField(des,ALLFROZEN,TRUE)
43   for i = 1..nfinal do
44     Ri.marked := TRUE
45     CAS(fld,old,new)                                                               ▷ mark step
46   ▷ Finalize all Ri in R, and unfreeze all Vi in V that are not in R                      ▷ update CAS
47   WriteField(des,STATE,Committed)                                                 ▷ commit step
48   return TRUE

```

Figure 12.5: Code for the *LLX* and *SCX* using the *mutable descriptor* ADT.

that O' returns FINALIZED (at line 12) in e' . However, this is immediate from the fact that $marked_1 = \text{TRUE}$ in O' in e (and, hence, in e').

Case 3: O' returns FAIL at line 13 in e . The proof is similar to the previous case, except $marked_1 = \text{FALSE}$ in O' in e , so when O' reaches line 12, it will enter the *else*-block and return FAIL in both e and e' .

12.4 Implementing the extended weak descriptor ADT

In this section, we give an efficient implementation of the extended weak descriptor ADT. Note, however, that this implementation uses largely known techniques (similar to [94]), and is not the main contribution of this work.

At a high level, each process p uses a *single* descriptor object $D_{T,p}$ in shared memory to represent *all* descriptors of type T that it ever creates. The descriptor object $D_{T,p}$ conceptually represents p 's *current* descriptor of type T . At different times in an execution, $D_{T,p}$ represents different *abstract descriptors* created by p . We store a sequence number in $D_{T,p}$ that is incremented every time p performs $CreateNew(T, -)$. Instead of using traditional descriptor pointers, we represent each descriptor pointer as a pair of bit fields stored in a single word. These bit fields contain the name of the process who owns the descriptor, and a sequence number that indicates which invocation of $CreateNew$ conceptually created this descriptor. When a descriptor pointer is passed to an operation O on the abstract descriptor, O compares the sequence number in des with the current sequence number in $D_{T,p}$ to determine whether the operation is valid or invalid. Thus, incrementing the sequence number in $D_{T,p}$ effectively makes all abstract descriptors of type T that were previously created by p *invalid*.

Detailed description Complete pseudocode appears in Figure 12.6. We start by describing the data types and shared variables. Each descriptor contains zero or more immutable fields, and zero or more mutable fields (which are determined by the *descriptor type*), as well as a sequence number field seq . Recall that $D_{T,p}$ represents different abstract descriptors at different times. Note that the immutable fields of $D_{T,p}$ are only immutable for as long as $D_{T,p}$ represents the same abstract descriptor. When $D_{T,p}$ is reused, so that it represents a different abstract descriptor, its immutable fields can be reinitialized. Usually very few bits are required for the mutable fields, since they exist solely to capture the state of an ongoing operation (and it is inefficient to frequently change the state of a descriptor). (Every lock-free algorithm we are aware of uses at most a small constant number of bits for its mutable fields.) Consequently, we think of the sequence field and the mutable fields of a descriptor d as being packed together in a single word *mutables* of d (with subfields for the sequence field and each mutable field). (Note that, if more space is needed for mutable fields in some future algorithm, we can eliminate this assumption about the size of *mutables*, as we explain below.) We use $d.f$ to denote an immutable field f of d , $d.\text{mutables}$ to denote the field *mutables* of d , and $d.\text{mutables}.f$ to denote a mutable field f of d .

Since *mutables* fits in a single word, it can be modified atomically using CAS. By having CAS atomically operate on a mutable field and the sequence number, we can ensure that a descriptor changes only if its sequence number has not changed.

We now describe the operations. An invocation of $CreateNew(T, \dots)$ by process p first increments the sequence number of $D_{T,p}$, then initializes all of its fields, then increments the sequence number again and returns a new descriptor pointer (with the up-to-date sequence number). Observe that the descriptor pointers returned by $CreateNew$ always have even sequence numbers, and the sequence number of a descriptor is odd while it is being initialized by $CreateNew$.

```

1  ▷ Data types
2  Descriptor of type  $T$  :
3       $mutables = \langle seq, mut_1, mut_2, \dots \rangle$                                 ▷ Mutable fields
4       $imm_1, imm_2, \dots$                                          ▷ Immutable fields

6  ▷ Shared variables
7       $D_{T,p}$  for each descriptor type  $T$  and process  $p$ 

9  ▷ ADT operations
10 CreateNew( $T, v_1, v_2, \dots$ ) by process  $p$  :
11      $oldseq := D_{T,p}.mutables.seq$ 
12      $D_{T,p}.mutables.seq := oldseq + 1$ 
13     for each field  $f$  in  $D_{T,p}$ 
14         let  $value$  be the corresponding value in  $\{v_1, v_2, \dots\}$ 
15         if  $f$  is immutable then
16              $D_{T,p}.f := value$ 
17         else
18              $D_{T,p}.mutables.f := value$ 
19      $D_{T,p}.mutables.seq := oldseq + 2$ 
20     return  $\langle p, oldseq + 2 \rangle$  ▷ Descriptor identifier

22 ReadField( $des, f, dv$ ) :
23      $\langle q, seq \rangle := des$ 
24     if  $f$  is immutable then
25          $result := D_{T,q}.f$ 
26     else
27          $result := D_{T,q}.mutables.f$ 
28     if  $seq \neq D_{T,q}.mutables.seq$  then return  $dv$ 
29     return  $result$ 

31 ReadImmutables( $des$ ) :
32      $\langle q, seq \rangle := des$ 
33     for each  $f$  in  $des$ 
34         if  $f$  is immutable then add  $D_{T,q}.f$  to  $result$ 
35     if  $seq \neq D_{T,q}.mutables.seq$  then return  $\perp$ 
36     return  $result$ 

38 WriteField( $des, f, value$ ) :
39      $\langle q, seq \rangle := des$ 
40     loop
41          $exp := D_{T,q}.mutables$ 
42         if  $exp.seq \neq seq$  then return
43          $new := exp$ 
44          $new.f := value$ 
45         if  $CAS(\&D_{T,q}.mutables, exp, new)$  then return

