Concurrent Wait-Free Red Black Trees

David Ferguson and Jacob Spigle

Abstract— **Our re-implementation of Natarajan, Savoie, and Mittal's wait-free algorithm [1] seeks to not only imitate the concurrent data structure presented, but also transform it into a transactional data structure using the RSTM library. We are implementing a concurrently managed red-black tree using wait-free algorithms designed and presented in [1]. We will be writing these algorithms using C++ as our programming language. Concurrency during manipulation of a tree data structure is not plausible without additional (and creative) data structures because of the multiple instructions that rotations perform during the re-balancing of the tree. Using the techniques outlined in [1], we will have a concurrent algorithm that makes progress, is linearizable, and correct. Experiments run by the authors of this implementation prove that their solution provides "significantly better performance"[1] than other attempts that preceded it, including both attempts at concurrency and lock-based implementations.**

I. Introduction

This wait-free implementation of the Red Black tree data structure boasts search(), insert(), update(), and delete() functions, all executed utilizing single-word compare-and-swap instructions. The data structure's concurrent implementation employs the use of "windows", which are overlapping snapshots of the current state of the Red Black tree within the scope of the windows' root node. Windows are copied from a global window, then edited locally before being pushed back to the global view of the Red Black tree itself, and pushing a modified window into the windows' origin will result in a correct, linearizable solution. This is because the window itself can be atomically swapped, where rotations are done inside a modified window, and using a single-word compare-and-swap, are placed back into the node where the window originated. This solution also strives for optimal concurrency by introducing an array that holds pending instructions (using the *announce* array) and decides whether or not a thread will assist by checking for conflicts with it's own update operation (using a *gate* variable, given to each record in the tree). A modify operation also is tasked to help another operation, choosing which to help in a round robin manner, so that it may also help during a search operation to ensure that it may eventually terminate [1]. This is necessary because this implementation avoids copying windows unnecessarily and instead traverses to the next window's root node when such a transaction would occur. Because this skipping occurs during a search operation, it would be possible for the operation to be overtaken,, These additions to the traditional sequential Red Black tree allow for an efficient algorithm that has outperformed other attempts at this implementation of the concurrent wait-free Red Black tree data structure.

Related Work: A previous attempt at creating a concurrent Red-Black tree data structure was implemented by Kim *et. al.* [3]. It is built upon Ma's [4] work with specifically insertions upon a lock-free Red-Black tree data structure, and [3] extends [4] to include both modification and deletion techniques. There have been a few more recent attempts at this wait-free implementation of manipulating red-black trees concurrently. Notably, in 2014, there was a thesis written proposing that instead of a Top-Down approach to obtaining ownership of nodes within the tree, that working from Bottom-Up approach would "[allow] operations interested in completely disparate portions of the tree to execute entirely uninhibited"[2].

Contributions: In this paper we seek to present our re-implementation of the Concurrent Wait-Free Red-Black Tree data structure presented in [1]. We also wish to present our solution to transforming [1] into a Transactional data structure by using the RSTM C++ Library presented by Marathe *et. al.* [6]. To do this, we will also need to employ POSIX Threads (pthreads) in our C++ code.

II. Preliminaries

The Concurrent Wait-Free Red Black Tree algorithm that was presented by [1] is derived from a few other presented algorithms. [1] presents a conglomeration of these algorithms in their modified state that creates the presented concurrent data structure.

Tsay & Li's Wait-Free Framework for Concurrent Tree-Based Data Structures

Tsay & Li's [5] TL-Framework is a construction that can be applied to any tree-based data structure attempting concurrency, provided the hardware supports load-linked, store-conditional and check-valid synchronization primitives. [5] presents the ability to perform wait-free top-down operations by utilizing 'windows', which are described as "rooted subtree[s] of [a] tree structure, that is, a small, contiguous piece of the tree" [1]. Essentially, as a process progresses through a tree while attempting to perform its' given operation, the process will continuously create copies of each node when

[1]D. Ferguson is a student at Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida, 32816-2450

[2]J. Spigle is a student at Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida, 32816-2450

utilizing this framework. Note that nodes inside the TL-Framework are outlined as dual-structured nodes. This means that a single node is composed of two nodes, a pointer node and a data node. As the process traverses the tree, it will be creating a 'window' accessible only by the local process. This is where operations will be able to perform their steps in what is called a 'window transaction', whether it be an insertion, update, or deletion, of a node, and switch the changed window back into the tree afterwards in a single atomic step, by simply changing the address that our pointer node is referencing to the new 'root' data node of the W_L we have modified. We call the current windows of the tree W_G for window global. W_G is the only window accessible from the root node of the tree. It is important to say here that [**?**] found some problems with a practical implementation of the TL-Framework and made a few adjustments to create their own backbone using what they call the MTL-Framework. With a firm grasp on the concept of windows, we dive into the steps that must be taken to perform any operation when utilizing the MTL-Framework as described in [1]:

1) Explore-Help-And-Copy & Transform-And-Lock: In these two steps, some process *p* traverses the Red Black Tree using the root node's W_G and each subsequent windows' global window. As *p* traverses W_G, it creates its' own localized copy of the window, we entitle its' window local, or W_L. Our process *p* is the only process able to access W_L, so we perform window transactions sequentially inside the window, thus modifying the window into what we call W_{ML} for window modified local. It is important to note that operations are owned by nodes, not processes. This is important because if some node *X*, traversed to by our process *p*, is owned by some operation , our process *p* will then 'help' by performing a window transaction in that window on 's behalf.

2) Install: The install step of performing operations in the TL-Framework are what bring our changes into our W_G view as well as allow us to continue forward traversing the tree when necessary. Because of the dual nature of our nodes in the TL-Framework, once a window transaction has completed we simply replace the current W_G with our modified W_{ML}. This is done in a single atomic instruction because the only change is where our windows' root pointer node is referencing. Upon success, our W_{ML} has now been inserted to the tree and becomes our new W_G, which is accessible from the root node and visible to all processes. We can note here that the old W_G is no longer accessible from the root, because it has been switched out, but all the references to those nodes are currently retained. As a side note, this brings us to another feature of the TL-Framework that we are utilizing, the idea of *active* and *passive* nodes. *Active* nodes are nodes that reside in the global window of the tree, while *passive* nodes are nodes that have not been dereferenced but also are not accessible from the root of our tree. This is important because if there is another operation attempting to perform in that window such as search, we must wait until it is safe to dereference the node, else we will incur a segmentation fault. When *p* has inserted W_{ML} into the W_G, it has also updated the operations' ownership from the root of the current window to the root of the next window. Because this happens technically in the W_{ML} and *p* is the sole accessor, the position is updated atomically.

3) Announce: Here, *p* updates the operation's new window position by updating the table (MT for modify table for modify operations, ST for search table for a search operation) entry for the process that originally injected the operation. Any process can do this because the operation's new window is now in the global eye and the information to do so is kept in the root node of the current window.

A. Red Black Trees

A red black tree is a specific tree structure and is a type of self-balancing binary search tree. It uses key-value pairs, where the value can be anything and the key is the variable that decides a node's place inside the tree data structure. This allows us to create template classes for the operations and other data structures to use inside our tree, so any datatype can be organized inside the tree as long as the desired structure is kept correct using the necessary key value. A red black tree supports four operations in our implementation, *search, insert, update,* and *delete*. The *search* operation finds the key and returns its' associated value inside the tree. *Insert* adds some key-value pair into the tree if it is not already present, and if it is present then the operation becomes an *update* operation, and updates the inserted key with the new value. *Delete* removes a given key from the red black tree if it is present. It is also important to note that in our implementation we refer to *insert, update,* and *delete* operations as *modify* operations. The reason for this is explained later, but the basic need is to separate *search* from the other available operations in our red black tree.

Tarjan's Top-Down Operations

[1] utilizes [7] to perform all of its' operations inside the Concurrent Wait-Free Red Black Tree. Tarjan's algorithms for Insertion and Deletion are necessary here because with the introduction of windows, we lose the ability to perform in a bottom-up manner globally across the W_G space. Tarjan allows the red black tree to work top-down in the global structure, but run window transactions bottom-up in W_L, because those operations are being run locally. The most important thing to note about Tarjan's work is the idea of an *external* red black tree. In this tree, all the tree's data is accessible in the leaf nodes. Any internal node has a correlating leaf node where the data is stored.