47 CASField( $des, f, fexp, fnew$ ) :
48      $\langle q, seq \rangle := des$ 
49     loop
50          $exp := D_{T,q}.mutables$ 
51         if  $exp.seq \neq seq$  then return  $\perp$ 
52         if  $exp.f \neq fexp$  then return  $exp.f$ 
53          $new := exp$ 
54          $new.f := fnew$ 
55         if  $CAS(\&D_{T,q}.mutables, exp, new)$  then
56             return  $fnew$ 

```

Figure 12.6: Pseudocode for the *extended weak descriptor* ADT implementation.

Consequently, while a descriptor is being initialized, its sequence number does not match any descriptor pointer in the system, so no process can read or modify the descriptor's fields.

Note that this approach of incrementing a sequence number twice has been used in different contexts such as in transactional memory, where the least significant bit represents whether the sequence number is locked or unlocked. Here, the idea is slightly different, since the least significant bit represents whether the descriptor is currently being reused and initialized, or is safe to access. (Nevertheless, in some sense, one can think of the bit indicating whether the descriptor is currently being initialized as a sort of lock. It does not prevent other processes from making progress (since operations on the descriptor will terminate, but will simply be invalid), but it does prevent them from accessing fields of the descriptor as they are being changed.)

An invocation of $\text{ReadField}(\text{des}, f, \text{default})$ by p reads the value v of the mutable or immutable field f from $D_{T,p}$ followed by its sequence number s . If s matches the sequence number in the descriptor pointer des , then v is returned. Otherwise, default is returned.

ReadImmutables is similar to ReadField , except it reads all immutable fields, instead of a single field, and it returns \perp instead of default .

An invocation I of $\text{WriteField}(\text{des}, f, \text{value})$ by p performs a sequence of one or more *attempts*. In each attempt, it reads the contents old of *mutables*, including the sequence number s , from $D_{T,p}$, then checks whether s matches the sequence number in the descriptor pointer des . If the sequence numbers do not match, then the abstract descriptor represented by des is invalid, so I returns without changing f . Otherwise, I uses CAS to try to change $D_{T,p}.\text{mutables}$ from old to new , which is a copy of old in which the contents of field f have been changed (locally) to contain value . Observe that this CAS will succeed only if the sequence number in $D_{T,p}.\text{mutables}$ matches the sequence number in des . If the CAS succeeds, then I returns. Otherwise, I performs another attempt.

Note that WriteField is less efficient than performing a direct write to memory. However, since mutable fields are used merely to encode the status of an ongoing operation, there are usually very few changes to a descriptor.

CASField is quite similar to WriteField . The only differences are (1) CASField has different return values and, (2) in each attempt, it performs an additional check to determine whether $old.f$ is equal to f_{exp} , and, if not, returns $old.f$.

Practical considerations One might wonder, in an algorithm with multiple types of descriptors, why the type of a descriptor is not also encoded in descriptor pointers. In algorithms that use multiple descriptor types, any time the original algorithm accesses a field of a descriptor, it typically must know what kind of descriptor it is accessing (if, for no other reason, to compute the address of the desired field within the descriptor). In such algorithms, it would not be necessary for descriptor pointers to carry this extra information. For algorithms that access descriptors without knowing their exact types, one can include the descriptor type in descriptor pointers.

Some lock-free algorithms “steal” up to three bits from pointers to encode additional information, typically to distinguish between application values and (potentially, various types of) descriptors. To accommodate such algorithms, one can slightly shrink the sequence number in our descriptor pointers, and reserve the three lowest-order bits for use by other algorithms.

One obvious way to store the descriptors for each thread is to create an array for each descriptor type, with a slot containing a descriptor for each process. In this kind of implementation, it is extremely important to pad each slot to avoid false sharing [112]. We suggest allocating at least two cache lines for each descriptor (128 bytes on modern Intel and AMD machines).

To improve efficiency, modern Intel and AMD processors implement a relaxed memory model called total store

order (TSO) that allows certain steps in a program to be executed out of order. Specifically, a read that occurs after a write in a program can actually be executed *before* the write, as long as the read and write are not accessing the same address. This can render a concurrent algorithm incorrect if it requires a write by a process p to be visible to other processes *before* p performs a subsequent read. One can prevent this reordering by placing a memory fence (or barrier) between the write and read. CAS instructions also act as memory fences. Our implementation does not require any memory fences (beyond those implied by CAS instructions). This is an attractive property, since memory fences incur high overhead.

Our implementation uses unbounded sequence numbers. However, in practice, sequence numbers are bounded, and they may wrap around. If wraparound occurs, then two invocations of *CreateNew* might return the same descriptor pointer. This can cause an *ABA problem* if the high-level algorithm that uses descriptors relies on the uniqueness of descriptor pointers returned by *CreateNew*.

We argue that the sequence number can be made sufficiently large on modern systems for this to be a non-issue. Consider a system with a 64-bit word size. Recall that a sequence number appears both in each descriptor pointer, and also in the *mutables* field of each descriptor. A descriptor pointer contains only a process name and a sequence number, so if n bits are reserved for the process name, then $64 - n$ bits remain for the sequence number. The *mutables* field contains the descriptor's mutable fields and a sequence number, so if m bits are reserved for mutable fields, then $64 - m$ bits remain for the sequence number. Thus, if we use 14-bit process names (as the Linux kernel does), and the mutable fields of each descriptor fit in at most 14 bits, then 50 bits remain for the sequence number. We are unaware of any algorithm that requires more than three bits for mutable fields in its descriptors, so this is realistic. In this case, a single process must perform 2^{50} operations to trigger even a single wraparound. If we assume that a single process can perform one million operations per second, this will take 35 years of continuous execution. If this is still a concern, then one can use double-wide CAS (DWCAS), which is implemented on modern Intel and AMD systems, instead of CAS, to atomically operate on two adjacent words (containing a much larger sequence number).

Although we are unaware of any current lock-free algorithms that use more than three bits for mutable fields in descriptors, some future algorithm may use more. If the mutable fields of a descriptor cannot fit in the same word as a sequence number, then our approach must be modified. If the mutable fields and a sequence number can fit in two adjacent words, then one can simply use DWCAS instead of CAS. Otherwise, one can store mutable fields in their own separate words, and *replicate* the sequence number, storing a copy in the word adjacent to each mutable field. To change a mutable field, one would then perform DWCAS on the word containing the mutable field, and its adjacent sequence number. When the descriptor is reused, instead of incrementing a single sequence number, one would increment all sequence numbers.

In order to choose how many bits should be devoted to the process name in descriptor pointers, one must know an upper bound on the number of processes. We stress that this is not an onerous constraint, because the upper bound does not need to be tight. Note that one need not initially allocate descriptors for all processes that *could* be running in the system. It is straightforward to allocate a descriptor for a process the first time it invokes *CreateNew* (potentially even in batches, to amortize the cost and improve control over memory layout).

Correctness. We now prove that our implementation is linearizable. We first give the linearization points for all operations.

- Each invocation of *CreateNew* is linearized at the increment of the sequence number at line 12.

- If an invocation I of $\text{ReadField}(des, f, dv)$ returns at line 28, then it is linearized at the read of the sequence number at the same line. If I returns at line 29, then it is linearized at the preceding read of the field f : for immutable fields this is line 25, and for mutable fields this is line 27.
- Each invocation of ReadImmutables is linearized at the read of the sequence number at line 35.
- If an invocation I of $\text{WriteField}(des, f, value)$ returns at line 42, then it is linearized at the last read of the sequence number at the same line. If I returns at line 45, then it is linearized at the successful CAS at the same line.
- If an invocation I of $\text{CASField}(des, f, fexp, fnew)$ returns at line 51, then it is linearized at the last read of the sequence number at the same line. If I returns at line 56, then it is linearized at the successful CAS at the previous line. If I returns at line 52, then it is linearized at the last read at the same line.

Observation 1 *The sequence number of $D_{T,p}$ (also denoted $D_{T,p}.\text{mutables.seq}$) is written only by p in invocations of $\text{CreateNew}(T, -)$.*

Observation 2 *Every descriptor pointer returned by CreateNew has an even sequence number, and the linearization point of CreateNew always changes the sequence number of the descriptor to an odd number.*

Observation 3 *The sequence number returned by a*

$\text{CreateNew}(T, -)$ operation by p is $2 + v$ where v is the sequence number returned by p 's previous $\text{CreateNew}(T, -)$ operation, or $v = 0$ if p has not performed $\text{CreateNew}(T, -)$.

We now prove that the above linearization points are correct. Let e be an execution of our implementation of extended weak descriptors. Let $O_1, O_2 \dots O_k$ be the extended weak descriptor operations executed in e in the order they are linearized.

Theorem 1 *The responses of $O_1, O_2 \dots O_k$ respect the semantics of the extended weak descriptor ADT.*

Proof: By strong induction on the sequence of extended weak descriptor operations that terminate in e . Base case: the claim vacuously holds when no operations have returned.

Induction step: assume the return values of $O_1, O_2 \dots O_{i-1}$ follow the semantics of the extended weak descriptor ADT. Let p be the process that performs O_i , and T be the type of descriptor on which O_i is performed.

Suppose O_i is a $\text{CreateNew}(T, -)$ operation. By Observation 3, O_i returns a unique descriptor pointer.

In each of the following cases, O_i takes a descriptor pointer des as one of its arguments. Let q and seq be the process name sequence number in des , respectively. Let O_{init} be the $\text{CreateNew}(T, -)$ by q that returned des . Since des is returned by O_{init} before it is passed to any operation, O_{init} is linearized before O_i .

Suppose O_i is a ReadField that returns the default value at line 28, a CASField that returns \perp at line 51 or a ReadImmutables that returns \perp at line 35. We argue that des is invalid when O_i is linearized. In each case, O_i returns after seeing that the sequence number of $D_{T,q}$ no longer contains seq . Thus, this sequence number must change after des is returned by O_{init} , and before O_i is linearized. By Observation 1, this change to the sequence number of $D_{T,q}$ must be performed by a $\text{CreateNew}(T, -)$ operation O_{change} by q (which occurs after O_{init} , and before O_i is linearized). O_{change} changes the sequence number twice, and is linearized at the first change. Thus, O_{change} is linearized after O_{init} and before O_i , which means that des is *invalid* when O_i is linearized.

Now suppose O_i is a *ReadField* that returns at line 29, a *CASField* that returns at line 56, or a *ReadImmutables* that returns at line 36. We first argue that des is valid when O_i is linearized. In each case, O_i sees that the sequence number of $D_{T,q}$ is seq at some time t , which is either when O_i is linearized, or is after O_i is linearized. By Observation 1 and Observation 3, whenever the sequence number of $D_{T,q}$ is changed, it is changed to a new value that it never previously contained. Thus, since the sequence number of $D_{T,q}$ contains seq when O_{init} terminates, and it contains seq at time t , it contains seq at all times after O_{init} terminates and before t . Hence, the sequence number of $D_{T,q}$ contains seq when O_i is linearized. By Observation 1, q does not perform any *CreateNew*($T, -$) after O_{init} and before t , so des is valid when O_i is linearized.

We now argue that the response of O_i is correct if it is a *ReadField*(des, f, dv). The proof is similar when O_i is a *CASField* or *ReadImmutables*.

If f is immutable, then it is changed only by *CreateNew*($T, -$) operations by q . Since des is valid when O_i is linearized, O_{init} performs the last change to f before O_i is linearized. Recall that O_i starts after O_{init} terminates. Thus, the write of f in O_{init} happens before the invocation of O_i , and O_i will return the value written to f by O_{init} .

If f is mutable, then let O_{change} be the operation that performs the last change to f before O_i is linearized, and v be the value that it stores in f . Observe that O_i returns v . We show that O_{change} is the last operation that changes f and is linearized before O_i . If O_{change} is the same as O_{init} , then we are done. Otherwise, since we have argued that q does not perform any *CreateNew*($T, -$) after O_{init} and before t , O_{change} must be a *WriteField* or *CASField*. In each case, O_{change} can change f only once, with a successful CAS (at line 55 or line 45). Since O_{change} is linearized at this CAS, it is linearized before O_i . Moreover, since we have assumed that O_{change} is the last operation to change f before O_i , no other operation that changes f linearized after O_{change} and before O_i . ■

Progress The proof of lock-free progress is straightforward. Suppose, to obtain a contradiction, that there is an execution in which processes take infinitely many steps, but only finitely many (extended weak descriptor) operations terminate. Then, after some time t , no operation terminates, which means there is at least one operation O in which a process takes infinitely many steps. By inspection of Figure 12.6, O must be a *WriteField* or *CASField* operation. Suppose O is a *WriteField* operation. Then, each time O executes line 42, it sees $old.seq = seq$, and each time it executes line 45, its CAS fails and returns old without changing $D_{T,q}.mutables$. Observe that the CAS will fail only if $D_{T,q}.mutables$ changes after it is read at line 41 and before the CAS at line 45. Thus, $D_{T,q}.mutables$ changes infinitely many times in the execution. Since $D_{T,q}.mutables$ can be changed only by *WriteField* or *CASField* operations, and any operation that changes $D_{T,q}.mutables$ immediately terminates, there must be infinitely many operations of *WriteField* or *CASField* that terminate, which is a contradiction. The proof is similar when O is a *CASField* operation.

12.5 Experiments

Our experiments were run on two large-scale systems. The first is a 2-socket Intel E7-4830 v3 with 12 cores per socket and 2 hyperthreads (HTs) per core, for a total of 48 threads. Each core has a private 32KB L1 cache and 256KB L2 cache (which is shared between HTs on a core). All cores on a socket share a 30MB L3 cache. The second is a 4-socket AMD Opteron 6380 with 8 cores per socket and 2 HTs per core, for a total of 64 threads. Each core has a private 16KB L1 data cache and 2MB L2 cache (which is shared between HTs on a core). All cores on a socket share a 6MB L3 cache.

Since both machines have multiple sockets and a non-uniform memory architecture (NUMA), in all of our experiments, we pinned threads to cores so that the first socket is filled first, then the second socket is filled, and so on. Furthermore, within each socket, each core has one thread pinned to it before hyperthreading is engaged. Consequently, our graphs clearly show the effects of hyperthreading and NUMA.

For example, on the Intel machine, from thread counts 1 to 12 all threads are running on a single socket and at most one thread is pinned to each core. (**socket 1: no HTs; socket 2: empty**). From 13 to 24, all threads are running on a single socket and cores either have one or two threads pinned to them (**socket 1: HTs; socket 2: empty**). From 25 to 36, each core on the first socket has two threads pinned to it, and the remaining threads are each pinned to unique cores on the second socket (**socket 1: HTs; socket 2: no HTs**). Finally, from 37 to 48, each core on the first socket has two threads pinned to it, and cores on the second socket have one or two threads pinned to them (**socket 1: HTs; socket 2: HTs**).

Both machines have 128GB of RAM. Each runs Ubuntu 14.04 LTS. All code was compiled with the GNU C++ compiler (G++) 4.8.4 with build target x86_64-linux-gnu and compilation options `-std=c++0x -mcpu=core2 -O3`. Thread support was provided by the POSIX Threads library. We used the Performance Application Programming Interface (PAPI) library [32] to collect statistics from hardware performance counters to determine cache miss rates, stall times, instructions retired, and so on.

The system (glibc) allocator was found to have poor scaling and overall performance. Instead, we used jemalloc 4.2.1, a fast user-space allocator designed to minimize contention and improve scalability [54]. The library was dynamically linked with `LD_PRELOAD`, which is the recommended method. This allocator was found to yield vastly superior performance for all algorithms, in all benchmarks. We also tried the tcmalloc allocator from Google’s Perftools library, which is another common choice for concurrency-friendly allocation. However, performance with tcmalloc was substantially worse for all algorithms than with jemalloc.

On the AMD machine, transparent huge-pages were disabled manually in the jemalloc implementation by changing the default allocation chunk size from 2^{21} to 2^{19} using the environment parameter setting `MALLOC_CONF=lg_chunk:19`. This maintained or improved the performance for all algorithms in all workloads, and did not change the performance relationship between any pair of algorithms. The same change did not improve performance on the Intel machine (for any algorithm or workload), so the original chunk size was used.

For read-heavy workloads, it was necessary to force distribution of pages across NUMA nodes to get consistently high performance. To achieve this, we used `numactl -interleave=all` for all workloads. (Doing this did not negatively impact the performance of any workload, but its benefit was less noticeable for write-heavy workloads.)

12.5.1 *k*-CAS microbenchmark

In order to compare our reusable descriptor technique with algorithms that reclaim descriptors, we implemented *k*-CAS with several memory reclamation schemes. Specifically, we implemented a lock-free memory reclamation scheme that aggressively frees memory called *hazard pointers* [99], a (blocking) epoch-based reclamation scheme called *DEBRA* [28], and reclamation using the read-copy-update (RCU) primitives [43] (also blocking). We use *Reuse* as shorthand for our reusable descriptor based algorithm, and *DEBRA*, *HP* and *RCU* to denote the algorithms that use DEBRA, hazard pointers and RCU, respectively.

The paper by Harris et al. also describes an optimization to reduce the number of DCSS descriptors that are allocated by embedding them in the *k*-CAS descriptor. We applied this optimization, and found that it did not signifi-

cantly improve performance. Furthermore, it complicated reclamation with hazard pointers. Thus, we did not use this optimization.

Methodology. We compared our implementations of k -CAS using a simple array-based microbenchmark. For each algorithm $A \in \{Reuse, DEBRA, HP, RCU\}$, array size $S \in \{2^{14}, 2^{20}, 2^{26}\}$ and k -CAS parameter $k \in \{2, 16\}$, we run ten timed *trials* for several thread counts n . In each trial, an array of a fixed size S is allocated and each entry is initialized to zero. Then, n concurrent threads run for one second, during which each thread repeatedly chooses k uniformly random locations in the array, reads those locations, and then performs a k -CAS (using algorithm A) to increment each location by one.

As a way of validating correctness in each trial, each thread keeps track of how many successful k -CAS operations it performs. At the end of the trial, the sum of entries in the array must be k times the total number of successful k -CAS operations over all threads.

Results. The results for this benchmark appear in Figure 12.7. Error bars are not drawn on the graphs, since more than 97% of the data points have a standard deviation that is less than 5% of the mean (making them essentially too small to see).

Overall, *Reuse* outperforms every other algorithm, in every workload, on both machines. Notably, on the Intel machine, its throughput is 2.2 *times* that of the next best algorithm at 48 threads with $k = 16$ and array size 2^{26} . On the AMD machine, its throughput is 1.7 times that of the next best algorithm at 64 threads with $k = 16$ and array size 2^{20} .

On the Intel machine, with $k = 2$, NUMA effects are quite noticeable for *Reuse* in the jump from 24 to 32 threads, as threads begin running on the second socket. According the statistics we collected with PAPI, this decrease in performance corresponds to an increase in cache misses. For example, with $k = 2$ and an array of size 2^{26} in the Intel machine, jumping from 24 threads to 25 increases the number of L3 cache misses per operation from 0.7 to 1.6 (with similar increases in L1 and L2 cache misses and pipeline stalls). We believe this is due to cross-socket cache invalidations.

From the three graphs for $k = 2$ on Intel, we can see that the effect is more severe with larger absolute throughput (since the additive overhead of a cache miss is more significant). Consequently, the effect is masked by the much smaller throughput of the slower algorithms, and by the substantially lower throughputs in the $k = 16$ case, except when the array is of size 2^{14} . In the array of size 2^{14} , contention is extremely high, since each of the 48 threads are accessing 16 k -CAS addresses, each of which causes contention on the entire cache line of 8 words, for a total of 6144 array entries contended at any given time. Thus, cache misses become a dominating factor in the performance on two sockets. These effects were not observed on the AMD machine. The number of cache misses is not significantly different when crossing socket boundaries, which suggests an architectural robustness to NUMA effects that is not seen on the Intel machine.

Interestingly, absolute throughputs on the AMD machine are larger with array size 2^{20} than with sizes 2^{14} and 2^{26} . This is because the 2^{20} array size represents a sweet spot with less contention than the 2^{14} size and better cache utilization than the 2^{26} size. For example, with 64 threads and $k = 16$, *Reuse* incurred approximately 50% more cache misses with size 2^{26} than with size 2^{20} , and approximately 50% of operations helped one another with size 2^{14} , whereas less than 1% of operations helped one another with size 2^{20} .

Note, however, that this is not true on the Intel machine. There, 2^{26} is almost always as fast as 2^{20} , because of the very large shared L3 cache (which is 5x larger than on the AMD machine). This is reflected in the increased number of cycles where the processor is stalled (e.g., waiting for cache misses to be served) when moving from size 2^{20} to 2^{26} .

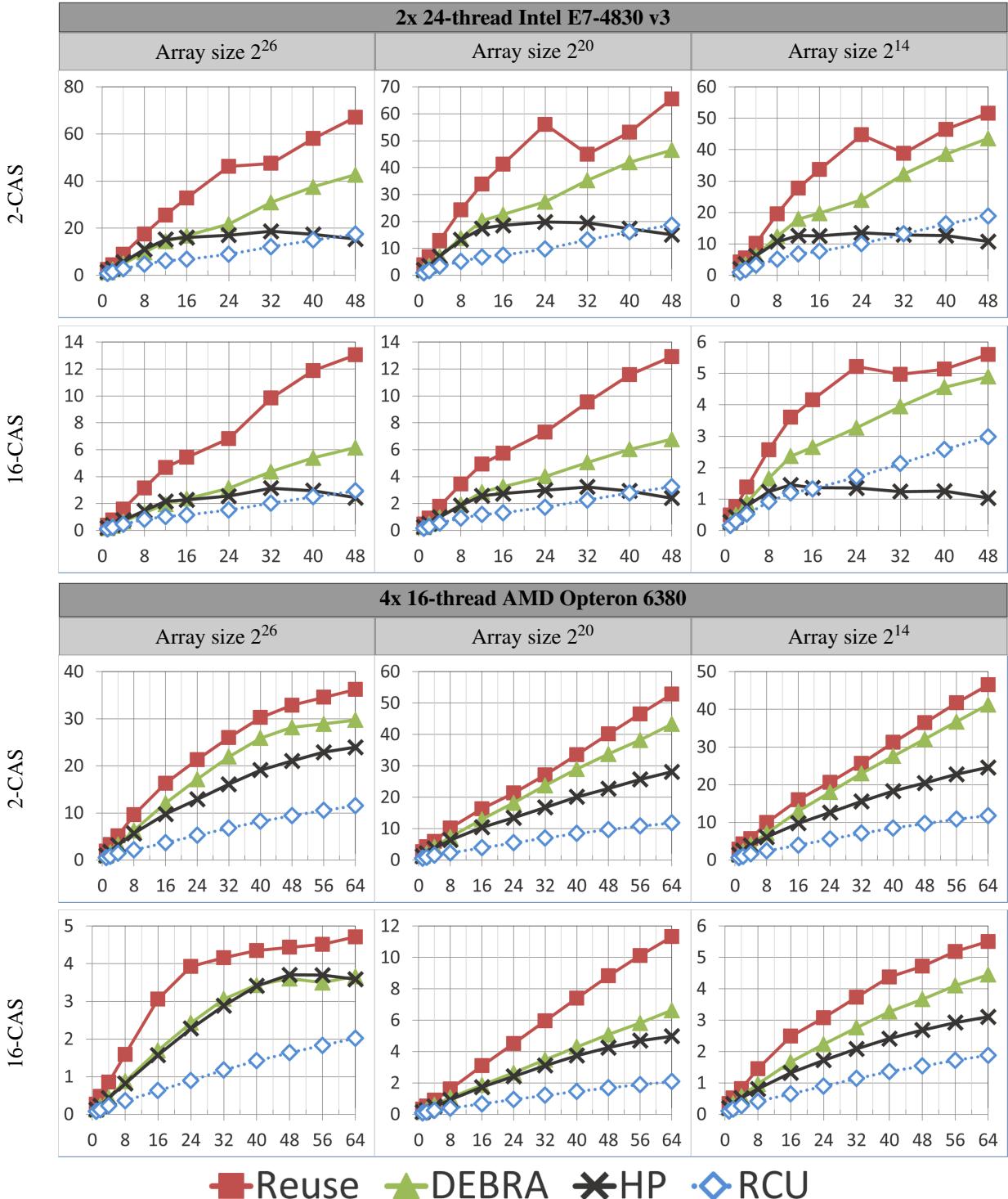


Figure 12.7: Results for a k -CAS microbenchmark. The x-axis represents the number of concurrent threads. The y-axis represents operations per microsecond.

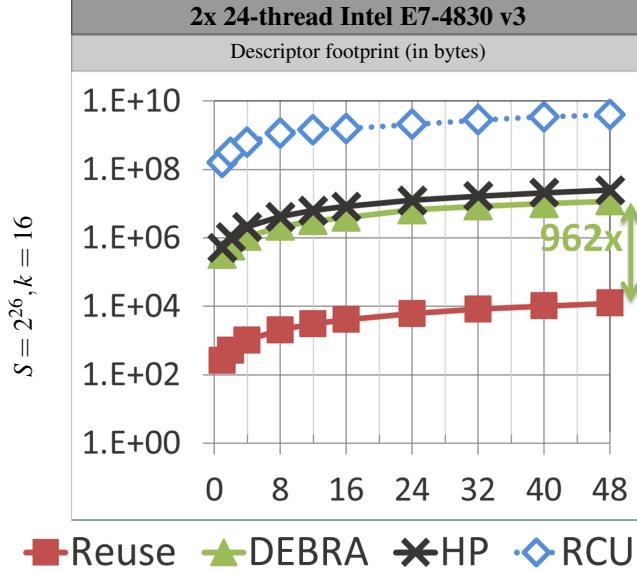


Figure 12.8: Memory usage for the k -CAS microbenchmark. The x-axis represents the number of concurrent threads. Note the logarithmic scale.

On the Intel machine, stalled cycles increase by 85% per operation, whereas on the AMD machine they increase by a whopping 450% per operation.

Memory usage We studied memory usage for all algorithms, in all workloads, on both systems, but we only show results for array size 2^{26} and $k = 16$, because the other graphs are very similar. These results appear in Figure 12.8. In particular, we are interested in the descriptor footprint, i.e., the maximum amount of memory ever occupied by descriptors in an execution. Unfortunately, computing the descriptor footprint exactly would require excessive synchronization between threads. Thus, we approximate the descriptor footprint by computing the descriptor footprint *for each thread*, and then summing those individual footprints. (This is only an approximation, since different threads may hit their peak memory usage for descriptors at different times.) The graph in Figure 12.8 contains the results of this approximation.

These results were obtained as follows. Each thread used three private variables: *totalFree*, *totalMalloc* and *maxFootprint*. Each time a thread invoked *free*, it incremented *totalFree* by the size of the descriptor being freed. Each time a thread invoked *malloc*, it incremented *totalMalloc* by the size of the descriptor being allocated, and then set *maxFootprint* = $\max\{\maxFootprint, \text{totalMalloc} - \text{totalFree}\}$. The per-thread *maxFootprints* are then summed to obtain the data points in the graph.

Note that the y-axis is a logarithmic scale. The results show that *DEBRA* and *HPs* use almost **three orders of magnitude** more memory than *Reuse* at their peaks, and *RCU* uses nearly three orders of magnitude more memory than *DEBRA* and *HPs*. *RCU*'s memory usage is significantly higher because reclamation is delayed significantly longer than in the other algorithms.

12.5.2 BST microbenchmark

Unlike in the k -CAS algorithm, where memory reclamation was only needed for descriptors, in the BST, memory reclamation is always needed for nodes. To compare our technique with different memory reclamation options, we implemented four variants of the BST algorithm: *DEBRA/DEBRA*, *DEBRA/Reuse*, *RCU/RCU* and *RCU/Reuse*. Here, an algorithm named X/Y uses X to reclaim nodes and Y for descriptors. For example, *DEBRA/Reuse* uses DEBRA to reclaim nodes and has reusable descriptors.

Methodology We compared our BST variants using a simple randomized microbenchmark. For each algorithm $A \in \{\text{DEBRA/DEBRA}, \text{DEBRA}/\text{Reuse}, \text{RCU/RCU}, \text{RCU}/\text{Reuse}\}$, key range size $K \in \{10^5, 10^6\}$ and update rate $U \in \{100, 10, 0\}$, we run ten timed *trials* for several thread counts n . Each trial proceeds in two phases: *prefilling* and *measuring*. In the prefilling phase, n concurrent threads perform 50% *Insert* and 50% *Delete* operations on keys drawn uniformly randomly from $[0, K]$ until the size of the tree converges to a steady state (containing approximately $K/2$ keys). Next, the trial enters the measuring phase, during which threads begin counting how many operations they perform. (These counts are eventually summed over all threads and reported in our graphs.) In this phase, each thread instead performs $(U/2)\%$ *Insert*, $(U/2)\%$ *Delete* and $(100 - U)\%$ *Find* operations on keys drawn uniformly from $[0, K]$ for one second.

As a way of validating correctness in each trial, each thread maintains a *checksum*. Each time a thread inserts a new key, it adds the key to its checksum. Each time a thread deletes a key, it subtracts the key from its checksum. At the end of the trial, the sum of all thread checksums must be equal to the sum of keys in the tree.

Results The results for this benchmark appear in Figure 12.9. The *Reuse* variants perform at least as well as the pure reclamation variants in every case, and significantly outperform the reclamation variants in the 100% update workload. Most notably, on the Intel machine with key range $[0, 10^6]$ and 48 threads, *DEBRA/Reuse* outperforms *DEBRA/DEBRA* by 57%, and *RCU/Reuse* outperforms *RCU/RCU* by 33%. As expected, *Reuse* does not perform significantly faster than the reclamation variants in the workloads with no updates. This is because searches do not create descriptors. However, crucially, our transformation does not impose any overhead on searches, either.

12.5.3 Studying sequence number wraparound

We performed experiments on the larger AMD machine to study how frequently errors occur when sequence numbers of varying bit-widths experience wraparound. For each bit-width $B \in \{2, 3, 4, \dots, 48\}$, we performed 200 trials in which 64 threads run for 100 milliseconds before terminating. Each trial was the same as a trial in our BST experiments with 100% updates and key range $[0, 10^5]$.

We identified three different types of errors in these trials. First, at the end of a trial, the sum of the checksums maintained by all threads would fail to match the sum of keys in the tree. Second, threads would enter infinite loops due to the tree structure being corrupted, e.g., because a cycle was introduced. (We identified this type of error by waiting until some thread had run twice as long as it should have.) Third, an invalid memory access would cause immediate program failure (e.g., due to segmentation fault or bus error).

For each B value, we divided the number of failed runs by 200 to estimate the probability of a trial failing. A graph showing the resulting estimated probability distribution appears in Figure 12.10. For small B values, trials

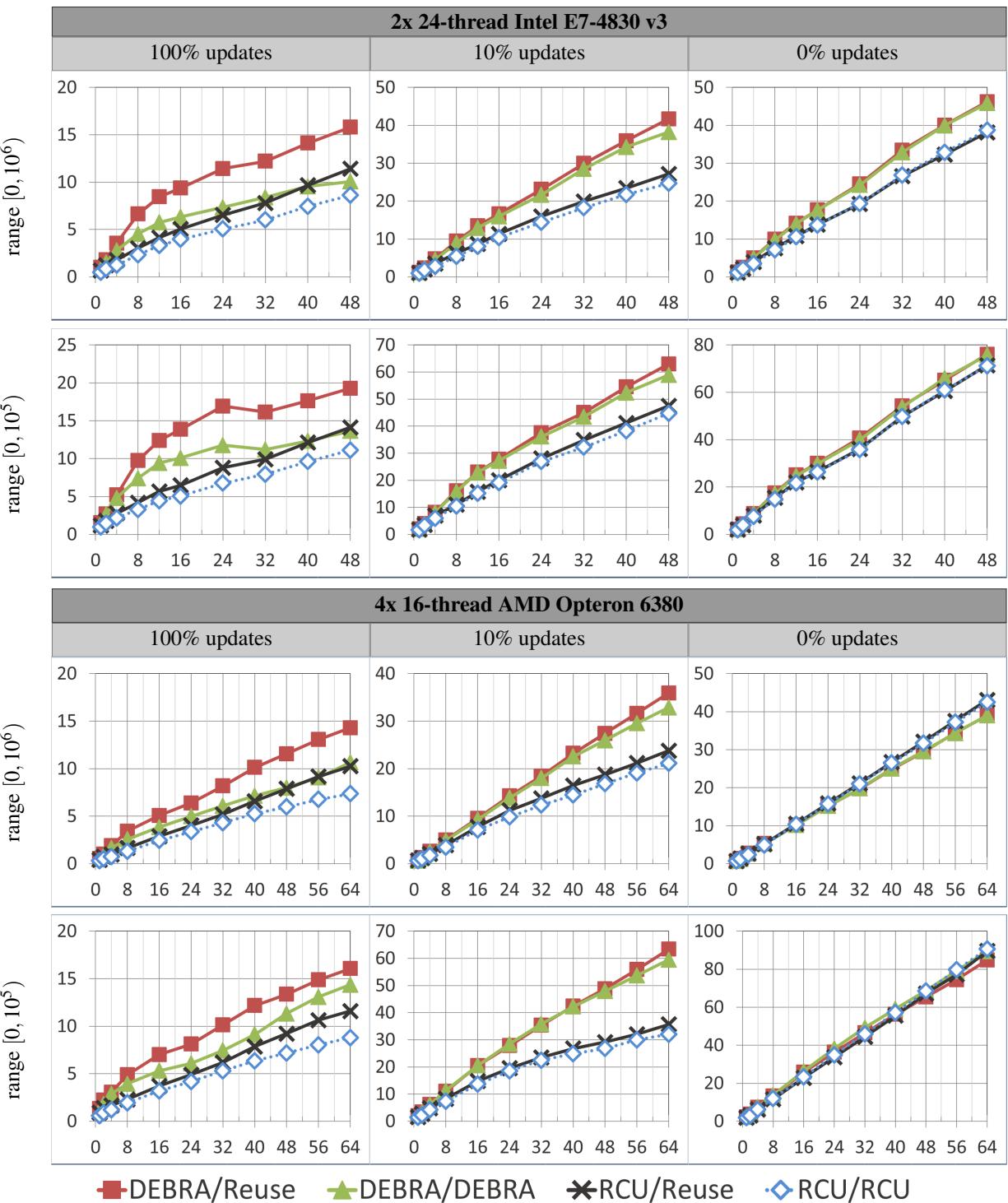


Figure 12.9: Results for a **BST microbenchmark**. The x-axis represents the number of concurrent threads. The y-axis represents operations per microsecond.

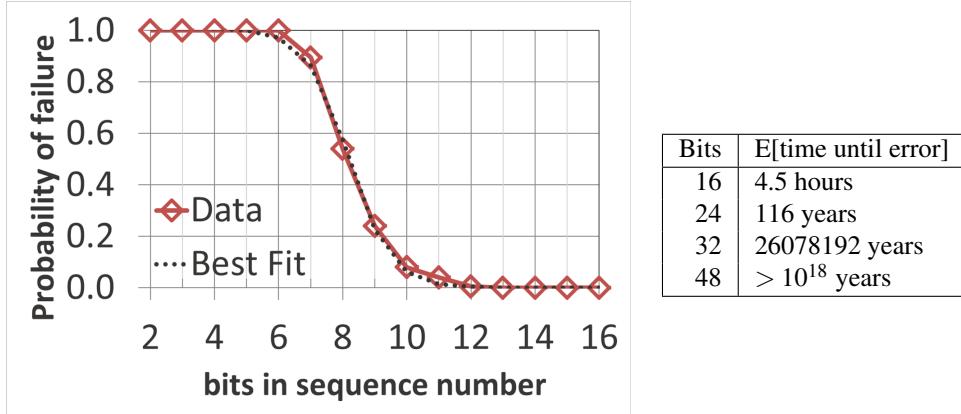


Figure 12.10: Experiment studying sequence number wraparound.

frequently experienced errors. However, for $B \geq 13$, we did not observe a single error in 200 trials (despite the fact that wraparound consistently occurred in every trial). For $B \geq 16$, trials were not sufficiently long for wraparound to consistently occur. The results appear in Figure 12.10.

As is common in physics when studying unknown functions, we make an educated guess that the distribution is sigmoidal, which means it is of the form $f(x) = a/(1 + e^{-b(x-c)})$ for constants a, b and c . We determined a sigmoidal curve of best fit from the data, obtaining the function $f(x) = 1/(1 + e^{1.53969(x-8.199181)})$, which is plotted as the *Best Fit* curve on the graph in Figure 12.10. As the graph shows, the error between the best fit curve and the measured data is extremely small. Although we do not have a justification for the shape of this distribution, we think it is worthwhile to put forth a hypothesis and study its consequences.

We used $f(x)$ to extrapolate on the data to estimate the expected time until an error occurs in this workload for several bit-widths that would be impractical to test experimentally. These extrapolations appear in the table on the right of Figure 12.10. They should be taken with a grain of salt, since the error in our estimation likely grows quickly with B . However, the extrapolations suggest that even $B = 32$ would be quite safe for this workload. To our knowledge, this kind of experimental exploration of the practicality of unbounded sequence numbers has not previously been done.

12.6 Related Work

Several papers have presented universal constructions or strong primitives for non-blocking algorithms in which operations create descriptors [72, 9, 2, 100, 62, 92, 74, 94, 13, 29]. A subset of these algorithms employ ad-hoc techniques for reusing descriptors [72, 9, 2, 100, 94, 92, 74]. The rest assume descriptors will be allocated for each operation and eventually reclaimed.

Most of the ad-hoc techniques for reusing descriptors have significant downsides. Some are complex and tightly integrated into the underlying algorithm, or rely on highly specific algorithmic properties (e.g., that descriptors contain only a single word). Others use synchronization primitives that atomically operate on large words, which are not available on modern systems, and are inefficient when implemented in software. Yet others introduce high space overhead (e.g., by attaching a sequence number to *every* memory word). Some techniques also incur significant runtime overhead (e.g., by invoking expensive synchronization primitives just to *read fields* of a descriptor). Furthermore, these techniques give, at best, a vague idea of how one might reuse descriptors for arbitrary algorithms, and it would

be difficult to determine how to use them in practice. Our work avoids all of these downsides, and provides a concrete approach for transforming a large class of algorithms.

Barnes [17] introduced a technique for producing non-blocking algorithms that can be more efficient (and sometimes simpler) than the universal constructions described above. With Barnes' technique, each operation creates a new descriptor. Creating a new descriptor for each operation allows his technique to avoid the ABA problem while remaining conceptually simple. Each operation conceptually locks each location it will modify by installing a pointer to its descriptor, and then performs its modifications and unlocks each location. Barnes' technique is the inspiration for the class WCA. Many algorithms have since been introduced using variants of this technique [62, 52, 13, 70, 113, 29, 30]. Several of these algorithms are quite efficient in practice despite the overhead of creating and reclaiming descriptors. Our technique can significantly improve the space and time overhead of such algorithms.

Recent work has identified ways to use hardware transactional memory (HTM) to reduce descriptor allocation [91]. Currently, HTM is supported only on recent Intel and IBM processors. Other architectures, such as AMD, SPARC and ARM have not yet developed HTM support. Thus, it is important to provide solutions for systems with no HTM support. Additionally, even with HTM support, our approach is useful. Current (and likely future) implementations of HTM offer no progress guarantees, so one must provide a lock-free fallback path to guarantee lock-free progress. The techniques in [91] accelerate the HTM-based *fast path*, but do nothing to reduce descriptor allocations on the fallback path. In some workloads, many operations run on the fallback path, so it is important for it to be efficient. Our work provides a way to accelerate the fallback path, and is orthogonal to work that optimizes the fast path.

The *long-lived renaming* (LLR) problem is related to our work (see [25] for a survey), but its solutions do not solve our problem. LLR provides processes with operations to *acquire* one unique resource from a pool of resources, and subsequently *release* it. One could imagine a scheme in which processes use LLR to reuse a small set of descriptors by invoking *acquire* instead of allocating a new descriptor, and eventually invoking *release*. Note, however, that a descriptor can safely be released only once it can no longer be accessed by any other process. Determining when it is safe to release a descriptor is as hard as performing general memory reclamation, and would also require delaying the release (and subsequent acquisition) of a descriptor (which would increase the number of descriptors needed). In contrast, our weak descriptors eliminate the need for memory reclamation, and allow immediate reuse.

12.7 Summary

We presented a novel technique for transforming algorithms that throw away descriptors into algorithms that reuse descriptors. Our experiments show that our transformation yields significant performance improvements for a lock-free k -CAS algorithm. Furthermore, our transformation reduces peak memory usage by nearly three orders of magnitude over the next best implementation.

We also applied our transformation to a lock-free implementation of *LLX* and *SCX*, and studied its performance by doing rigorous experiments on a lock-free binary search tree that uses *LLX* and *SCX*. These experiments demonstrated a significant performance advantage for our transformed algorithm in workloads that perform many updates. Our transformed *LLX* and *SCX* algorithm has the potential to accelerate many algorithms that use *LLX* and *SCX*.

We believe our transformation can be used to improve the performance and memory usage of many other algorithms that throw away descriptors. Moreover, we hope that our extended weak descriptor ADT will aid in the design of more efficient, complex algorithms, by allowing algorithm designers to benefit from the conceptual simplicity of throwing away descriptors without paying the practical costs of doing so.

Chapter 13

Accelerating the template with HTM

In this chapter, we study how the new hardware transactional memory (HTM) capabilities found in recent processors (e.g., by Intel and IBM) can be used to produce significantly faster implementations of the tree update template. By accelerating the tree update template, we also provide a way to accelerate all of the data structures that have been implemented with it. Since library data structures are reused many times, even minor performance improvements confer a large benefit.

HTM allows a programmer to run blocks of code in transactions, which either commit and take effect atomically, or abort and have no effect on shared memory. Although transactional memory was originally intended to *simplify* concurrent programming, researchers have since realized that HTM can also be used effectively to *improve the performance of existing concurrent code* [91, 118, 93]: Hardware transactions typically have very little overhead, so they can often be used to replace other, more expensive synchronization mechanisms. For example, instead of performing a sequence of CAS primitives, it may be faster to perform reads, if-statements and writes inside a transaction. Note that this represents a *non-standard use of HTM*: we are *not* interested in its ease of use, but, rather, in its ability to reduce synchronization costs.

Although hardware transactions are fast, it is surprisingly difficult to obtain the full performance benefit of HTM. Here, we consider Intel’s HTM, which is a *best-effort* implementation. This means it offers *no guarantee* that transactions will ever commit. Even in a single threaded system, a transaction can repeatedly abort because of internal buffer overflows, page faults, interrupts, and many other events. So, to guarantee progress, any code that uses HTM must also provide a *software fallback path* to be executed if a transaction fails. The design of the fallback path profoundly impacts the performance of HTM-based algorithms.

Allowing concurrency between two paths. Consider an operation O that is implemented using the tree update template. One natural way to use HTM to accelerate O is to use the original operation as a fallback path, and then obtain an HTM-based fast path by wrapping O in a transaction, and performing optimizations to improve performance [91]. We call this the **2-path concurrent** algorithm (*2-path con*). Since the fast path is just an optimized version of the fallback path, transactions on the fast path and fallback path can safely run concurrently. If a transaction aborts, it can either be retried on the fast path, or be executed on the fallback path. Unfortunately, supporting concurrency between the fast path and fallback path can add significant overhead on the fast path.

The first source of overhead is *instrumentation* on the fast path that manipulates the *meta-data* used by the fallback path to synchronize processes. For example, lock-free algorithms often create a *descriptor* for each update operation (so that processes can determine how to help one another make progress), and store pointers to these descriptors in Data-records, where they act as locks. The fast path must also manipulate these descriptors and pointers so that the fallback path can detect changes made by the fast path.

The second source of overhead comes from constraints imposed by algorithmic assumptions made on the fallback path. The tree update template implementation in [30] assumes that only child pointers can change, and all other fields of nodes, such as keys and values, are never changed (after a node is created). Changes to these other *immutable* fields must be made by replacing a node with a new copy that reflects the desired change. Because of this assumption on the fallback path, transactions on the fast path *cannot* directly change any field of a node other than its child pointers. This is because the fallback path has no mechanism to detect such a change (and may, for example, erroneously delete a node that is concurrently being modified by the fast path). Thus, just like the fallback path, the fast path must replace a node with a new copy to change one of its immutable fields, which can be much less efficient than changing the field directly.

Disallowing concurrency between two paths. To avoid the overheads described above, concurrency is often *disallowed* between the fast path and fallback path. The simplest example of this approach is a technique called **transactional lock elision** (TLE) [106, 107]. TLE is used to implement an operation by wrapping its sequential code in a transaction, and falling back to acquire a global lock after a certain number of transactional attempts. At the beginning of each transaction, a process reads the state of the global lock and aborts the transaction if the lock is held (to prevent inconsistencies that might arise because the fallback path is not atomic). Once a process begins executing on the fallback path, all concurrent transactions abort, and processes wait until the fallback path is empty before retrying their transactions.

If transactions never abort, then *TLE represents the best performance we can hope to achieve*, because the fallback path introduces almost no overhead and synchronization is performed entirely by hardware. Note, however, that TLE is not lock-free. Additionally, in workloads where operations periodically run on the fallback path, performance can be very poor.

As a toy example, consider a TLE implementation of a binary search tree, with a workload consisting of insertions, deletions and *range queries*. A range query returns all of the keys in a range $[lo, hi]$. Range queries access many memory locations, and cause frequent transactional aborts due to internal processor buffer overflows (capacity limits). Thus, range queries periodically run on the fallback path, where they can lead to numerous performance problems. Since the fallback path is sequential, range queries (or any other long-running operations) cause a severe *concurrency bottleneck*, because they prevent transactions from running on the fast path while they slowly complete, serially.

One way to mitigate this bottleneck is to replace the sequential fallback path in TLE with a lock-free algorithm, and replace the global lock with a fetch-and-increment object F that counts how many operations are running on the fallback path. Instead of aborting if the lock is held, transactions on the fast path abort if F is non-zero. We call this the **2-path non-concurrent** algorithm (*2-path con*). In this algorithm, if transactions on the fast path retry only a few times before moving to the fallback path, or do not wait between retries for the fallback path to become empty, then the lemming effect [48] can occur. (The lemming effect occurs when processes on the fast path rapidly fail and move to the fallback path, simply because other processes are on the fallback path.) This can cause the algorithm to run only as fast as the (much slower) fallback path. However, if transactions avoid the lemming effect by retrying many times before moving to the fallback path, and waiting between retries for the fallback path to become empty, then processes can spend most of their time *waiting*. The performance problems discussed up to this point are summarized on the left side of Figure 13.1.

The problem with two paths. In this paper, we study two different types of workloads: **light workloads**, in which transactions rarely run on the fallback path, and **heavy workloads**, in which transactions more frequently run on the fallback path. In light workloads, algorithms that allow concurrency between paths perform very poorly (due to high overhead) in comparison to algorithms that disallow concurrency. However, in heavy workloads, algorithms that disallow concurrency perform very poorly (since transactions on the fallback path prevent transactions from running on the fast path) in comparison to algorithms that allow concurrency between paths. Consequently, all two path algorithms have workloads that yield poor performance. Our experiments confirm this, showing surprisingly poor performance for two path algorithms in many cases.

Using three paths. We introduce a technique that simultaneously achieves high performance for both light and heavy workloads by using three paths: an HTM fast path, an HTM middle path and a non-transactional fallback path. (See the illustration on the right side of Figure 13.1.) Each operation begins on the fast path, and moves to the middle path

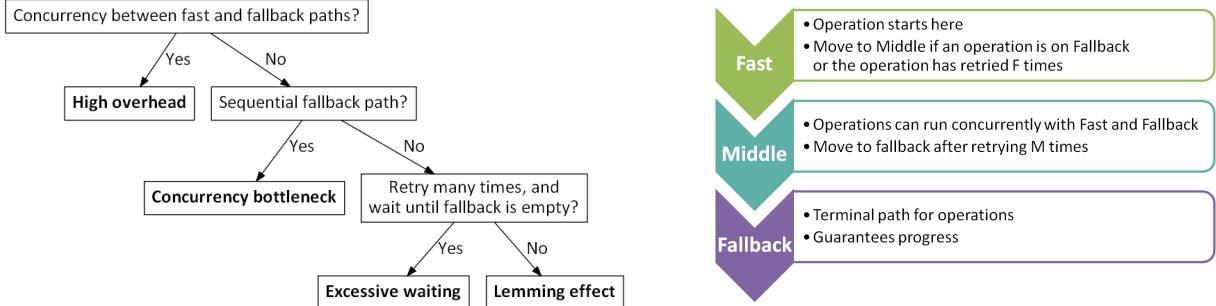


Figure 13.1: (Left) Performance problems affecting two-path algorithms. (Right) Using three execution paths.

after it retries F times. An operation on the middle path moves to the fallback path after retrying M times on the middle path. The fast path does not manipulate any synchronization meta-data used by the fallback path, so operations on the fast path and fallback path cannot run concurrently. Thus, whenever an operation is on the fallback path, all operations on the fast path move to the middle path. The middle path manipulates the synchronization meta-data used by the fallback path, so operations on the middle path and fallback path can run concurrently. Operations on the middle path can also run concurrently with operations on the fast path (since conflicts are resolved by the HTM system). We call this the **3-path** algorithm (*3-path*).

We briefly discuss why this approach avoids the performance problems described above. Since transactions on the fast path do not run concurrently with transactions on the fallback path, transactions on the fast path run with *no instrumentation overhead*. When a transaction is on the fallback path, transactions can freely execute on the middle path, *without waiting*. The *lemming effect* does not occur, since transactions do not have to move to the fallback path simply because a transaction is on the fallback path. Furthermore, we enable a high degree of concurrency, because the fast and middle paths can run concurrently, and the middle and fallback paths can run concurrently.

We performed experiments to evaluate our new template algorithms by comparing them with the original template algorithm. In order to compare the different template algorithms, we used each algorithm to implement two data structures: a binary search tree (BST) and a relaxed (a,b) -tree. We then ran microbenchmarks to compare the performance (operations per second) of the different implementations in both light and heavy workloads. The results show that our new template algorithms offer significant performance improvements. For example, on an Intel system with 72 concurrent processes, our best implementation of the relaxed (a,b) -tree outperformed the implementation using the original template algorithm by an average of 410% over all workloads.

Contributions

- We present four accelerated implementations of the tree update template that explore the design space for HTM-based implementations: *2-path con*, *TLE*, *2-path \overline{con}* , and *3-path*.
- We highlight the importance of studying both light and heavy workloads in the HTM setting. Each serves a distinct role in evaluating algorithms: light workloads demonstrate the potential of HTM to improve performance by reducing overhead, and heavy workloads capture the performance impact of interactions between different execution paths.
- We demonstrate the effectiveness of our approach by accelerating two different lock-free data structures: an unbalanced BST, and a relaxed (a,b) -tree. Experimental results show a significant performance advantage for our accelerated implementations.

<pre> 1 Private variable for process p: $attempts_p, tagseq_p$ 2 $SCX(V, R, fld, new)$ by process p 3 onAbort: ▷ jump here on transaction abort 4 if we jumped here after an explicit abort then return FALSE 5 if $attempts_p < AttemptLimit$ then 6 $attempts_p := attempts_p + 1$ 7 $retval := SCX_{HTM}(V, R, fld, new)$ ▷ Fast 8 else 9 $retval := SCX_O(V, R, fld, new)$ ▷ Fallback 10 if $retval$ then $attempts_p := 0$ 11 return $retval$ </pre>	<pre> 12 $SCX_{HTM}(V, R, fld, new)$ by process p 13 Let $infoFields$ be a pointer to a table in p's private memory 14 containing, 15 for each r in V, the value of $r.info$ read by p's last $LLX(r)$ 16 Let old be the value for fld returned by p's last $LLX(r)$ 17 Begin hardware transaction 18 $tagseq_p := tagseq_p + 2^{\lceil \log n \rceil}$ 19 for each $r \in V$ do 20 Let $r.info$ be the pointer indexed by r in $infoFields$ 21 if $r.info \neq rinfo$ then Abort hardware transaction 22 (explicitly) 23 for each $r \in V$ do $r.info := tagseq_p$ 24 for each $r \in R$ do $r.marked := \text{TRUE}$ write new to the field pointed to by fld Commit hardware transaction return TRUE </pre>
--	---

Figure 13.2: HTM-based implementation of SCX .

The remainder of this chapter is structured as follows. We describe an HTM-based implementation of LLX and SCX in Section 13.1. In Section 13.2, we describe our four template implementations, and argue correctness and progress. In Section 13.3, we describe two data structures that we use in our experiments. Experimental results are presented in Section 13.4. In Section 13.6, we describe a way to reclaim memory more efficiently for 3 -path algorithms. Related work is surveyed in Section 13.7. In Section 13.8, we describe how our approach could be used to accelerate data structures that use the read-copy-update (RCU) or k -compare-and-swap primitives. Finally, we summarize in Section 13.9.

13.1 HTM-based LLX and SCX

In this section, we describe an HTM-based implementation of LLX and SCX . This implementation is used by our first accelerated template implementation, 2 -path *con*, which is described in Section 13.2.

In the following, we use SCX_O and LLX_O to refer to the original lock-free implementation of LLX and SCX . We give an implementation of SCX that uses an HTM-based fast path called SCX_{HTM} , and SCX_O as its fallback path. Hardware transactions are instrumented so they can run concurrently with processes executing SCX_O . This algorithm guarantees lock-freedom and achieves a high degree of concurrency. Pseudocode appears in Figure 13.2. At a high level, an SCX_{HTM} by a process p starts a transaction, then attempts to perform a highly optimized version of SCX_O . Each time a transaction executed by p aborts, control jumps to the *onAbort* label, at the beginning of the SCX procedure. If a process *explicitly* aborts a transaction at line 19, then SCX returns FALSE at line 4. Each process has a budget $AttemptLimit$ that specifies how many times it will attempt hardware transactions before it will fall back to executing SCX_O .

In SCX_O , SCX -records are used (1) to facilitate helping, and (2) to lock Data-records and detect changes to them. In particular, SCX_O guarantees the following property. **P1:** between any two changes to (the user-defined fields of) a Data-record u , a new SCX -record pointer is stored in $u.info$. However, SCX_{HTM} does not create SCX -records. In a transactional setting, helping causes unnecessary aborts, since executing a transaction that performs the same work as a running transaction will cause at least one (and probably both) to abort. Helping in transactions is also *not necessary* to guarantee progress, since progress is guaranteed by the fallback path. So, to preserve property P1, we give each

process p a *tagged sequence number* $tseq_p$ that contains the process name, a sequence number, and a *tag* bit. The *tag* bit is the least significant bit. On modern systems where pointers are word aligned, the least significant bit in a pointer is always zero. Thus, the tag bit allows a process to distinguish between a tagged sequence number and a pointer. In SCX_{HTM} , instead of having p create a new SCX -record and store pointers to it in Data-records to lock them, p increments its sequence number in $tseq_p$ and stores $tseq_p$ in Data-records. Since no writes performed by a transaction T can be seen until it commits, it never actually needs to hold any locks. Thus, every value of $tseq_p$ stored in a Data-record represents an unlocked value, and writing $tseq_p$ represents p locking and immediately unlocking a node.

After storing $tseq_p$ in each $r \in V$, SCX_{HTM} finalizes each $r \in R$ by setting $r.marked := \text{TRUE}$ (mimicking the behaviour of SCX_O). Then, it stores *new* in the field pointed to by *fld*, and commits. Note that eliminating the creation of SCX -records on the fast path also eliminates the need to *reclaim* any created SCX -records, which further reduces overhead.

The SCX_{HTM} algorithm also necessitates a small change to LLX_O , to handle tagged sequence numbers. An invocation of $LLX_O(r)$ reads a pointer *rinfo* to an SCX -record, follows *rinfo* to read one of its fields, and uses the value it reads to determine whether r is locked. However, *rinfo* may now contain a tagged sequence number, instead of a pointer to an SCX -record. So, in our modified algorithm, which we call LLX_{HTM} , before a process tries to follow *rinfo*, it first checks whether *rinfo* is a tagged sequence number, and, if so, behaves as if r is unlocked. The code for LLX_{HTM} appears in Figure 13.6.

13.1.1 Correctness and Progress

In this section, we prove correctness and progress for our HTM-based implementation of LLX and SCX . We do this by showing that one can start with LLX_O and SCX_O , and obtain our HTM-based implementation by applying a sequence of transformations. Intuitively, these transformations preserve the semantics of SCX and maintain backwards compatibility with SCX_O so that the transformed versions can be run concurrently with invocations of SCX_O . More formally, for each execution of a transformed algorithm, there is an execution of the original algorithm in which: the same operations are performed, they are linearized in the same order, and they return the same results. For each transformation, we sketch the correctness and progress argument, since the transformations are simple and a formal proof would be overly pedantic.

Adding transactions For the first transformation, we replaced the invocation of *Help* in SCX_O with the body of the *Help* function, and wrapped the code in a transaction. Since the fast path simply executes the fallback path algorithm in a transaction, the correctness of the resulting algorithm is immediate from the correctness of the original LLX and SCX algorithm.

We also observe that it is not necessary to commit a transaction that sets the *state* of its SCX -record to *Aborted* and returns *FALSE*. The only effect that committing such a transaction would have on shared memory is changing some of the *info* fields of Data-records in its V sequence to point to its SCX -record. In SCX_O , *info* fields serve two purposes. First, they provide pointers to an SCX -record while its SCX is in progress (so it can be helped). Second, they act as locks that grant exclusive access to an SCX_O , and allow an invocation of SCX_O to determine whether any user-defined fields of a Data-record r have changed since its linked $LLX(r)$ (using property P1). However, since the effects of a transaction are not visible until it has already committed, a transaction no longer needs help by the time it modified