1) Insertion: Inserting into an *external* red black tree performs the following steps. Following the access path (the path from our tree's root to the node that was inserted/deleted) from some current node *X*, we make *X* black if it is red or its' children to black if both are red, and we look out for a few cases:

- *(a)* we reach an external node, in which case we create an internal node in its' place, with its' children being the original external node and the incoming inserted node, and push the key of the smallest of the two into the new

internal node. Note that this is also how we enforce the idea of the *external* red black tree, because upon insertion into a node's left or right child, the external node itself will also be copied into the internal node created.
- *(b)* we reach a black node with a black child, in which case we replace the current node with its' parent before repeating the internal node creation and population outlined in *(a)*.
- *(c)* we encounter four black nodes, each with two red children, in succession. Here we color the last black node red and its' children black, then insert the node into the tree using the general step outlined in *(a)*, before moving upwards inside the window to adjust any imbalances before inserting back into the global window.

2) Deletion: Deletion in a top-down fashion is similar to our insertion method, with a few key changes. For one, we have to use something called a *short node* definition, so whenever our access path does not have the correct number of black nodes leading to an external node we know that this needs to be adjusted at that root node before pushing back into our global window. For deletion, we make the current node *X* our root, and color the root red if it is black with two black children. Then we traverse down the access path and look for:

- *(a)* an external node, in which case we found our item and we replace the external node's parent with its' sibling. Note that when this step has concluded, if the node we just replaced was black then we have a *short node* situation, so we then traverse up the window, converting each short node's sibling to red and creating a *short node* in the parent. To reconcile this, we perform one of the operations shown in [7]'s Figure 5.
- *(b)* a node that is red, has a red child, or has a red grandchild. Here we replace such a node with our current node *X* before performing the general step outlined in *(a)*.
- *(c)* three black nodes with black children as well as black grandchildren in succession. In this case we color the bottom-most node and its' sibling red (making its' parent a *short node*), and follow the steps specific to getting rid of the *short node* condition outlined in *(a)* to get rid of the *short node*. Then we replace our *X* with the bottom-most node and repeat steps in *(a)*.

III. Concurrent Wait-Free Red Black Trees

[1] presents an algorithm for creating and maintaining a concurrent wait-free red black tree data structure. This section will outline the changes and additions that were necessary to create this algorithm, and later we will expand on our own implementation of the data strucuture.

A. Data Structures Used

[1] utilizes 5 main data structures within their algorithm, outlined below:

- *Nodes:* For "nodes" we recognize the dual structure that is shown in [5] (TL-Framework). Nodes here are made up of both pointer nodes and data nodes.
  - *Pointer Node:* A pointer node is a single word that holds a tag, or *Flag*, that indicates if the node is FREE or OWNED, as well as the address of the data node that it points to.
  - *Data Node:* A data node holds all the information relevant to a nodes as it would be in a red black tree. *Color, key, left,* and *right* define a data node and its' position, with *left* and *right* pointing to other data nodes' pointer nodes. Two additional fields are the *valData*, which is utilized to keep an external record of a nodes value, and *opData*, which holds information of the operation that is being performed in the root window expanding from this node. The *next* field indicates the next operation location after this nodes' operation has been completed, and thus it is a single word that holds this operation's progress, *status* (WAITING, IN_PROGRESS, COMPLETED), as well as *move* which is the address of the next node that holds the next operation.
- *Value Record:* The value record holds a data node's value external from the tree, because the TL-Framework calls for the copying of nodes, and we need to keep up to date with the value accordingly. It simply holds the actual *value* and a *gate* variable (used for deciding which process' operation will perform an *update* on the value record.
- *Operation Record:* An operation record holds the information necessary for a given process to help another process' injected operation by performing an instruction laid out in this data structure inside the 'root' (root of a given window) data node that this operation record spawns from. It holds *type, key, value* for the instruction, the *pid* (process ID) that spawned this operation record, as well as a *state* single word variable that holds the *status* of the operation being performed (WAITING, IN_PROGRESS, COMPLETED) and the *position* which is the current window in which this operation is being performed. Note that *position* could also be the address of a value record when referring to a *search* or *update* operation.
- *Tables:* [1] also utilizes two table data structures for keeping track of the *search* and *modify* operations being run and what process they have spawned from. Both the search table (ST) and the modify table (MT) are 2D arrays that hold the opData that was last injected by any process. Both tables have size of NUM_THREADS and thus allow us to choose which process to help when traversing the red black tree by incrementing through the array in a round robin manner. This will be explained in our implementation later as well.