```

1  SCX1(V, R, fld, new) by process p
2  Let infoFields be a pointer to a table in p's private memory containing,
   for each r in V, the value of r.info read by p's last LLX(r)
3  Let old be the value for fld returned by p's last LLX(r)
4  Begin hardware transaction
5  scxPtr := pointer to new SCX-record(V, R, fld, new, old, InProgress, FALSE, infoFields)
6  ▷ Freeze all Data-records in scxPtr.V to protect their mutable fields from being changed by other SCXs
7  for each r in scxPtr.V enumerated in order do
8    Let rinfo be the pointer indexed by r in scxPtr.infoFields
9    if not CAS(r.info, rinfo, scxPtr) then                                ▷ freezing CAS
10   if r.info ≠ scxPtr then
11     ▷ Could not freeze r because it is frozen for another SCX
12     if scxPtr.allFrozen = TRUE then                                         ▷ frozen check step
13       ▷ the SCX has already completed successfully
14       Commit hardware transaction
15       return TRUE
16     else Abort hardware transaction (explicitly)
17   ▷ Finished freezing Data-records (Assert: state ∈ {InProgress, Committed})
18   scxPtr.allFrozen := TRUE                                              ▷ frozen step
19   for each r ∈ scxPtr.R do r.marked := TRUE                            ▷ mark step
20   CAS(scxPtr.fld, scxPtr.old, scxPtr.new)                                ▷ update CAS
21   ▷ Finalize all r in R, and unfreeze all r in V that are not in R
22   scxPtr.state := Committed                                              ▷ commit step
23   Commit hardware transaction
24   return TRUE

```

Figure 13.3: HTM-based SCX: after adding transactions.

any *info* field. And, since an SCX_O that sets the *state* of its SCX-record to Aborted does not change any user-defined field of a Data-record, these changes to *info* fields are not needed to preserve property P1. The only consequence of changing these *info* fields is that other invocations of SCX_O might needlessly fail and return FALSE, as well. So, instead of setting *state* = Aborted and committing, we *explicitly abort* the transaction and return FALSE. Figure 13.3 shows the result of this transformation: SCX_1 . (Note that aborting transactions does not affect correctness—only progress.)

1	Private variable for process p: attempts _p
2	SCX(V, R, fld, new) by process p
3	onAbort:
4	if we jumped here after an explicit abort in the code then return FALSE
5	if attempts _p < AttemptLimit then
6	attempts _p := attempts _p + 1
7	retval := SCX ₁ (V, R, fld, new)
8	else
9	retval := SCX _O (V, R, fld, new)
10	if retval then attempts _p := 0
11	return retval

Figure 13.4: How the HTM-based SCX_1 is used to provide lock-free SCX.

Of course, we must provide a fallback code path in order to guarantee progress. Figure 13.4 shows how SCX_1 (the fast path) and SCX_O (the fallback path) are used together to implement lock-free SCX. In order to decide when each code path should be executed, we give each process p a private variable $attempts_p$ that contains the number of times p has attempted a hardware transaction since it last performed an SCX_1 or SCX_O that succeeded (i.e., returned TRUE). The SCX procedure checks whether $attempts_p$ is less than a (positive) threshold $AttemptLimit$. If so, p increments $attempts_p$ and invokes SCX_1 to execute a transaction on the fast path. If not, p invokes SCX_O (to guarantee progress).

Whenever p returns TRUE from an invocation of SCX_1 or SCX_O , it resets its budget $attempts_p$ to zero, so it will execute on the fast path in its next SCX . Each time a transaction executed by p aborts, control jumps to the `onAbort` label, at the beginning of the SCX procedure. If a process explicitly aborts a transaction it is executing (at line 16 in SCX_1), then control jumps to the `onAbort` label, and the SCX returns FALSE at the next line.

Progress We briefly recall the progress property for LLX and SCX . Specifying a progress guarantee for LLX and SCX operations is subtle, because if processes repeatedly perform LLX on Data-records that have been finalized, or repeatedly perform failed LLX s, then they may never be able to invoke SCX . In particular, it is not sufficient to simply prove that LLX s return snapshots infinitely often, since *all* of the LLX s in a sequence must return snapshots before a process can invoke SCX . To simplify the progress guarantee for LLX and SCX , we make a definition. An SCX -*Update* algorithm is one that performs LLX s on a sequence V of Data-records and invokes $SCX(V, R, fld, new)$ if they all return snapshots. The progress guarantee in Chapter 3 was then stated as follows.

PROG: Suppose that (a) there is always some non-finalized Data-record reachable by following pointers from an entry point, (b) for each Data-record r , each process performs finitely many invocations of $LLX(r)$ that return FINALIZED, and (c) processes perform infinitely many executions of SCX -*Update* algorithms. Then, infinitely many invocations of SCX succeed.

LLX_O and SCX_O satisfy PROG. We argue that PROG is satisfied by the implementation of LLX and SCX in Figure 13.4. To obtain a contradiction, suppose the antecedent of PROG holds, but only finitely many invocations of SCX return TRUE. Then, after some time t , no invocation of SCX returns TRUE.

Case 1: Suppose processes take infinitely many steps in transactions. By inspection of the code, each transaction is wait-free, and SCX returns TRUE immediately after a transaction commits. Since no transaction commits after t , there must be infinitely many aborts. However, each process can perform at most $AttemptLimit$ aborts since the last time it performed an invocation of SCX that returned TRUE. So, only finitely many aborts can occur after t —a contradiction.

Case 2: Suppose processes take only finitely many steps in transactions. Then, processes take only finitely many steps in SCX_1 . It follows that, after some time t' , no process takes a step in SCX_1 . Therefore, in the suffix of the execution after t' , processes only take steps in SCX_O and LLX_O . However, since LLX_O and SCX_O satisfy PROG, infinitely many invocations of SCX must succeed after t' , which is a contradiction.

Eliminating *most* accesses to fields of SCX -records created on the fast path In LLX_O and SCX_O , helping is needed to guarantee progress, because otherwise, an invocation of SCX_O that crashes while one or more Data-records are frozen for it could cause every invocation of LLX_O to return FAIL (which, in turn, could prevent processes from performing the necessary linked invocations of LLX_O to invoke SCX_O). However, as we mentioned above, since transactions are atomic, a process cannot see any of their writes (including the contents of any SCX -record they create and publish pointers to) until they have committed, at which point they no longer need help. Thus, it is not necessary to help transactions in SCX_1 .¹

In fact, it is easy to see that processes will not help any SCX -record created by a transaction in SCX_1 . Observe that each transaction in SCX_1 sets the *state* of its SCX -record to Committed before committing. Consequently, if an invocation of LLX_O reads $r.info$ and obtains a pointer $rinfo$ to an SCX -record created by a transaction in SCX_1 , then $rinfo$ has *state* Committed. Therefore, by inspection of the code, LLX_O will not invoke $Help(rinfo)$.

¹In fact, helping transactions would be *actively harmful*, since performing the same modifications to shared memory as an in-flight transaction will cause it to abort. This leads to very poor performance, in practice.

```

1 Private variable for process  $p$ :  $\text{attempts}_p$ 
2  $\text{SCX}_2(V, R, \text{fld}, \text{new})$  by process  $p$ 
3 Let  $\text{infoFields}$  be a pointer to a table in  $p$ 's private memory containing,
   for each  $r$  in  $V$ , the value of  $r.\text{info}$  read by  $p$ 's last  $\text{LLX}(r)$ 
4 Let  $\text{old}$  be the value for  $\text{fld}$  returned by  $p$ 's last  $\text{LLX}(r)$ 
5 Begin hardware transaction
6  $\text{scxPtr} :=$  pointer to new SCX-record( $-,-,-,-,-$ , InProgress,  $-,-$ )
7 ▷ Freeze all Data-records in  $V$  to protect their mutable fields from being changed by other SCXs
8 for each  $r$  in  $V$  enumerated in order do
9   Let  $r.\text{info}$  be the pointer indexed by  $r$  in  $\text{infoFields}$ 
10  if not  $\text{CAS}(r.\text{info}, r.\text{info}, \text{scxPtr})$  then ▷ freezing CAS
11    if  $r.\text{info} \neq \text{scxPtr}$  then Abort hardware transaction (explicitly)
12  ▷ Finished freezing Data-records
13  for each  $r \in R$  do  $r.\text{marked} := \text{TRUE}$  ▷ Finalize each  $r \in R$ 
14   $\text{CAS}(\text{fld}, \text{old}, \text{new})$  ▷ mark step
15   $\text{scxPtr.state} := \text{Committed}$  ▷ update CAS
16  Commit hardware transaction ▷ commit step
17  return  $\text{TRUE}$ 

```

Figure 13.5: HTM-based SCX: after eliminating **most** accesses to fields of SCX-records created on the fast path.

Since LLX_O never invokes $\text{Help}(r.\text{info})$ for any $r.\text{info}$ created by a transaction in SCX_1 , most fields of an SCX-record created by a transaction are accessed only by the process that created the SCX-record. The only field that is accessed by other processes is the state field (which is accessed in LLX_O). Therefore, it suffices for a transaction in SCX_1 to initialize only the state field of its SCX-record. As we will see, any accesses to the other fields can simply be eliminated or replaced with locally available information.

Using this knowledge, we transform SCX_1 in Figure 13.3 into a new procedure called SCX_2 in Figure 13.5. First, instead of initializing the entire SCX-record when we create a new SCX-record at line 5 in SCX_1 , we initialize only the state field. We then change any steps that read fields of the SCX-record (lines 7, 8, 12, 19 and 20 in SCX_1) to use locally available information, instead.

Next, we eliminate the *frozen step* at line 18 in SCX_1 , which changes the allFrozen field of the SCX-record. Recall that allFrozen is used by SCX_O to prevent helpers from making conflicting changes to the state field of its SCX-record. When a *freezing CAS* fails in an invocation S of SCX_O (at line 26 of Help in Figure 3.4), it indicates that either S will fail due to contention, or another process had already helped S to complete successfully. The allFrozen bit allows a process to distinguish between these two cases. Specifically, it is proved in [29] that a process will see $\text{allFrozen} = \text{TRUE}$ at line 26 of Help if and only if another process already helped S complete and set $\text{allFrozen} := \text{TRUE}$. However, since we have argued that processes never help transactions (and, in fact, no other process can even access the SCX-record until the transaction that created it has committed), allFrozen is always FALSE at the corresponding step (line 12) in SCX_1 . This observation allows us to eliminate the entire *if* branch at line 12 in SCX_1 .

Clearly, this transformation preserves PROG. Note that SCX_2 (and each of the subsequent transformed variants) is used in the same way as SCX_1 : Simply replace SCX_1 in Figure 13.4 with SCX_2 .

Completely eliminating accesses to fields of SCX-records created on the fast path We now describe a transformation that completely eliminates all accesses to the state fields of SCX-records created by transactions in SCX_2 (i.e., the last remaining accesses by transactions to fields of SCX-records).

We transform SCX_2 into a new procedure SCX_3 , which appears in Figure 13.6. First, the commit step in SCX_2 is eliminated. Whereas in SCX_2 , we stored a pointer to the SCX-record in $r.\text{info}$ for each $r \in V$ at line 10, we store a

```

1 Private variable for process  $p$ :  $attempts_p$ 
2  $LLX_{HTM}(r)$  by process  $p$  ▷ Precondition:  $r \neq \text{NIL}$ 
3    $marked_1 := r.\text{marked}$                                      ▷ order of lines 3–6 matters
4    $rinfo := r.\text{info}$ 
5   *  $state := (rinfo \& 1) ? \text{Committed} : rinfo.\text{state}$            ▷ if  $rinfo$  is tagged, take  $state$  to be Committed
6    $marked_2 := r.\text{marked}$ 
7   if  $state = \text{Aborted}$  or ( $state = \text{Committed}$  and not  $marked_2$ ) then          ▷ if  $r$  was not frozen at line 5
8     read  $r.m_1, \dots, r.m_y$  and record the values in local variables  $m_1, \dots, m_y$ 
9     if  $r.info = rinfo$  then                                         ▷ if  $r.info$  points to the same SCX-record as on line 4
10       store  $\langle r, rinfo, \langle m_1, \dots, m_y \rangle \rangle$  in  $p$ 's local table
11       return  $\langle m_1, \dots, m_y \rangle$ 
12   if  $state = \text{InProgress}$  then  $\text{Help}(rinfo)$ 
13   if  $marked_1$  then
14     return  $\text{FINALIZED}$ 
15   else
16     return  $\text{FAIL}$ 

17  $SCX_3(V, R, fld, new)$  by process  $p$ 
18 Let  $infoFields$  be a pointer to a table in  $p$ 's private memory containing,
19   for each  $r$  in  $V$ , the value of  $r.info$  read by  $p$ 's last  $LLX(r)$ 
20 Let  $old$  be the value for  $fld$  returned by  $p$ 's last  $LLX(r)$ 
21 Begin hardware transaction
22  $scxPtr :=$  pointer to new SCX-record( $-, -, -, -, -, -, -, -$ )
23 ▷ Freeze all Data-records in  $V$  to protect their mutable fields from being changed by other SCXs
24 for each  $r$  in  $V$  enumerated in order do
25   Let  $rinfo$  be the pointer indexed by  $r$  in  $infoFields$ 
26   if not  $\text{CAS}(r.info, rinfo, (scxPtr \& 1))$  then                                ▷ freezing CAS
27     if  $r.info \neq (scxPtr \& 1)$  then Abort hardware transaction (explicitly)
28   ▷ Finished freezing Data-records
29   for each  $r \in R$  do  $r.\text{marked} := \text{TRUE}$     ▷ Finalize each  $r \in R$            ▷ mark step
30    $\text{CAS}(fld, old, new)$                                          ▷ update CAS
31   Commit hardware transaction
32   return  $\text{TRUE}$ 

```

Figure 13.6: HTM-based SCX: after completely eliminating accesses to fields of SCX-records created on the fast path.

tagged pointer to the SCX-record at line 25 in SCX_3 . A tagged pointer is simply a pointer that has its least significant bit set to one. Note that, on modern systems where pointers are word aligned, the least significant bit in a pointer to an SCX-record will be zero. Thus, the least significant bit in a tagged pointer allows processes to distinguish between a tagged pointer (which is stored in $r.info$ by a transaction) from a regular pointer (which is stored in $r.info$ by an invocation of SCX_O). Line 26 in SCX_3 is also updated to check for a tagged pointer in $r.info$.

In order to deal with tagged pointers, we transform LLX_O into new procedure called LLX_{HTM} , that is used instead of LLX_O from here on. Any time an invocation of LLX_O would follow a pointer that was read from an $info$ field $r.info$, LLX_{HTM} first checks whether the value $rinfo$ read from the $info$ field is a pointer or a tagged pointer. If it is a pointer, then LLX_{HTM} proceeds exactly as in LLX_O . However, if $rinfo$ is a tagged pointer, then LLX_{HTM} proceeds as if it had seen an SCX-record with $state$ Committed (i.e., whose SCX has already returned TRUE). We explain why this is correct. If $rinfo$ contains a tagged pointer, then it was written by a transaction T that committed (since it changed shared memory) at line 30 in SCX_3 , just before returning TRUE. Observe that, in SCX_2 , the $state$ of the SCX-record is set to Committed just before TRUE is returned. In other words, if not for this transformation, T would have set the $state$ of its SCX-record to Committed. So, clearly it is correct to treat $rinfo$ as if it were an SCX-record with $state = \text{Committed}$.

Since this transformation simply changes the *representation* of an SCX-record D with $state = \text{Committed}$ that is

created by a transaction (and does not change how the algorithm behaves when it encounters D), it preserves PROG.

Eliminating the creation of SCX-records on the fast path Since transactions in SCX_3 are not helped, we would like to eliminate the *creation* of SCX-records in transactions, altogether. However, since SCX-records are used as part of the *freezing* mechanism in SCX_O on the fallback path, we cannot simply eliminate the steps that freeze Data-records, or else transactions on the fast path will not synchronize with SCX_O operations on the fallback path. Consider an invocation S of SCX_O by a process p that creates an SCX-record D , and an invocation L of $LLX(r)$ linked to S . When S uses CAS to freeze r (by changing $r.info$ from the value seen by L to D), it interprets the success of the CAS to mean that r has not changed since L (relying on property P1). If a transaction in SCX_3 changes r without changing $r.info$ (to a new value that has never before appeared in $r.info$), then it would violate P1, rendering this interpretation invalid. Thus, transactions in $SCX_3(V, R, fld, new)$ must change $r.info$ to a new value, for each $r \in V$.

We transform SCX_3 into a new procedure SCX_r , which appears in Figure 13.7. We now explain what a transaction T in an invocation S of SCX_4 by a process p does instead of creating an SCX-record and using it to freeze Data-records. We give each process p a *tagged sequence number* $tseq_p$, which consists of three bit fields: a tag-bit, a process name, and a sequence number. The tag-bit, which is the least significant bit, is always one. This tag-bit distinguishes tagged sequence numbers from pointers to SCX-records (similar to tagged pointers, above). The process name field of $tseq_p$ contains p . The sequence number is a non-negative integer that is initially zero. Instead of creating a new SCX-record (at line 21 in SCX_3), S increments the sequence number field of $tseq_p$. Then, instead of storing a pointer to an SCX-record in $r.info$ for each $r \in V$ (at line 25 in SCX_3), T stores $tseq_p$. (Line 26 is also changed accordingly.) The combination of the process name and sequence number bit fields ensure that whenever T stores $tseq_p$ in an $info$ field, it is storing a value that has never previously been contained in that field.²