B. Modified Search Operation

If we utilized the MTL-Framework for each operation we would have every process helping every operation in its' path. There are some cases where this behavior is unnecessary and actually very strange when thinking about it in a concurrent setting. The first thought is our *search* operation. Using MTL, *search* would occupy a node while traversing a window, and other processes that traverse to that node would then be expected to help the search function. If we were to have two search functions attempting to run in the same part of the tree, those search functions would be in conflict with each other. [**?**] observes that with the new MTL-Framework, a window transaction is atomic because of the compare and swap that occurs on each pointer node that occupies the W_G space to point to the new modified W_{ML}, all in a single atomic step. So our operations will only be able to view all or none of the modifications passed into the global space. Another observation is that because these windows are not pushed back into the global space until all transformations have completed, the resulting red black tree will always show correct and legal. So moving forward, [1] forgoes the process of helping other operations when performing *search*, and is important because we also can now ignore the copying of nodes during a *search* operation which make the MTL-Framework so expensive. Now the *search* operation can simply traverses the tree unobstructed. It is important to mention that *search* still runs the risk of encountering *passive* nodes in this implementation, but we call these Search-Misses (finds the node, but the node is no longer part of the tree), and utilizing this new *search* is still valuable.

C. Modified Modify Operations

The overhead of a modify operation can be very expensive when performing insertion and deletions to the bottom of a tree that potentially have already been run by the time we reach that window. There are a few ways we can reduce the cost of this.

Because of our new *fast-search* operation, we can perform preemptive checks to the state of the tree before we attempt a *modify* operation. Doing this, our *insert/update* and *delete* operations have phases when they are performed:

- *Insert:*
  - *P1: Fast-Search* traverses the tree to see if the key already exists within the tree. Upon finding it, the *insert* operation simply updates the value record that was returned by *search*.
  - *P2:* If *fast-search* did not locate the key, *insert* will perform normally and run the expensive MTL-Framework version.
  - *P3:* If the *insert* operation locates the key in the tree, it then updates the associated value record using Chuong *et al.*'s wait-free algorithm [8]. This is possible due to a few modifications to [**?**]: all value records share one *announce* array (used for processes to announce their operations to one another), however, each value record holds its' own *gate* variable (used to force processes to agree on the next operation to update this value record). In the modified version, Chuong *et al.*'s algorithm will only help update a value record if it is in direct conflict with its' own update operation. This is possible because the current process would then have to store the address of the record in the *announce* array, and then use the value record's *gate* to decide which operation updates the value record.
- *Delete:*
  - *P1: Fast-Search* traverses the tree to see if the key has been deleted from the tree. If it has, then another *delete* operation has already completed and we can stop.
  - *P2:* If *fast-search* was able to locate the key, *delete* will perform normality and run the expensive MTL-Framework version.

With utilizing [7]'s top-down operations, we will see the invariants required for balance when traversing the tree and inside certain windows. Because they are already being followed, it is a waste to acquire ownership of the root node of a window, copy the nodes using the MTL-Framework, then obtain ownership of the subsequent window's root node before dropping the previous window's root. Nothing has happened in this instance, and it is called a *trivial transaction*. We are able to skip these transactions entirely by skipping a window and sliding down to the next root by obtaining ownership of it. Note that acquiring the next windows' root is not atomic (an instruction to acquire ownership, another to release ownership), and thus we could encounter an instance where our *fast-search* has been overtaken. This is a problem because in the case that *search* is continuously overtaken, its' key moving farther and farther down the tree for infinity, *search* may never terminate. To ensure that it does, [1] includes the stipulation that a *modify* operation may have to help a *search* operation complete. It does this by helping each process in a round-robin manner.

IV. Our Re-Implementation

Our goal was to recreate the concurrent wait-free red black tree algorithm presented in [1] using the C++ programming language, while also using the POSIX Threads (pthread.h library) for our our creation & management of the threads involved. Upon completion of bringing the algorithms presented in [1] into C++, we then transformed our code to utilize the RSTM library (a software transactional memory library that allows only one change in a critical piece(s) of memory for any *n* threads that currently access that piece of memory's *read/write* set).

A. Details of Code

```
template <class V>
class ValueRecord
{
public:
```

```
        V *mValue;
        uint32_t mGate;

        ValueRecord(V *value, uint32_t gate)
        {
                mValue = value;
                mGate = gate;
        }
};
```

```
#include <memory>
#include <pthread.h>
#include <climits>

#define PointerNode PackedPointer
```

B. Changes

C. Obstacles

V. Proofs

A. Progress Guarantee

B. Correctness Condition

C. Synchronization Techniques

VI. Experimentation

A. Experimental Setup

We utilized an Intel 4-core i5 chip running on a 64-bit Linux Ubuntu 16.04 Virtual Machine given 4GB of RAM hosted on a 64-bit Windows 10 operating system for conducting experiments with our transactional data structure. Our implementation was written in C++ and utilizes POSIX threads for thread creation and management as well as an RSTM library for calls to memory and dynamic memory allocation. For each of our experiments we took the following parameters into account:

- *Number of Operations:* We kept the number of operations run on our red black tree static throughout all of our experimentation, keeping it at 500,000 operations.
- *Distribution of Operations:* We varied the distribution of each of our operations (*search, insert/update,* and *delete*) as such:
  - *(a) Write Dominant:* In this scenario, we have 50% of our operations running *insert/update* operations, while 25% run *delete* operations and the remaining 25% running our *search* operations.
  - *(b) Read Dominant:* In this scenario, we have 50% of our operations running *search* operations, while 25% run *insert/update* operations and the remaining 25% running our *delete* operations.
  - *(c) Mixed Workload:* In this scenario, we have an equal 33% split between our *search, insert/update* and *delete* operations.
- *Degrees of Contention:* The number of threads in each test phase of our experimentation varied from 1, 2, 4, and 8 threads sharing the same data structure, each running our 500,000 operations.