```

1 Private variable for process  $p$ :  $tseq_p$ 
2  $SCX_4(V, R, fld, new)$  by process  $p$ 
3 Let  $infoFields$  be a pointer to a table in  $p$ 's private memory containing,
   for each  $r$  in  $V$ , the value of  $r.info$  read by  $p$ 's last  $LLX(r)$ 
4 Let  $old$  be the value for  $fld$  returned by  $p$ 's last  $LLX(r)$ 
5 Begin hardware transaction
6  $tseq_p := tseq_p + 2^{\lceil \log n \rceil}$  ▷ increment  $p$ 's tagged sequence number
7 ▷ Freeze all Data-records in  $V$  to protect their mutable fields from being changed by other SCXs
8 for each  $r$  in  $V$  enumerated in order do
9   Let  $r.info$  be the pointer indexed by  $r$  in  $infoFields$ 
10  if not  $CAS(r.info, r.info, tseq_p)$  then ▷ freezing CAS
11    if  $r.info \neq tseq_p$  then Abort hardware transaction (explicitly)
12  ▷ Finished freezing Data-records
13  ▷ Finalize each  $r \in R$ , update  $fld$ , and unfreeze all  $r \in (V \setminus R)$ 
14  for each  $r \in R$  do  $r.marked := \text{TRUE}$  ▷ mark step
15   $CAS(fld, old, new)$  ▷ update CAS
16  Commit hardware transaction
17  return  $\text{TRUE}$ 

```

Figure 13.7: HTM-based SCX: after eliminating SCX-record creation on the fast path.

²Technically, with a finite word size it is possible for a sequence number to overflow and wrap around, potentially causing P1 to be violated. On modern systems with a 64-bit word size, we suggest representing a tagged sequence number using 1 tag-bit, 15 bits for the process name (allowing up to 32,768 concurrent processes) and 48 bits for the sequence number. In order for a sequence number to experience wraparound, a *single process* must then perform 2^{48} operations. According to experimental measurements for several common data structures on high performance systems, this would take at least a decade of continuous updates. Moreover, if wraparound is still a concern, one can replace the freezing CAS steps in SCX_O with double-wide CAS instructions (available on all modern systems) which atomically operate on 128-bits, making wraparound virtually impossible.

Observe that LLX_{HTM} does not require any further modification to work with tagged sequence numbers, since it distinguishes between tagged sequence numbers and SCX-records using the tag-bit (the exact same way it distinguished between tagged pointers and pointers to SCX-records). Moreover, it remains correct to treat tagged sequence numbers as if they are SCX-records with *state* Committed (for the same reason it was correct to treat tagged pointers that way). Progress is preserved for the same reason as it was in the previous transformation: we are simply changing the *representation* of SCX-records with *state* = Committed that are created by transactions.

Note that this transformation eliminates not only the *creation* of SCX-records, but also the need to *reclaim* those SCX-records. Thus, it can lead to significant performance improvements.