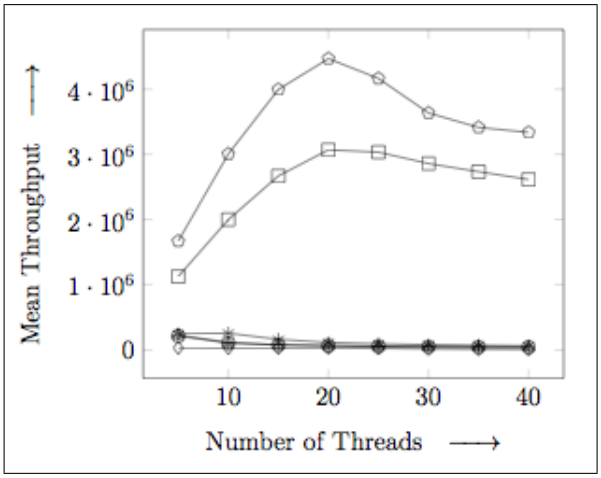

Fig. 1. (a) Write Dominant — Concurrent

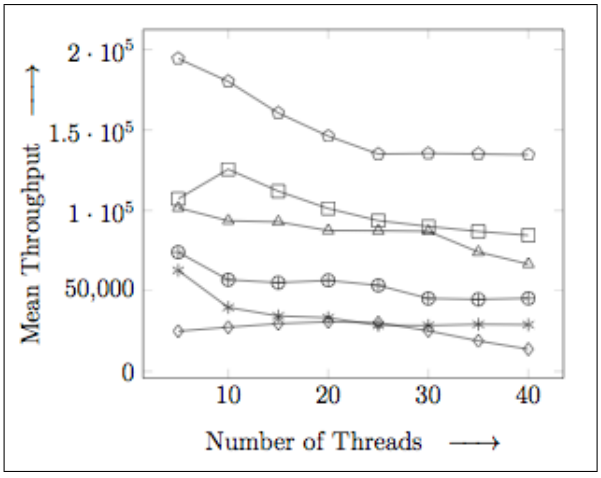

Fig. 2. (a) Write Dominant — Transactional Operations : 2

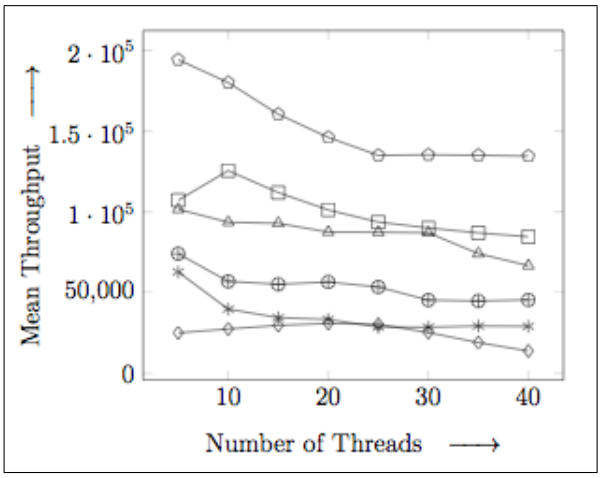

Fig. 3. (a) Write Dominant — Transactional Operations: 3

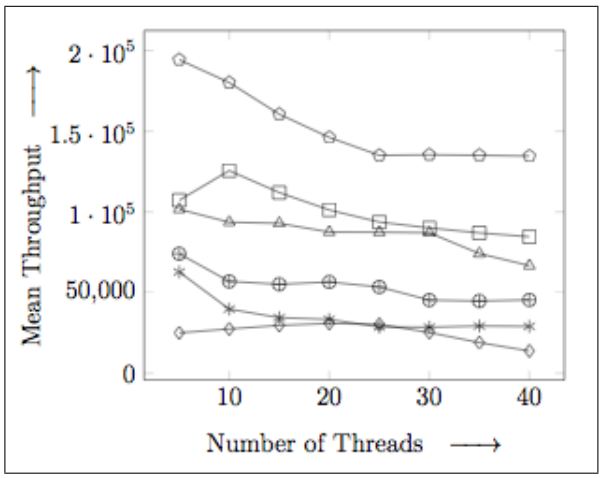

Fig. 4. (b) Search Dominant — Concurrent

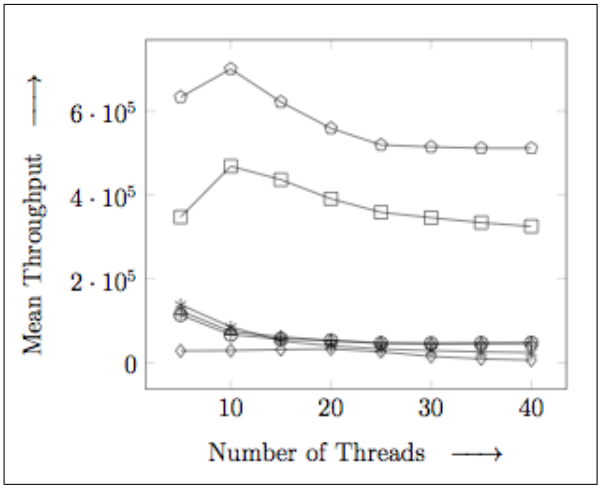

Fig. 7. (c) Mixed Workload — Concurrent

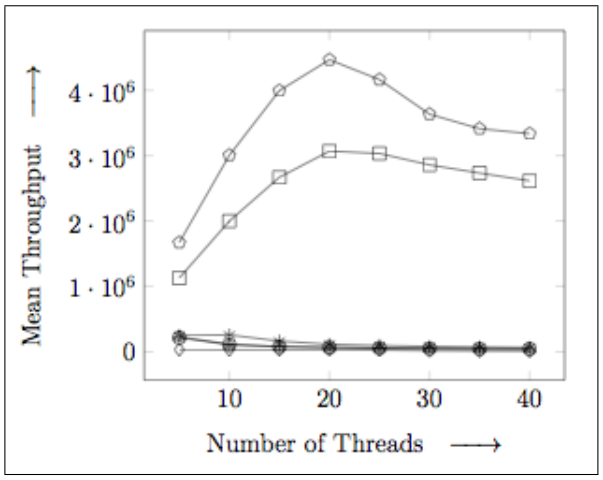

Fig. 5. (b) Search Dominant — Transactional Operations : 2

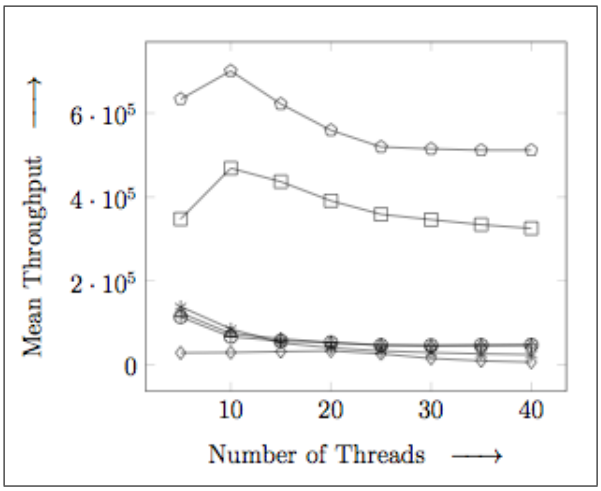

Fig. 8. (c) Mixed Workload — Transactional Operations : 2

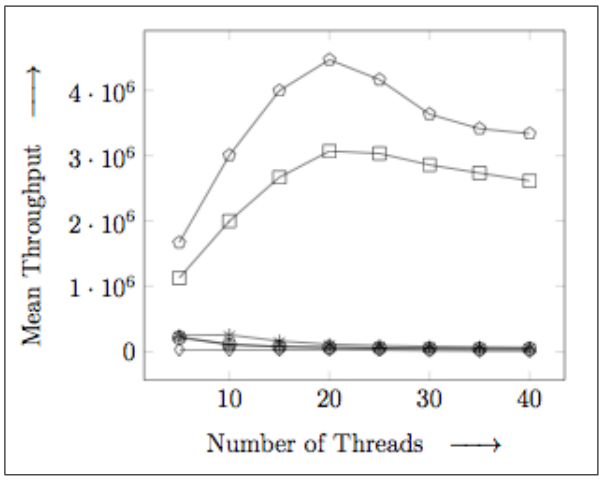

Fig. 6. (b) Search Dominant — Transactional Operations : 3

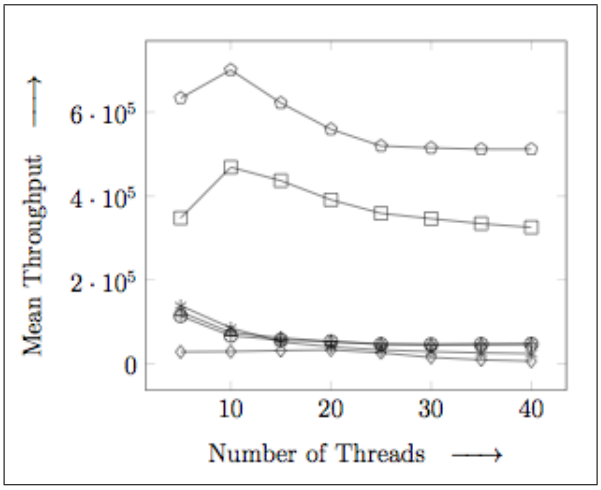

Fig. 9. (c) Mixed Workload — Transactional Operations : 3

B. Experimental Results

VII. CONCLUSIONS

Our presentation of both the concurrent and transactional data structure implementations deriving from [1] demonstrates the linearizibility of the resulting C++ program. Searching in the Red-Black tree using these algorithms will be significantly less expensive due to the lack of helping. Modify operations that are performed in separate windows are able to execute concurrently without loss of correctness at any point because of the single word compare and swap instructions used when swapping out windows during a window transaction.