```

1 Private variable for process  $p$ :  $tseq_p$ 
2  $SCX_5(V, R, fld, new)$  by process  $p$ 
3   Let  $infoFields$  be a pointer to a table in  $p$ 's private memory containing,
4     for each  $r$  in  $V$ , the value of  $r.info$  read by  $p$ 's last  $LLX(r)$ 
5   Let  $old$  be the value for  $fld$  returned by  $p$ 's last  $LLX(r)$ 
6   Begin hardware transaction
7    $tseq_p := tseq_p + 2^{\lceil \log n \rceil}$  ▷ increment  $p$ 's tagged sequence number
8   ▷ Freeze all Data-records in  $V$  to protect their mutable fields from being changed by other SCXs
9   for each  $r$  in  $V$  enumerated in order do
10    Let  $r.info$  be the pointer indexed by  $r$  in  $infoFields$ 
11    if  $r.info = rinfo$  then  $r.info := tseq_p$ 
12    else Abort hardware transaction (explicitly)
13   ▷ Finished freezing Data-records
14   ▷ Finalize each  $r \in R$ , update  $fld$ , and unfreeze all  $r \in (V \setminus R)$ 
15   for each  $r \in R$  do  $r.marked := \text{TRUE}$  ▷ mark step
16   if  $fld = old$  then  $fld := new$ 
17   Commit hardware transaction
     return  $\text{TRUE}$ 
```

Figure 13.8: HTM-based SCX: after replacing CAS with sequential code and optimizing.

Simple optimizations Since any code executed inside a transaction is atomic, we are free to replace atomic synchronization primitives inside a transaction with sequential code, and reorder the transaction's steps in any way that does not change its sequential behaviour. We now describe how to transform SCX_4 by performing two simple optimizations.

For the first optimization, we replace each invocation of $CAS(x, o, n)$ with sequential code: **if** $x = 0$ **then** $x := n$, $result := \text{TRUE}$ **else** $result := \text{FALSE}$. If the CAS is part of a condition for an if-statement, then we execute this code just before the if-statement, and replace the invocation of CAS with $result$. We then eliminate any *dead code* that cannot be executed. Figure 13.8 shows the transformed procedure, SCX_5 .

More concretely, in place of the CAS at line 10 in SCX_4 , we do the following. First, we check whether $r.info = rinfo$. If so, we set $r.info := tseq_p$ and continue to the next iteration of the loop. Suppose not. If we were naively transforming the code, then the next step would be to check whether $r.info$ contains $tseq_p$. However, p is the only process that can write $tseq_p$, and it only writes $tseq_p$ just before continuing to the next iteration. Thus, $r.info$ cannot possibly contain $tseq_p$ in this case, which makes it unnecessary to check whether $r.info = tseq_p$. Therefore, we execute the *else*-case, and explicitly abort the transaction. Observe that, if SCX_5 is used to replace SCX_1 in Figure 13.4, then this explicit abort will cause SCX to return FALSE (right after it jumps to the *onAbort* label). In place of the CAS at line 15 in SCX_4 , we can simply check whether fld contains old and, if so, write new into fld .

In fact, it is not necessary to check whether fld contains old , because the transaction will have aborted if fld was changed after old was read from it. We explain why. Let S be an invocation of SCX_5 (in Figure 13.8) by a

process p , and let r be the Data-record that contains fld . Suppose S executes line 15 in SCX_5 , where it checks whether $fld = old$. Before invoking S , p performs an invocation L of $LLX(r)$ linked to S . Subsequently, p reads old while performing S . After that, p freezes r while performing S . If r changes after L , and before p executes line 10, then p will see $r.info \neq rinfo$ when it executes line 10 (by property P1, which has been preserved by our transformations). Consequently, p will fail to freeze r , and S will perform an explicit abort and return FALSE, so it will *not* reach line 15, which contradicts our assumption (so this case is impossible). On the other hand, if r changes after p executes line 10, and before p executes line 15, then the transaction will abort due to a data conflict (detected by the HTM system). Therefore, when p executes line 15, fld must contain old .

For the second optimization, we split the loop in Figure 13.8 into two. The first loop contains all of the steps that check whether $r.info = rinfo$, and the second loop contains all of the steps that set $r.info := tseq_p$. This way, all of the writes to $r.info$ occur after all of the reads and if-statements. The advantage of delaying writes for as long as possible in a transaction is that it reduces the probability of the transaction causing other transactions to abort. As a minor point, whereas the loop in SCX_O iterated over the elements of the sequence V in a particular order to guarantee progress, it is not necessary to do so here, since progress is guaranteed by the fallback path, not the fast path. Clearly, this transformation does not affect correctness or progress.

13.2 Accelerated template implementations

13.2.1 The 2-path *con* algorithm

We now use our HTM-based LLX and SCX to obtain an HTM-based implementation of a template operation O . The fallback path for O is simply a lock-free implementation of O using LLX_O and SCX_O . The fast path for O starts a transaction, then performs the same code as the fallback path, except that it uses the HTM-based LLX and SCX . Since the *entire operation* is performed inside a transaction, we can optimize the invocations of SCX_{HTM} that are performed by O as follows. Lines 15 and 23 can be eliminated, since SCX_{HTM} is already running inside a large transaction. Additionally, lines 17-19 can be eliminated, since the transaction will abort due to a data conflict if $r.info$ changes after it is read in the (preceding) linked invocation of $LLX(r)$, and before the transaction commits. The proof of correctness and progress for 2-path *con* follows immediately from the proof of the original template and the proof of the HTM-based LLX and SCX implementation.

Note that it is not necessary to perform the entire operation in a single transaction. In Section 13.5, we describe a modification that allows a read-only *searching* prefix of the operation to be performed before the transaction begins.

13.2.2 The TLE algorithm

To obtain a TLE implementation of an operation O , we simply take *sequential code* for O and wrap it in a transaction on the fast path. The fallback path acquires and releases a global lock instead of starting and committing a transaction, but otherwise executes the same code as the fast path. To prevent the fast path and fallback path from running concurrently, transactions on the fast path start by reading the lock state and aborting if it is held. An operation attempts to run on the fast path up to *AttemptLimit* times (waiting for the lock to be free before each attempt) before resorting to the fallback path. The correctness of TLE is trivial. Note, however, that TLE only satisfies deadlock-freedom

(not lock-freedom).

13.2.3 The 2-path \overline{con} algorithm

We can improve concurrency on the fallback path and guarantee lock-freedom by using a lock-free algorithm on the fallback path, and a global fetch-and-increment object F instead of a global lock. Consider an operation O implemented with the tree update template. We describe a 2-path \overline{con} implementation of O . The fallback path increments F , then executes the lock-free tree update template implementation of O , and finally decrements F . The fast path executes *sequential code* for O in a transaction. To prevent the fast path and fallback path from running concurrently, transactions on the fast path start by reading F and aborting if it is nonzero. An operation attempts to run on the fast path up to *AttemptLimit* times (waiting for F to become zero before each attempt) before resorting to the fallback path. (Note that F can actually be a *counter* object, instead of a fetch-and-increment object. The former is somewhat weaker, and can be implemented using only registers.)

Recall that operations implemented using the tree update template can only change a single pointer atomically (and can perform multiple changes atomically only by creating a connected set of new nodes that reflect the desired changes). Thus, each operation on the fallback path simply creates new nodes and changes a single pointer (and assumes that all other operations also behave this way). However, since the fast path and fallback path do not run concurrently, the fallback path does *not* impose this requirement on the fast path. Consequently, the fast path can make (multiple) direct changes to nodes. Unfortunately, as we described above, this algorithm can still suffer from concurrency bottlenecks.

13.2.4 The 3-path algorithm

One can think of the 3-path algorithm as a kind of hybrid between the 2-path *con* and 2-path \overline{con} algorithms that obtains their benefits while avoiding their downsides. Consider an operation O implemented with the tree update template. We describe a 3-path implementation of O . As in 2-path \overline{con} , there is a global fetch-and-increment object F , and the fast path executes *sequential code* for O in a transaction. The middle path and fallback path behave like the fast path and fallback path in the 2-path *con* algorithm, respectively. Each time an operation begins (resp., stops) executing on the fallback path, it increments (resp., decrements) F . (If the scalability of fetch-and-increment is of concern, then a *scalable non-zero indicator* object [53] can be used, instead.) This prevents the fast and fallback paths from running concurrently. As we described above, operations begin on the fast path, and move to the middle path after *FastLimit* attempts, or if they see $F \neq 0$. Operations move from the middle path to the fallback path after *MiddleLimit* attempts. Note that an operation never waits for the fallback path to become empty—it simply moves to the middle path.

Since the fast path and fallback path do not run concurrently, the fallback path does not impose any overhead on the fast path, except checking if $F = 0$ (offering low overhead for light workloads). Additionally, when there are operations running on the fallback path, hardware transactions can continue to run on the middle path (offering high concurrency for heavy workloads).

Correctness The correctness argument for 3-path is straightforward. The goal is to prove that all template operations are linearizable, regardless of which path they execute on. Recall that the fallback path and middle path behave like the fast path and fallback path in 2-path *con*, and the correctness of 2-path *con* was proved above. It follows that, if there are no operations on the fast path, then the correctness of operations on the middle path and fallback path is immediate from the correctness of 2-path *con*. Of course, whenever there is an operation executing on the fallback path, no operation can run on the fast path. Since operations on the fast path and middle path run in transactions, they are atomic, and any conflicts between the fast path and middle path are handled automatically by the HTM system.

Therefore, all template operations are linearizable.

Progress The progress argument for *3-path* relies on a three simple assumptions.

- A1.** The sequential code for an operation executed on the fast path must terminate after a finite number of steps if it is run on a static tree (which does not change during the operation).
- A2.** In an operation executed on the middle path or fallback path, the search phase must terminate after a finite number of steps if it is run on a static tree.
- A3.** In an operation executed on the middle path or fallback path, the update phase can modify only a finite number of nodes.

We give a simple proof that *3-path* satisfies lock-freedom. To obtain a contradiction, suppose there is an execution in which after some time t , some process takes infinitely many steps, but no operation terminates. Thus, the tree does not change after t . We first argue that no process takes infinitely many steps in a transaction T . If T occurs on the fast path, then A1 guarantees it will terminate. If T occurs on the middle path, then A2 and A3 guarantee that it will terminate. Therefore, eventually, processes only take steps on the fallback path. Progress then follows from the fact that the original tree update template implementation (our fallback path) is lock-free.

13.3 Example data structures

We use two data structures as examples to help us study our accelerated template algorithms: an unbalanced BST (similar to the Chromatic tree in Chapter 6, but with no rebalancing), and the relaxed (a,b) -tree described in Chapter 8. In this section, we briefly describe these algorithms, and give additional details on their *3-path* implementations.

Each data structure implements the ordered dictionary ADT, which stores a set of keys, and associates each key with a value. An ordered dictionary offers four operations: $\text{INSERT}(key, value)$, $\text{DELETE}(key)$, $\text{SEARCH}(key)$ and $\text{RANGEQUERY}(lo, hi)$. Recall that both data structures are *leaf-oriented* (also called *external*), which means that all of the keys in the dictionary are stored in the leaves of the tree, and internal nodes contain *routing* keys which simply direct searches to the appropriate leaf. This is in contrast to *node-oriented* or *internal* trees, in which internal nodes also contain keys in the set.

13.3.1 Unbalanced BST

Fallback path. The fallback path consists of a lock-free implementation of the operations in Figure 13.9 using the (original) tree update template. Note that this implementation is similar to the Chromatic tree in Chapter 6, except it does not perform any rebalancing. As required by the template, these operations change child pointers, but do not change the key or value fields of nodes directly. Instead, to replace a node's key or value, the node is replaced by a new copy. If key is not already in the tree, then $\text{Insert}(key, value)$ inserts a new leaf and internal node. Otherwise, $\text{Insert}(key, value)$ replaces the leaf containing key with a new leaf that contains the updated value. $\text{Delete}(key)$ replaces the leaf l being deleted and its parent with a new copy of the sibling of l .

It may seem strange that Delete creates a new copy of the deleted leaf's sibling, instead of simply reusing the existing sibling (which is not changed by the deletion). Recall that this comes from a requirement of the tree update template: each invocation of $\text{SCX}(V, R, fld, new)$ must change the field fld to a value that it has *never previously contained* (to avoid the ABA problem).

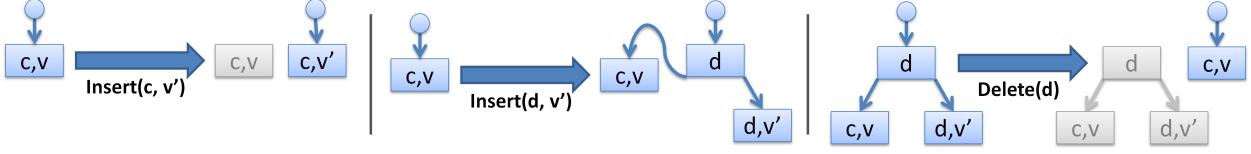


Figure 13.9: Fallback path operations for the unbalanced BST.

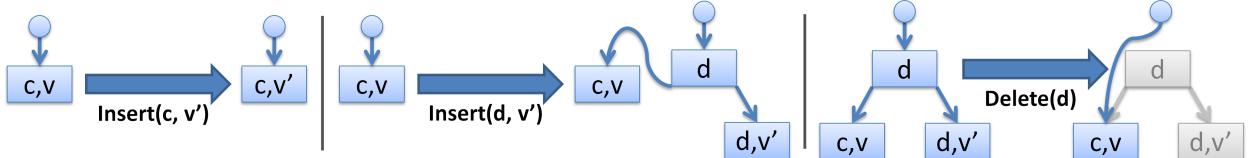


Figure 13.10: Fast path operations for the unbalanced BST. (*Insert(d, v')* is the same as on the fallback path.)

Middle path. The middle path is the same as the fallback path, except that each operation is performed in a large transaction, and the HTM-based implementation of *LLX* and *SCX* is used instead of the original implementation.

Fast path. The fast path is a sequential implementation of the BST, where each operation is executed in a transaction. Figure 13.10 shows the insertion and deletion operations on the fast path. Unlike on the fallback path, operations on the fast path directly modify the keys and values of nodes, and, hence, can avoid creating nodes in some situations. If *key* is already in the tree, then *Insert(key, value)* directly changes the value of the leaf that contains *key*. Otherwise, *Insert(key, value)* creates a new leaf and internal node and attaches them to the tree. *Delete(key)* changes a pointer to remove the leaf containing *key* and its parent from the tree.

How the fast path improves performance. The first major performance improvement on the fast path comes from a reduction in node creation. Each invocation of *Insert(key, value')* that sees *key* in the tree can avoid creating a new node by writing *value'* directly into the node that already contains *key*. In contrast, a new node had to be created on the middle path, since the middle path runs concurrently with the fallback path, which assumes that the keys and values of nodes do not change. Additionally, each invocation of *Delete* that sees *key* in the tree can avoid creating a new copy of the sibling of the deleted leaf. This optimization was not possible on the middle path, because the fallback path assumes that each successful operation writes a pointer to a newly created node. The second major improvement comes from the fact that reads and writes suffice where invocations of LLX and SCX were needed on the other paths.

13.3.2 Relaxed (a,b) -tree

Recall that a relaxed (a,b) -tree is an advanced balanced tree in which nodes contain at most b keys, and, if there are no ongoing updates (insertions and deletions), nodes contain at least a keys, and all leaves have the same depth.

Fallback path. The fallback path is the lock-free relaxed (a,b) -tree in Chapter 8, which uses the (original) tree update template. For convenience, we briefly recall how its operations work. For simplicity, we omit a description of the rebalancing steps here, and only describe the insertion and deletion operations. If *key* is in the tree, then *Insert(key, value)* replaces the leaf containing *key* with a new copy that contains $(key, value)$. Suppose *key* is not in the tree. Then, *Insert* finds the leaf *u* where the key should be inserted. If *u* is not full (has degree less than b), then it is replaced with a new copy that contains $(key, value)$. Otherwise, *u* is replaced by a subtree of three new nodes: one parent and two children. The two new children evenly share the key-value pairs of *u* and $(key, value)$. The new parent

p contains only a single routing key and two pointers (to the two new children), and is *tagged*, which indicates that the subtree rooted at p is too tall, and rebalancing should be performed to shrink its height. $\text{Delete}(key)$ replaces the leaf containing key with a new copy new that has key deleted. If the degree of new is smaller than a , then rebalancing must be performed.

Middle path. This path is obtained from the fallback path the same way as in the unbalanced BST.

Fast path. The fast path is a sequential implementation of a relaxed (a,b) -tree whose operations are executed inside transactions. Like the external BST, the major performance improvement over the middle path comes from the facts that (1) operations create fewer nodes, and (2) reads and writes suffice where LLX and SCX were needed on the other paths. In particular, $\text{Insert}(key,value)$ and $\text{Delete}(key)$ simply directly modify the keys and values of leaves, instead of creating new nodes, except in the case of an Insert into a full node u . In that case, two new nodes are created: a parent and a sibling for u . (Recall that this case resulted in the creation of three new nodes on the fallback path and middle path.) Note that reducing node creation is more impactful for the relaxed (a,b) -tree than for the unbalanced BST, since nodes are much larger. As a minor point, we found that it was faster in practice to perform *rebalancing steps* by creating new nodes, and simply replacing the old nodes with the new nodes that reflect the desired change (instead of rebalancing by directly changing the keys, values and pointers of nodes).

13.4 Experimental results

We used two different Intel systems for our experiments: a dual-socket 12-core E7-4830 v3 with hyperthreading for a total of 48 hardware threads (running Ubuntu 14.04LTS), and a dual-socket 18-core E5-2699 v3 with hyperthreading for a total of 72 hardware threads (running Ubuntu 15.04). Each machine had 128GB of RAM. We used the scalable thread-caching allocator (tcmalloc) from the Google perftools library. All code was compiled on GCC 4.8+ with arguments `-std=c++0x -O2 -mcx16`. (Using the higher optimization level `-O3` did not significantly improve performance for any algorithm, and decreased performance for some algorithms.) On both machines, we *pinned* threads such that we saturate one socket before scheduling any threads on the other.

Data structure parameters. Recall that nodes in the relaxed (a,b) -tree contain up to b keys, and, when there are no ongoing updates, they contain at least a keys (where $b \geq 2a - 1$). In our experiments, we fix $a = 6$ and $b = 16$. With $b = 16$, each node occupies four consecutive cache lines. Since $b \geq 2a - 1$, with $b = 16$, we must have $a \leq 8$. We chose to make a slightly smaller than 8 in order to exploit a performance tradeoff: a smaller minimum degree may slightly increase depth, but decreases the number of rebalancing steps that are needed to maintain balance.

Template implementations studied. We implemented each of the data structure with four different template implementations: *3-path*, *2-path con*, *TLE* and the original template implementation, which we call *Non-HTM*. (*2-path con* is omitted, since it performed similarly to *TLE*, and cluttered the graphs.) The *2-path con* and *TLE* implementations perform up to 20 attempts on the fast path before resorting to the fallback path. *3-path* performs up to 10 attempts (each) on the fast path and middle path. We implemented memory reclamation using DEBRA [28], an epoch based reclamation scheme. A more efficient way to reclaim memory for *3-path* is proposed in Section 13.6

13.4.1 Light vs. Heavy workloads

Methodology. We study two workloads: in **light**, n processes perform updates (50% insertion and 50% deletion), and in **heavy**, $n - 1$ processes perform updates, and one thread performs 100% range queries (RQs). For each workload

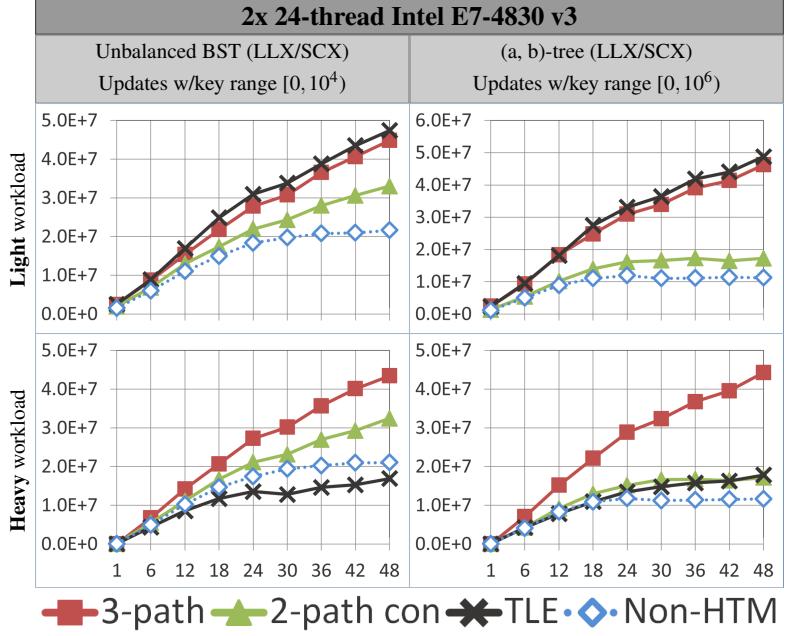


Figure 13.11: Results (48-thread system) showing throughput (operations per second) versus the number of concurrent processes.

and data structure implementation, and a variety of thread counts, we perform a set of five randomized trials. In each trial, n processes perform either updates or RQs (as appropriate for the workload) for one second, and counted the number of completed operations. Updates are performed on keys drawn uniformly randomly from a fixed key range $[0, K]$. RQs are performed on ranges $[lo, lo + s]$ where lo is uniformly random in $[0, K]$ and s is chosen, according to a probability distribution described below, from $[1, 1000]$ for the BST and $[1, 10000]$ for the (a, b) -tree. (We found that nodes in the (a, b) -tree contained approximately 10 keys, on average, so the respective maximum values of s for the BST and (a, b) -tree resulted in range queries returning keys from approximately the same number of nodes in both data structures.) To ensure that we are measuring steady-state performance, at the start of each trial, the data structure is prefilled by having threads perform 50% insertions and 50% deletions on uniform keys until the data structure contains approximately half of the keys in $[0, K]$.

We verified the correctness of each data structure after each trial by computing *key-sum hashes*. Each thread maintains the sum of all keys it successfully inserts, minus the sum of all keys it successfully deletes. At the end of the trial, the total of these sums over all threads must match the sum of keys in the tree.

Probability distribution of s . We chose the probability distribution of s to produce many small RQs, and a smaller number of very large ones. To achieve this, we chose s to be $\lfloor x^2 S \rfloor + 1$, where x is a uniform real number in $[0, 1]$, and $S = 1000$ for the BST and $S = 10000$ for the (a, b) -tree. By squaring x , we bias the uniform distribution towards zero, creating a larger number of small RQs.

Results. We briefly discuss the results from the 48 thread machine, which appear in Figure 13.11. The BST and the relaxed (a, b) -tree behave fairly similarly. Since the (a, b) -tree has large nodes, it benefits much more from a low-overhead fast path (in TLE or 3-path) which can avoid creating new nodes during updates. In the light workloads, 3-path performs significantly better than 2-path con (which has more overhead) and approximately as well as TLE.

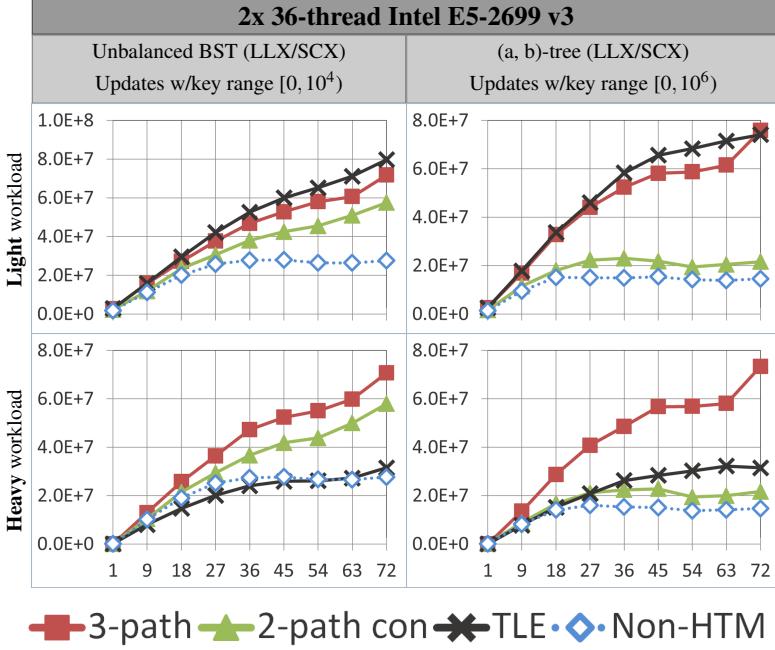


Figure 13.12: Supplementary results from the 2x36 thread Intel E5-2699 v3.

On average, the *3-path* algorithms completed 2.1x as many operations as their *non-HTM* counterparts (and with 48 concurrent processes, this increases to 3.0x, on average). In the heavy workloads, *3-path* significantly outperforms *TLE* (completing 2.0x as many operations, on average), which suffers from *excessive waiting*. Interestingly, *3-path* is also significantly faster than *2-path con* in the heavy workloads. This is because, even though RQs are always being performed, some RQs can succeed on the fast path, so many update operations can still run on the fast path in *3-path*, where they incur much less overhead (than they would in *2-path con*).

Supplementary experimental results from the 72-thread machine appear in Figure 13.12

13.4.2 Code path usage and abort rates

To gain further insight into the behaviour of our accelerated template implementations, we gathered some additional metrics about the experiments described above. Here, we only describe results from the 48-thread Intel machine. (Results from the 72-thread Intel machine were similar.)

Operations completed on each path. We started by measuring how often operations completed successfully on each execution path. This revealed that operations almost always completed on the fast path. Broadly, over all thread counts, the minimum number of operations completed on the fast path in any trial was 86%, and the average over all trials was 97%.

In each trial that we performed with 48 concurrent threads, at least 96% of operations completed on the fast path, *even in the workloads with RQs*. Recall that RQs are the operations most likely to run on the fallback path, and they are only performed by a single thread, so they make up a relatively small fraction of the total operations performed in a trial. In fact, our measurements showed that the number of operations which completed on the fallback path was never more than a fraction of one percent in our trials with 48 concurrent threads.

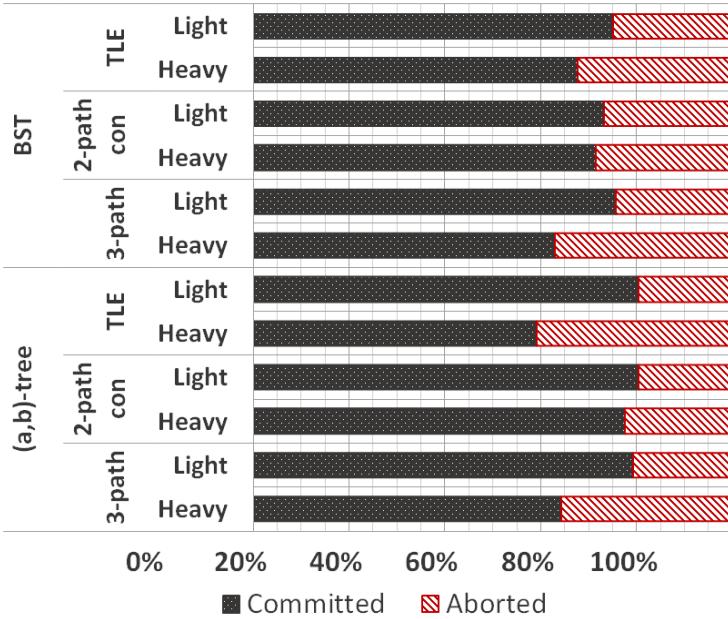


Figure 13.13: Summary of how many transactions commit vs. how many abort in our experiments on the 48-thread Intel machine.

In light of this, it might be somewhat surprising that the performance of TLE was so much worse in heavy workloads than light ones. However, the cost of serializing threads is high, and this cost is compounded by the fact that the operations which complete on the fallback path are often long-running. Of course, in workloads where more operations run on the fallback path, the advantage of improving concurrency between paths would be even greater.

Commit/abort rates. We also measured how many transactions committed and how many aborted, on each execution path, in each of our trials. Figure 13.13 summarizes the average commit/abort rates for each data structure, template implementation and workload. Since nearly all operations completed on the fast path, we decided not to distinguish between the commit/abort rate on the fast path and the commit/abort rate on the middle path.

13.4.3 Comparing with hybrid transactional memory

Hybrid transactional memory (hybrid TM) combines hardware and software transactions to hide the limitations of HTM and guarantee progress. This offers an alternative way of using HTM to implement concurrent data structures. Note, however, that state of the art hybrid TMs use locks. So, **they cannot be used to implement lock-free data structures**. Regardless, to get an idea of how such implementations would perform, relative to our accelerated template implementations, we implemented the unbalanced BST using Hybrid NOrec, which is arguably the fastest hybrid TM implementation with readily available code [109].

If we were to use a precompiled library implementation of Hybrid NOrec, then the unbalanced BST algorithm would have to perform a library function call for *each read and write to shared memory*, which would incur significant overhead. So, we directly compiled the code for Hybrid NOrec into the code for the BST, allowing the compiler to inline the Hybrid NOrec functions for reading and writing from shared memory into our BST code, eliminating this

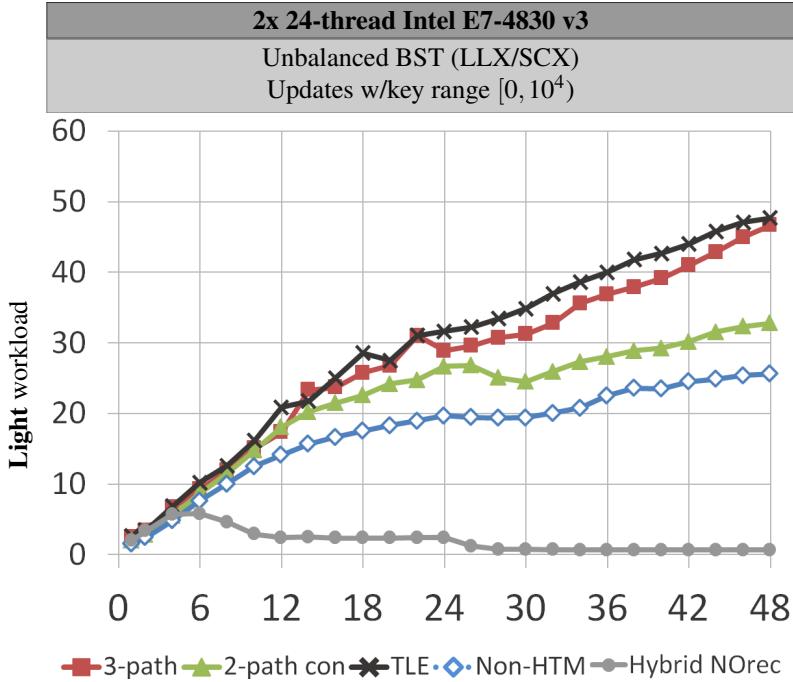


Figure 13.14: Results showing throughput (operations per second) versus number of processes for an unbalanced BST implemented with different tree update template algorithms, and with the hybrid TM algorithm *Hybrid NOrec*.

overhead. Of course, if one intended to use hybrid TM in practice (and not in a research prototype), one would use a precompiled library, with all of the requisite overhead. Thus, the following results are quite charitable towards hybrid TMs.

We implemented the BST using Hybrid NOrec by wrapping sequential code for the BST operations in transactions, and manually replacing each read from (resp., write to) shared memory with a read (resp., write) operation provided by Hybrid NOrec. Figure 13.14 compares the performance of the resulting implementation to the other BST implementations discussed in Section 13.4.

The BST implemented with Hybrid NOrec performs relatively well with up to six processes. However, beyond six processes, it experiences severe negative scaling. The negative scaling occurs because Hybrid NOrec increments a global counter in each updating transaction (i.e., each transaction that performs at least one write). This global contention hotspot in updating transactions causes many transactions to abort, simply because they contend on the global counter (and not because they conflict on any data in the tree). However, even without this bottleneck, Hybrid NOrec would still perform poorly in heavy workloads, since it incurs very high instrumentation overhead for software transactions (which must acquire locks, perform repeated validation of read-sets, maintain numerous auxiliary data structures for read-sets and write-sets, and so on). Note that this problem is not unique to Hybrid NOrec, as every hybrid TM must use a software TM as its fallback path in order to guarantee progress. In contrast, in our template implementations, the software-only fallback path is a fast lock-free algorithm.

13.5 Modifications for performing searches outside of transactions

In this section, we describe how the *3-path* implementations of the unbalanced BST and relaxed (a,b) -tree can be modified so that each operation attempt on the fast path or middle path performs its search phase *before* starting a transaction (and only performs its update phase in a transaction). (The same technique also applies to the *2-path con* implementations.) First, note that the lock-free search procedure for each of these data structures is actually a standard, sequential search procedure. Consequently, a simple sequential search procedure will return the correct result, regardless of whether it is performed inside a transaction. (Generally, whenever we produce a *3-path* implementation starting from a lock-free fallback path, we will have access to a correct non-transactional search procedure.)

The difficulty is that, when an operation starts a transaction and performs its update phase, it may be working on a part of the tree that was deleted by another operation. One can imagine an operation O_d that deletes an entire subtree, and an operation O_i that inserts a node into that subtree. If the search phase of O_i is performed, then O_d is performed, then the update phase of O_i is performed, then O_i may erroneously insert a node into the deleted subtree.

We fix this problem as follows. Whenever an operation O on the fast path or middle path removes a node from the tree, it sets a *marked* bit in the node (just like operations on the fallback path do). Whenever O first accesses a node u in its transaction, it checks whether u has its *marked* bit set, and, if so, aborts immediately. This way, O 's transaction will commit only if every node that it accessed is in the tree.

We found that this modification yielded small performance improvements (on the order of 5-10%) in our experiments. The reason this improves performance is that fewer memory locations are tracked by the HTM system, which results in fewer capacity aborts. We briefly discuss why the performance benefit is small in our experiments. The relaxed (a,b) -tree has a very small height, because it is balanced, and its nodes contain many keys. The BST also has a fairly small height (although it is considerably taller than the relaxed (a,b) -tree), because processes in our experiments perform insertions and deletions on uniformly random keys, which leads to trees of logarithmic height with high probability. So, in each case, the sequence of nodes visited by searches is relatively small, and is fairly unlikely to cause capacity aborts.

The performance benefit associated with this modification will be greater for data structures, operations or workloads in which an operation's search phase will access a large number of nodes. Additionally, IBM's HTM implementation in their POWER8 processors is far more prone to capacity aborts than Intel's implementation, since a transaction *will* abort if it accesses more than 64 different cache lines [103]. (In contrast, in Intel's implementation, a transaction can potentially commit after accessing tens of thousands of cache lines.) Thus, this modification could lead to significantly better performance on POWER8 processors.

13.6 Reclaiming memory more efficiently

In Chapter 11, we explained that lock-free memory reclamation is hard because a process can always be poised to access a node just after it is freed. The penalty for accessing a freed node can be data corruption or a program crash (e.g., due to a segmentation fault). Once a process has removed a node from a tree, it knows that no other process can reach the node by following pointers starting from the root of the tree. However, a process may still be able to reach the removed node by starting from a pointer in its *private memory*. Thus, lock-free memory reclamation schemes must implement special mechanisms to determine when it is safe to free a node (because the node cannot no longer be reached by any process).

However, advanced memory reclamation schemes are unnecessary if (1) *all* accesses to nodes are performed inside transactions, and (2) at the end of each operation on the data structure, the process performing the operation discards all pointers in its private memory that point to nodes. Consider an implementation of a tree that satisfies (1) and (2). Let O be an operation that removes a node u from the tree, and then frees u immediately thereafter. We argue that, if an operation O' accesses u , and O subsequently removes u from the tree before O' has terminated, then O' will abort. (Note that, if O' terminates before u is removed from the tree, then there is no problem.) By (2), O' must obtain a pointer to u by traversing the tree, starting at the root. Consider the sequence of pointers followed by O' as it traverses the tree to reach u . At least one of these pointers must change after O' reads it during its traversal, and before O frees u (or else u would still be in the tree after O removed it—a contradiction). Therefore, O' will abort due to a data conflict. Consequently, in such a data structure, reclaiming memory is as easy as invoking `free()` immediately after a node is removed.

In our three path algorithms, the fast path can only run concurrently with the middle path (but not the fallback path), and the fast path and middle path both satisfy (1) and (2). Thus, memory can be reclaimed on the fast path simply by using `free()` immediately after removing a node inside a transaction. Our performance experiments did *not* implement this optimization, but doing so would likely further improve the performance of the three path algorithms.

13.7 Related work

Hybrid TMs share some similarities to our work, since they all feature multiple execution paths. The first hybrid TM algorithms allowed HTM and STM transactions to run concurrently [78, 41]. Hybrid NOrec [40] and Reduced hardware NOrec [95] are hybrid TMs that both use global locks on the fallback path, eliminating any concurrency. We discuss two additional hybrid TMs, Phased TM [90] (PhTM) and Invyswell [33], in more detail.

PhTM alternates between five *phases*: HTM-only, STM-only, concurrent HTM/STM, and two global locking phases. Roughly speaking, PhTM’s HTM-only phase corresponds to our uninstrumented fast path, and its concurrent HTM/STM phase corresponds to our middle HTM and fallback paths. However, their STM-only phase (which allows no concurrent hardware transactions) and global locking phases (which allow no concurrency) have no analogue in our approach. In heavy workloads, PhTM must oscillate between its HTM-only and concurrent HTM/STM phases to maximize the performance benefit it gets from HTM. When changing phases, PhTM typically waits until all in-progress transactions complete before allowing transactions to begin in the new mode. Thus, after a phase change has begun, and before the next phase has begun, there is a window during which new transactions must wait (reducing performance). One can also think of our three path approach as proceeding in two phases: one with concurrent fast/middle transactions and one with concurrent middle/fallback transactions. However, in our approach, “phase changes” do not introduce any waiting, and there is always concurrency between two execution paths.

Invyswell is closest to our three path approach. At a high level, it features an HTM middle path and STM slow path that can run concurrently (sometimes), and an HTM fast path that can run concurrently with the middle path (sometimes) but not the slow path, and two global locking fallback paths (that prevent any concurrency). Invyswell is more complicated than our approach, and has numerous restrictions on when transactions can run concurrently. Our three path methodology does not have these restrictions. The HTM fast path also uses an optimization called lazy subscription. It has been shown that lazy subscription can cause opacity to be violated, which can lead to data corruption or program crashes [45].

Hybrid TM is very general, and it pays for its generality with high overhead. Consequently, data structure designers

can extract far better performance for library code by using more specialized techniques. Additionally, we stress that state-of-the-art hybrid TMs use locks, so they cannot be used to produce lock-free data structures.

Different options for concurrency have recently begun to be explored in the context of TLE. Refined TLE [47] and Amalgamated TLE [4] both improve the concurrency of TLE when a process is on the fallback path by allowing HTM transactions to run concurrently with a *single process* on the fallback path. Both of these approaches still serialize processes on the fallback path. They also use locks, so they cannot be used to produce lock-free data structures.

Timnat, Herlihy and Petrank [118] proposed using a strong synchronization primitive called *multiword compare-and-swap* (k -CAS) to obtain fast HTM algorithms. They showed how to take an algorithm implemented using k -CAS and produce a two-path implementation that allows concurrency between the fast and fallback paths. One of their approaches used a lock-free implementation of k -CAS on the fallback path, and an HTM-based implementation of k -CAS on the fast path. They also experimented with two-path implementations that do not allow concurrency between paths, and found that allowing concurrency between the fast path and fallback path introduced significant overhead. Makreshanski, Levandoski and Stutsman [93] also independently proposed using HTM-based k -CAS in the context of databases.

Liu, Zhou and Spear [91] proposed a methodology for accelerating concurrent data structures using HTM, and demonstrated it on several lock-free data structures. Their methodology uses an HTM-based fast path and a non-transactional fallback path. The fast path implementation of an operation is obtained by encapsulating part (or all) of the operation in a transaction, and then applying sequential optimizations to the transactional code to improve performance. Since the optimizations do not change the code's logic, the resulting fast path implements the same logic as the fallback path, so both paths can run concurrently. Consequently, the fallback path imposes overhead on the fast path.

Some of the optimizations presented in that paper are similar to some optimizations in our HTM-based implementation of *LLX* and *SCX*. For instance, when they applied their methodology to the lock-free unbalanced BST of Ellen et al. [52], they observed that helping can be avoided on the fast path, and that the descriptors which are normally created to facilitate helping can be replaced by a small number of statically allocated descriptors. However, they did not give details on exactly how these optimizations work, and did not give correctness arguments for them. In contrast, our optimizations are applied to a more complex algorithm, and are proved correct.

Multiversion concurrency control (MVCC) is another way to implement range queries efficiently [20, 14]. At a high level, it involves maintaining multiple copies of data to allow read-only transactions to see a consistent view of memory and serialize even in the presence of concurrent modifications. However, our approach could also be applied to operations that *modify* a range of keys, so it is more general than MVCC.

13.8 Other uses for the *3-path* approach

13.8.1 Accelerating data structures that use read-copy-update (RCU)

In this section, we sketch a *3-path* algorithm for an ordered dictionary implemented with a node-oriented unbalanced BST that uses the RCU synchronization primitives. The intention is for this to serve as an example of how one might use the *3-path* approach to accelerate a data structure that uses RCU.

RCU is both a programming paradigm and a set of synchronization primitives. The paradigm organizes operations into a search/reader phase and an (optional) update phase. In the update phase, all modifications are made on a *new*

copy of the data, and the old data is atomically replaced with the new copy. In this work, we are interested in the *RCU primitives* (rather than the paradigm).

Semantics of RCU primitives and their uses. The basic RCU synchronization primitives are *rcu_begin*, *rcu_end* and *rcu_wait* [43]. Operations invoke *rcu_begin* and *rcu_end* at the beginning and end of the search phase, respectively. The interval between an invocation of *rcu_begin* and the next invocation of *rcu_end* by the same operation is called a *read-side critical section*. An invocation of *rcu_wait* blocks until all read-side critical sections that started before the invocation of *rcu_wait* have ended. One common use of *rcu_wait* is to wait, after a node has been deleted, until no readers can have a pointer to it, so that it can safely be freed. It is possible to use RCU as the sole synchronization mechanism for an algorithm, if one is satisfied with allowing many concurrent readers, but only a single updater at a time. If multiple concurrent updaters are required, then another synchronization mechanism, such as fine-grained locks, must also be used. However, one must be careful when using locks with RCU, since locks cannot be acquired inside a read-side critical section without risking deadlock.

The CITRUS data structure. We consider how one might accelerate a node-oriented BST called CITRUS [10], which uses the RCU primitives, and fine-grained locking, to synchronize between threads. First, we briefly describe the implementation of CITRUS. At a high level, RCU is used to allow operations to search without locking, and fine-grained locking is used to allow multiple updaters to proceed concurrently.

The main challenge in the implementation of CITRUS is to prevent race conditions between searches (which do not acquire locks) and deletions. When an internal node u with two children is deleted in an internal BST, its key is replaced by its successor's key, and the successor (which is a leaf) is then deleted. This case must be handled carefully, or else the following can happen. Consider concurrent invocations D of *Delete(key)* and S of *Search(key')*, where key' is the successor of key . Suppose S traverses past the node u containing key , and then D replaces u 's key by key' , and deletes the node containing key' . The search will then be unable to find key' , even though it has been in the tree throughout the entire search. To avoid this problem in CITRUS, rather than changing the key of u directly, D replaces u with a new copy that contains key' . After replacing u , D invokes *rcu_wait* to wait for any ongoing searches to finish, before finally deleting the leaf containing key' . The primary sources of overhead in this algorithm are invocations of *rcu_wait*, and lock acquisition costs.

Fallback path. The fallback path uses the implementation of CITRUS in [10] (additionally incrementing and decrementing the global fetch-and-add object F , as described in Section 13.2).

Middle path. The middle path is obtained from the fallback path by wrapping each fallback path operation in a transaction and optimizing the resulting code. The most significant optimization comes from an observation that the invocation of *rcu_wait* in *Delete* is unnecessary since transactions make the operation atomic. Invocations of *rcu_wait* are the dominating performance bottleneck in CITRUS, so this optimization greatly improves performance. A smaller improvement comes from the fact that transactions can avoid acquiring locks. Transactions on the middle path must ensure that all objects they access are not locked by other operations (on the fallback path), or else they might modify objects locked by operations on the fallback path. However, it is not necessary for transaction to actually *acquire* locks. Instead, it suffices for a transaction to simply *read* the lock state for all objects it accesses (before accessing them) and ensure that they are not held by another process. This is because transactions *subscribe* to each memory location they access, and, if the value of the location (in this case, the lock state) changes, then the transaction will abort.

Fast path. The fast path is a sequential implementation of a node-oriented BST whose operations are executed in transactions. As in the other 3-path algorithms, each transaction starts by reading F , and aborts if it is nonzero. This prevents operations on the fast path and fallback path from running concurrently. There are two main differences

between fast path and the middle path. First, the fast path does not invoke *rcu_begin* and *rcu_end*. These invocations are unnecessary, because operations on the fast path can run concurrently *only* with other operations on the fast path or middle path, and *neither* path depends on RCU for its correctness. (However, the middle path *must* invoke these operations, because it runs concurrently with the fallback path, which relies on RCU.) The second difference is that the fast path does not need to read the lock state for any objects. Any conflicts between operations on the fast and middle path are resolved directly by the HTM system.

13.8.2 Accelerating data structures that use k -CAS

In this section, we sketch how one might produce a *3-path* implementation of a lock-free algorithm that uses the k -CAS synchronization primitive.

A k -CAS operation takes, as its arguments, k memory locations, expected values and new values, and atomically: reads the memory locations and, if they contain their expected values, writes new values into each of them. We consider the implementation of k -CAS in Chapter 12. For convenience, we give a brief overview of the implementation, here. At a high level, a k -CAS creates a descriptor object that describes the k -CAS operation, then uses CAS to store a pointer to this descriptor in each memory location that it operates on. Then, it uses CAS to change each memory location to contain its new value. While a k -CAS is in progress, some fields may contain pointers to descriptor objects, instead of their regular values. Consequently, reading a memory location becomes more complicated: it requires reading the location, then testing whether it contains a pointer to a descriptor object, and, if so, helping the k -CAS operation that it represents, before finally returning a value.

Fallback path. An operation on the fallback path simply increments a global fetch-and-increment object F , performs the appropriate operation of the lock-free algorithm, then decrements F .

Middle path. The middle path is similar to the fallback path, except that the CAS-based implementation of k -CAS is replaced with a straightforward implementation using HTM (following the approach in [118]). This HTM-based implementation performs the entire k -CAS atomically, and does not need to create a descriptor. However, to facilitate concurrency with the lock-free implementation of k -CAS, the HTM-based k -CAS *does* check whether each address it reads contains pointer to a descriptor, and, if so, helps the operation that created the descriptor before trying again.

Fast path. The fast path is a sequential implementation wrapped in a transaction. The main optimization on the fast path comes from the fact that, since there are no concurrent operations on the fallback path, there are no pointers to k -CAS descriptors in shared memory. Consequently, operations on the fast path do not need to check whether any values they read from shared memory are actually pointers to k -CAS descriptors, which significantly reduces overhead.