APPENDIX

REFERENCES

[1] A. Natarajan, L. Savoie, & N. Mittal 2013. 'Concurrent Wait-Free Red Black Trees'. The University of Texas at Dallas, Richardson, TX 75080, USA.

[2] V. Kubushyn 2014. 'Concurrent Localized Wait-Free Operations on a Red Black Tree'. University of Nevada, Las Vegas.

[3] J. H. Kim, H. Cameron, & P. Graham 2006. 'Lock-Free Red-Black Trees Using CAS'. *Concurrency and Computation: Practice and Experience*, 1-40.

[4] J. Ma, 2003. 'Lock-Free Insertions on Red-Black Trees. MSc thesis. University of Manitoba.

[5] J.J. Tsay & H.C. Li 1994. 'Lock-Free Concurrent Tree Structures for Multiprocessor Systems'. *Parallel and Distributed Systems, 1994*, 554-549. IEEE.

[6] V.J. Marathe, M.F. Spear, C. Heriot, A. Acharya, D. Eisenstat, W.N. Scherer III & M.L. Scott. 'Lowering the Overhead of Nonblocking Software Transactional Memory'. Computer Science Department, University of Rochester.

[7] R.E. Tarjan 1985. 'Efficient Top-Down Updating of Red-Black Trees'. Computer Science Department, Princeton University, Princeton, NJ 08544.

[8] P. Chuong, F. Ellen, & V. Ramachandran 2010. 'A Universal Construction for Wait-Free Transaction Friendly Data Structures'. In: *Proceedings of the 22nd ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*. Thira, Santorini, Greece, pp. 335-344.