13.9 Summary

In this chapter, we explored the design space for HTM-based implementations of the tree update template of Brown et al. and presented four accelerated implementations. We discussed performance issues affecting HTM-based algorithms with two execution paths, and developed an approach that avoids them by using three paths. We used our template implementations to accelerate two different lock-free data structures, and performed experiments that showed significant performance improvements over several different workloads. This makes our implementations an attractive option for producing fast concurrent data structures for inclusion in libraries, where performance is critical.

Our accelerated data structures each perform an entire operation inside a single transaction (except on the fallback code path, where no transactions are used). We discussed how one can improve efficiency by performing the read-only *searching* part of an operation non-transactionally, and simply using a transaction to perform any modifications to the data structure. Our *3-path* approach may also have other uses. As an example, we sketched an accelerated *3-path* implementation of a node-oriented BST that uses the read-copy-update (RCU) synchronization primitives. We suspect that a similar approach could be used to accelerate other data structures that use RCU. Additionally, we describe how one might produce a *3-path* implementation of a lock-free algorithm that uses the *k*-CAS synchronization primitive.

Bibliography

- [1] M AdelsonVelskii and Evgenii Mikhailovich Landis. An algorithm for the organization of information. Technical report, DTIC Document, 1963.
- [2] Yehuda Afek, Dalia Dauber, and Dan Touitou. Wait-free made fast. In *Proceedings of the Twenty-seventh Annual ACM Symposium on Theory of Computing*, STOC '95, pages 538–547, 1995.
- [3] Yehuda Afek, Haim Kaplan, Boris Korenfeld, Adam Morrison, and Robert E. Tarjan. CBTTree: A practical concurrent self-adjusting search tree. In *Proc. 26th International Symposium on Distributed Computing*, volume 7611 of *LNCS*, pages 1–15, 2012.
- [4] Yehuda Afek, Alexander Matveev, Oscar R Moll, and Nir Shavit. Amalgamated lock-elision. In *Distributed Computing*, pages 309–324. Springer, 2015.
- [5] Yehuda Afek, Michael Merritt, Gadi Taubenfeld, and Dan Touitou. Disentangling multi-object operations. In *Proc. 16th ACM Symposium on Principles of Distributed Computing*, pages 111–120, 1997.
- [6] Zahra Aghazadeh, Wojciech Golab, and Philipp Woelfel. Making objects writable. In *Proceedings of the 2014 ACM symposium on Principles of distr. comp.*, pages 385–395. ACM, 2014.
- [7] Dan Alistarh, Patrick Eugster, Maurice Herlihy, Alexander Matveev, and Nir Shavit. Stacktrack: An automated transactional approach to concurrent memory reclamation. In *Proceedings of the 9th European Conference on Comp. Sys.*, page 25. ACM, 2014.
- [8] Dan Alistarh, William M. Leiserson, Alexander Matveev, and Nir Shavit. Threadscan: Automatic and scalable memory reclamation. In *Proceedings of the 27th ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '15, pages 123–132, New York, NY, USA, 2015. ACM.
- [9] James H. Anderson and Mark Moir. Universal constructions for multi-object operations. In *Proceedings of the Fourteenth Annual ACM Symposium on Principles of Distributed Computing*, PODC '95, pages 184–193, 1995.
- [10] Maya Arbel and Hagit Attiya. Concurrent updates with RCU: Search tree as an example. In *Proceedings of the 2014 ACM Symposium on Principles of Distributed Computing*, PODC '14, pages 196–205, New York, NY, USA, 2014. ACM.
- [11] David M Arnow and Aaron M Tenenbaum. An empirical comparison of B-trees, compact B-trees and multiway trees. In *ACM SIGMOD Record*, volume 14:2, pages 33–46. ACM, 1984.

- [12] David M Arnow, Aaron M Tenenbaum, and Connie Wu. P-trees: Storage efficient multiway trees. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–121. ACM, 1985.
- [13] Hagit Attiya and Eshcar Hillel. Highly concurrent multi-word synchronization. *Theoretical Computer Science*, 412(12):1243–1262, 2011.
- [14] Hagit Attiya and Eshcar Hillel. A single-version stm that is multi-versioned permissive. *Theory of Computing Systems*, 51(4):425–446, 2012.
- [15] Ricardo A. Baeza-Yates and P-A Larson. Performance of B+-trees with partial expansions. *Knowledge and Data Eng., IEEE Transactions on*, 1(2):248–257, 1989.
- [16] Oana Balmau, Rachid Guerraoui, Maurice Herlihy, and Igor Zablotchi. Fast and robust memory reclamation for concurrent data structures. In *Proceedings of the 28th ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA ’16, pages 349–359, New York, NY, USA, 2016. ACM.
- [17] Greg Barnes. A method for implementing lock-free shared-data structures. In *Proceedings of the Fifth Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA ’93, pages 261–270, 1993.
- [18] R Bayer and E McCreight. Organization and maintenance of large indexes. Technical Report D1-82-0989, Boeing Scientific Research Laboratories, 1970.
- [19] Rudolf Bayer. Symmetric binary B-trees: Data structure and maintenance algorithms. *Acta Informatica*, 1(4):290–306, 1972.
- [20] Philip A Bernstein and Nathan Goodman. Multiversion concurrency control-theory and algorithms. *ACM Transactions on Database Systems (TODS)*, 8(4):465–483, 1983.
- [21] Luc Bougé, Joaquim Gabarró, Xavier Messeguer, and Nicolas Schabanel. Height-relaxed AVL rebalancing: A unified, fine-grained approach to concurrent dictionaries. Technical Report LSI-98-12-R, Universitat Politècnica de Catalunya, 1998. Available from http://www.lsi.upc.edu/dept/techreps/l1listat_detallat.php?id=307.
- [22] Joan Boyar, Rolf Fagerberg, and Kim S. Larsen. Amortization results for chromatic search trees, with an application to priority queues. *Journal of Computer and System Sciences*, 55(3):504–521, December 1997.
- [23] Anastasia Braginsky, Alex Kogan, and Erez Petrank. Drop the anchor: lightweight memory management for non-blocking data structures. In *Proceedings of the 25th ACM symposium on Parallelism in alg. and arch.*, pages 33–42. ACM, 2013.
- [24] Anastasia Braginsky and Erez Petrank. A lock-free B+tree. In *Proc. 24th ACM Symposium on Parallelism in Algorithms and Architectures*, pages 58–67, 2012.
- [25] Alex Brodsky, Faith Ellen, and Philipp Woelfel. Fully-adaptive algorithms for long-lived renaming. *Distributed Computing*, 24(2):119, 2011.

- [26] Hervé Brönnimann, Jyrki Katajainen, and Pat Morin. Putting your data structure on a diet. *CPH STL Rep*, 1, 2007.
- [27] Nathan G. Bronson, Jared Casper, Hassan Chafi, and Kunle Olukotun. A practical concurrent binary search tree. In *Proc. 15th ACM Symposium on Principles and Practice of Parallel Programming*, pages 257–268, 2010.
- [28] Trevor Brown. Reclaiming memory for lock-free data structures: There has to be a better way. In *Proceedings of the 2015 ACM Symposium on Principles of Distributed Computing*, PODC ’15, pages 261–270, 2015.
- [29] Trevor Brown, Faith Ellen, and Eric Ruppert. Pragmatic primitives for non-blocking data structures. In *Proceedings of the 2013 ACM Symposium on Principles of Distributed Computing*, PODC ’13, pages 13–22, 2013. Full version available from <http://tbrown.pro>.
- [30] Trevor Brown, Faith Ellen, and Eric Ruppert. A general technique for non-blocking trees. In *Proceedings of the 19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP ’14, pages 329–342, 2014. Full version available from <http://tbrown.pro>.
- [31] Trevor Brown and Joanna Helga. Non-blocking k-ary search trees. In *Proc. 15th International Conf. on Principles of Distributed Systems*, volume 7109 of *LNCS*, pages 207–221, 2011.
- [32] Shirley Browne, Jack Dongarra, Nathan Garner, George Ho, and Philip Mucci. A portable programming interface for performance evaluation on modern processors. *International Journal of High Performance Computing Applications*, 14(3):189–204, 2000.
- [33] Irina Calciu, Justin Gottschlich, Tatiana Shpeisman, Gilles Pokam, and Maurice Herlihy. Invyswell: a hybrid transactional memory for haswell’s restricted transactional memory. In *Proceedings of the 23rd international conference on Parallel architectures and compilation*, pages 187–200. ACM, 2014.
- [34] Phong Chuong, Faith Ellen, and Vijaya Ramachandran. A universal construction for wait-free transaction friendly data structures. In *Proc. 22nd ACM Symposium on Parallelism in Algorithms and Architectures*, pages 335–344, 2010.
- [35] Nachshon Cohen and Erez Petrank. Automatic memory reclamation for lock-free data structures. In *Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications*, OOPSLA 2015, pages 260–279, New York, NY, USA, 2015. ACM.
- [36] Nachshon Cohen and Erez Petrank. Efficient memory management for lock-free data structures with optimistic access. In *Proceedings of the 27th ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA ’15, pages 254–263, New York, NY, USA, 2015. ACM.
- [37] Tyler Crain, Vincent Gramoli, and Michel Raynal. A speculation-friendly binary search tree. In *Proc. 17th ACM Symp. on Principles and Practice of Parallel Programming*, pages 161–170, 2012.
- [38] Tyler Crain, Vincent Gramoli, and Michel Raynal. A contention-friendly binary search tree. In *Euro-Par*, pages 229–240, 2013.
- [39] Karel Culik II, Th Ottmann, and Derick Wood. Dense multiway trees. *ACM Transactions on Database Systems (TODS)*, 6(3):486–512, 1981.

- [40] Luke Dalessandro, Francois Carouge, Sean White, Yossi Lev, Mark Moir, Michael L Scott, and Michael F Spear. Hybrid norec: A case study in the effectiveness of best effort hardware transactional memory. *ACM SIGPLAN Notices*, 46(3):39–52, 2011.
- [41] Peter Damron, Alexandra Fedorova, Yossi Lev, Victor Luchangco, Mark Moir, and Daniel Nussbaum. Hybrid transactional memory. In *ACM Sigplan Notices*, volume 41:11, pages 336–346. ACM, 2006.
- [42] Damian Dechev, Peter Pirkelbauer, and Bjarne Stroustrup. Understanding and effectively preventing the ABA problem in descriptor-based lock-free designs. In *Proc. 13th IEEE Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing*, pages 185–192, 2010.
- [43] Mathieu Desnoyers, Paul E. McKenney, Alan S. Stern, Michel R. Dagenais, and Jonathan Walpole. User-level implementations of Read-Copy Update. *IEEE Transactions on Parallel and Distributed Systems*, 23(2):375–382, 2012.
- [44] David L Detlefs, Paul A Martin, Mark Moir, and Guy L Steele Jr. Lock-free reference counting. *Distributed Computing*, 15(4):255–271, 2002.
- [45] Dave Dice, T Harris, Alex Kogan, Yossi Lev, and Mark Moir. Pitfalls of lazy subscription. In *Proceedings of the 6th Workshop on the Theory of Transactional Memory, Paris, France*, 2014.
- [46] Dave Dice, Maurice Herlihy, and Alex Kogan. Fast non-intrusive memory reclamation for highly-concurrent data structures. In *Proceedings of the 2016 ACM SIGPLAN International Symposium on Memory Management, ISMM 2016*, pages 36–45, New York, NY, USA, 2016. ACM.
- [47] Dave Dice, Alex Kogan, and Yossi Lev. Refined transactional lock elision. In *Proceedings of the 21st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPoPP ’16*, pages 19:1–19:12, New York, NY, USA, 2016. ACM.
- [48] Dave Dice, Yossi Lev, Mark Moir, and Daniel Nussbaum. Early experience with a commercial hardware transactional memory implementation. In *Proceedings of the 14th International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS XIV*, pages 157–168, New York, NY, USA, 2009. ACM.
- [49] Dana Drachsler, Martin Vechev, and Eran Yahav. Practical concurrent binary search trees via logical ordering. In *Proceedings of the 19th annual ACM symposium on Principles and practice of parallel programming*, pages 343–356. ACM, 2014.
- [50] Aleksandar Dragojević, Maurice Herlihy, Yossi Lev, and Mark Moir. On the power of hardware transactional memory to simplify memory management. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing*, pages 99–108. ACM, 2011.
- [51] Faith Ellen, Panagiota Fatourou, Joanna Helga, and Eric Ruppert. The amortized complexity of non-blocking binary search trees. In *Proc. of ACM Symp. on Princ. of Distr. Comp., PODC ’14*, pages 332–340, New York, NY, USA, 2014. ACM.

- [52] Faith Ellen, Panagiota Fatourou, Eric Ruppert, and Franck van Breugel. Non-blocking binary search trees. In *Proceedings of the 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, PODC '10, pages 131–140, 2010. Full version available as Technical Report CSE-2010-04, York University.
- [53] Faith Ellen, Yossi Lev, Victor Luchangco, and Mark Moir. Snzi: Scalable nonzero indicators. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, pages 13–22. ACM, 2007.
- [54] Jason Evans. A scalable concurrent malloc (3) implementation for freebsd. In *Proc. of the BSDCan Conference, Ottawa, Canada*, 2006.
- [55] Keir Fraser. *Practical lock-freedom*. PhD thesis, University of Cambridge, 2004.
- [56] Keir Fraser and Tim Harris. Concurrent programming without locks. *ACM Trans. Comput. Syst.*, 25(2), May 2007.
- [57] Anders Gidenstam, Marina Papatriantafilou, Håkan Sundell, and Philippas Tsigas. Efficient and reliable lock-free memory reclamation based on reference counting. *Parallel and Distributed Systems, IEEE Transactions on*, 20(8):1173–1187, 2009.
- [58] D.H. Greene and D.E. Knuth. *Mathematics for the analysis of algorithms* (2nd). Progress in computer science. Birkhäuser, 1982.
- [59] Leo J. Guibas and Robert Sedgewick. A dichromatic framework for balanced trees. In *Proc. 19th IEEE Symposium on Foundations of Computer Science*, pages 8–21, 1978.
- [60] Bernhard Haeupler, Siddhartha Sen, and Robert E. Tarjan. Rank-balanced trees. *ACM Trans. Algorithms*, 11(4):30:1–30:26, June 2015.
- [61] Timothy L. Harris. A pragmatic implementation of non-blocking linked-lists. In *Proceedings of the 15th International Conference on Distributed Computing*, DISC '01, pages 300–314, 2001.
- [62] Timothy L. Harris, Keir Fraser, and Ian A. Pratt. A practical multi-word compare-and-swap operation. In *Proceedings of the 16th International Conference on Distributed Computing*, DISC '02, pages 265–279, 2002.
- [63] Thomas E Hart, Paul E McKenney, Angela Demke Brown, and Jonathan Walpole. Performance of memory reclamation for lockless synchronization. *Journal of Parallel and Distributed Computing*, 67(12):1270–1285, 2007.
- [64] Meng He and Mengdu Li. Deletion without rebalancing in non-blocking binary search trees. In *Proceedings of the 20th International Conference on Principles of Distributed Systems*, 2016.
- [65] Steve Heller, Maurice Herlihy, Victor Luchangco, Mark Moir, William N. Scherer, and Nir Shavit. A lazy concurrent list-based set algorithm. In *Proceedings of the 9th International Conference on Principles of Distributed Systems*, OPODIS'05, pages 3–16, Berlin, Heidelberg, 2006. Springer-Verlag.
- [66] Maurice Herlihy, Victor Luchangco, Paul Martin, and Mark Moir. Nonblocking memory management support for dynamic-sized data structures. *ACM Transactions on Computer Systems (TOCS)*, 23(2):146–196, 2005.

- [67] Maurice Herlihy, Victor Luchangco, Mark Moir, and William N Scherer III. Software transactional memory for dynamic-sized data structures. In *Proceedings of the twenty-second annual symposium on Principles of distributed computing*, pages 92–101. ACM, 2003.
- [68] Maurice Herlihy and J. Eliot B. Moss. Transactional memory: Architectural support for lock-free data structures. In *Proc. 20th Annual International Symposium on Computer Architecture*, pages 289–300, 1993.
- [69] Maurice P. Herlihy and Jeannette M. Wing. Linearizability: a correctness condition for concurrent objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463–492, July 1990.
- [70] Shane V. Howley and Jeremy Jones. A non-blocking internal binary search tree. In *Proceedings of the Twenty-fourth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA ’12, pages 161–171, 2012.
- [71] Shou-Hsuan S. Huang. Height-balanced trees of order (β, γ, δ) . *ACM Trans. Database Syst.*, 10(2):261–284, June 1985.
- [72] Amos Israeli and Lihu Rappoport. Disjoint-access-parallel implementations of strong shared memory primitives. In *Proceedings of the Thirteenth Annual ACM Symposium on Principles of Distributed Computing*, PODC ’94, pages 151–160, 1994.
- [73] Lars Jacobsen and Kim S. Larsen. Variants of (a, b)-trees with relaxed balance. *Int. J. Found. Comput. Sci.*, 12(4):455–478, 2001.
- [74] Prasad Jayanti and Srdjan Petrovic. Efficiently implementing a large number of LL/SC objects. In *Proc. 9th International Conference on Principles of Distributed Systems*, volume 3974 of *LNCS*, pages 17–31, 2005.
- [75] Richard Jones and Rafael Lins. Garbage collection: Algorithms for automatic dynamic memory management. *John Wiley & Sons Ltd., England*, 1996.
- [76] Michael Kerrisk. *The Linux programming interface*. No Starch Press, 2010.
- [77] Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998.
- [78] Sanjeev Kumar, Michael Chu, Christopher J Hughes, Partha Kundu, and Anthony Nguyen. Hybrid transactional memory. In *Proceedings of the eleventh ACM SIGPLAN symposium on Principles and practice of parallel programming*, pages 209–220. ACM, 2006.
- [79] H.T. Kung and Philip L. Lehman. Concurrent manipulation of binary search trees. *ACM Transactions on Database Systems*, 5(3):354–382, 1980.
- [80] Klaus Küspert. Storage utilization in B*-trees with a generalized overflow technique. *Acta Informatica*, 19(1):35–55, 1983.
- [81] Kim S. Larsen. Avl trees with relaxed balance. In *Parallel Processing Symposium, 1994. Proceedings., Eighth International*, pages 888–893. IEEE, 1994.

- [82] Kim S. Larsen. Amortized constant relaxed rebalancing using standard rotations. *Acta Informatica*, 35(10):859–874, 1998.
- [83] Kim S. Larsen. AVL trees with relaxed balance. *Journal of Computer and System Sciences*, 61(3):508–522, December 2000.
- [84] Kim S. Larsen and Rolf Fagerberg. B-trees with relaxed balance. In *Proc. 9th International Symposium on Parallel Processing*, pages 196–202, 1995.
- [85] Kim S. Larsen and Rolf Fagerberg. Efficient rebalancing of b-trees with relaxed balance. *Int. J. Found. Comput. Sci.*, 7(2):169–186, 1996.
- [86] Kim S. Larsen, Thomas Ottmann, and Eljas Soisalon-Soininen. Relaxed balance for search trees with local rebalancing. *Acta Informatica*, 37(10):743–763, 2001.
- [87] Kim S. Larsen, Eljas Soisalon-Soininen, and Peter Widmayer. Relaxed balance through standard rotations. In Frank Dehne, Andrew Rau-Chaplin, Jörg-Rüdiger Sack, and Roberto Tamassia, editors, *Algorithms and Data Structures*, volume 1272 of *Lecture Notes in Computer Science*, pages 450–461. Springer Berlin Heidelberg, 1997.
- [88] Doug Lea. Concurrentskiplistmap. <http://fuseyism.com/classpath/doc/java/util/concurrent/ConcurrentSkipListMap-source.html>.
- [89] Hyonho Lee. Fast local-spin abortable mutual exclusion with bounded space. In *Principles of Distributed Systems*, pages 364–379. Springer, 2010.
- [90] Yossi Lev, Mark Moir, and Dan Nussbaum. Phtm: Phased transactional memory. In *Workshop on Transactional Computing (Transact)*, 2007.
- [91] Yujie Liu, Tingzhe Zhou, and Michael Spear. Transactional acceleration of concurrent data structures. In *Proc. of 27th ACM Sym. on Parallelism in Algorithms and Arch.*, SPAA ’15, pages 244–253, New York, NY, USA, 2015. ACM.
- [92] Victor Luchangco, Mark Moir, and Nir Shavit. Nonblocking k -compare-single-swap. *Theory of Computing Systems*, 44(1):39–66, January 2009.
- [93] Darko Makreshanski, Justin Levandoski, and Ryan Stutsman. To lock, swap, or elide: on the interplay of hardware transactional memory and lock-free indexing. *Proceedings of the VLDB Endowment*, 8(11):1298–1309, 2015.
- [94] Virendra Jayant Marathe and Mark Moir. Toward high performance nonblocking software transactional memory. PPoPP ’08, pages 227–236, New York, NY, USA, 2008. ACM.
- [95] Alexander Matveev. Private communication.
- [96] Paul E. McKenney and John D Slingwine. Read-copy update: Using execution history to solve concurrency problems. In *Parallel and Distributed Computing and Systems*, pages 509–518, 1998.

- [97] Maged Michael. Hazard pointers: Safe memory reclamation for lock-free objects. *Parallel and Distributed Systems, IEEE Transactions on*, 15(6):491–504, 2004.
- [98] Maged M. Michael. High performance dynamic lock-free hash tables and list-based sets. In *Proceedings of the Fourteenth Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA ’02, pages 73–82, 2002.
- [99] Maged M. Michael. Hazard pointers: Safe memory reclamation for lock-free objects. *IEEE Trans. Parallel Distrib. Syst.*, 15(6):491–504, June 2004.
- [100] Mark Moir. Practical implementations of non-blocking synchronization primitives. In *Proceedings of the Sixteenth Annual ACM Symposium on Principles of Distributed Computing*, PODC ’97, pages 219–228, 1997.
- [101] Aravind Natarajan and Neeraj Mittal. Fast concurrent lock-free binary search trees. In *Proceedings of the 19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP ’14, pages 317–328, 2014.
- [102] Aravind Natarajan, Lee Savoie, and Neeraj Mittal. Concurrent wait-free red black trees. In *Proc. 15th International Symposium on Stabilization, Safety and Security of Distributed Systems*, volume 8255 of *LNCS*, pages 45–60, 2013.
- [103] Andrew T Nguyen. *Investigation of Hardware Transactional Memory*. PhD thesis, Massachusetts Institute of Technology, 2015.
- [104] Otto Nurmi and Eljas Soisalon-Soininen. Chromatic binary search trees: A structure for concurrent rebalancing. *Acta Informatica*, 33(6):547–557, 1996.
- [105] Otto Nurmi, Eljas Soisalon-Soininen, and Derick Wood. Relaxed avl trees, main-memory databases and concurrency. *International journal of computer mathematics*, 62(1-2):23–44, 1996.
- [106] Ravi Rajwar and James R Goodman. Speculative lock elision: Enabling highly concurrent multithreaded execution. In *Proceedings of the 34th annual ACM/IEEE international symposium on Microarchitecture*, pages 294–305. IEEE Computer Society, 2001.
- [107] Ravi Rajwar and James R Goodman. Transactional lock-free execution of lock-based programs. *ACM SIGOPS Operating Systems Review*, 36(5):5–17, 2002.
- [108] Arunmoezhi Ramachandran and Neeraj Mittal. A fast lock-free internal binary search tree. In *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, ICDCN ’15, pages 37:1–37:10, New York, NY, USA, 2015. ACM.
- [109] Torvald Riegel, Patrick Marlier, Martin Nowack, Pascal Felber, and Christof Fetzer. Optimizing hybrid transactional memory: The importance of nonspeculative operations. In *Proceedings of the 23rd ACM Symposium on Parallelism in Algorithms and Architectures*, pages 53–64. ACM, 2011.
- [110] Arnold L. Rosenberg and Lawrence Snyder. Compact B-trees. In *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’79, pages 43–51, New York, NY, USA, 1979. ACM.

- [111] Martin Schoeberl and Wolfgang Puffitsch. Nonblocking real-time garbage collection. *ACM Transactions on Embedded Computing Systems (TECS)*, 10(1):6, 2010.
- [112] ML Scott and WJ Bolosky. False sharing and its effect on shared memory performance. In *Proceedings of the USENIX Symposium on Experiences with Distributed and Multiprocessor Systems (SEDMIS)*, page 57, 1993.
- [113] Niloufar Shafiei. Non-blocking patricia tries with replace operations. In *Proceedings of the 2013 IEEE 33rd International Conference on Distributed Computing Systems*, ICDCS '13, pages 216–225, 2013.
- [114] Nir Shavit and Dan Touitou. Software transactional memory. *Distributed Computing*, 10(2):99–116, February 1997.
- [115] Michael Spiegel and Paul F. Reynolds, Jr. Lock-free multiway search trees. In *Proc. 39th International Conference on Parallel Processing*, pages 604–613, 2010.
- [116] B. Srinivasan. An adaptive overflow technique to defer splitting in B-trees. *The Computer Journal*, 34(5):397–405, 1991.
- [117] Håkan Sundell. Wait-free multi-word compare-and-swap using greedy helping and grabbing. *International Journal of Parallel Programming*, 39(6):694–716, December 2011.
- [118] Shahar Timnat, Maurice Herlihy, and Erez Petrank. A practical transactional memory interface. In *Euro-Par 2015: Parallel Processing*, pages 387–401. Springer, 2015.
- [119] Jyh-Jong Tsay and H.-C. Li. Lock-free concurrent tree structures for multiprocessor systems. In *Proc. International Conference on Parallel and Distributed Systems*, pages 544–549, 1994.
- [120] John Turek, Dennis Shasha, and Sundeep Prakash. Locking without blocking: Making lock based concurrent data structure algorithms nonblocking. In *Proc. 11th ACM Symposium on Principles of Database Systems*, pages 212–222, 1992.
- [121] John D. Valois. Lock-free linked lists using compare-and-swap. In *Proceedings of the Fourteenth Annual ACM Symposium on Principles of Distributed Computing*, PODC '95, pages 214–222, 1995.