Efficient concurrent search trees using portable fine-grained locality

Phuong Hoai Ha, Otto J. Anshus, Ibrahim Umar

Abstract—Concurrent search trees are crucial data abstractions widely used in many important systems such as databases, file systems and data storage. Unlike other fundamental abstractions for energy-efficient computing, concurrent search trees should support both high concurrency and fine-grained data locality in a platform-independent manner. However, existing portable fine-grained locality-aware search trees such as ones based on the van Emde Boas layout (VEB-based trees) poorly support concurrent update operations while existing highly-concurrent search trees such as non-blocking search trees do not consider fine-grained data locality. In this paper, we first present a novel methodology to achieve both portable fine-grained data locality and high concurrency for search trees. Based on the methodology, we devise a novel locality-aware concurrent search tree called GreenBST. To the best of our knowledge, GreenBST is the first practical search tree that achieves both portable fine-grained data locality and high concurrency. We analyze and compare GreenBST energy efficiency (in operations/Joule) and performance (in operations/second) with seven prominent concurrent search trees on a high performance computing (HPC) platform (Intel Xeon), an embedded platform (ARM), and an accelerator platform (Intel Xeon Phi) using parallel micro-benchmarks (Synchrobench). Our experimental results show that GreenBST achieves the best energy efficiency and performance on all the different platforms. GreenBST achieves up to 50% more energy efficiency and 60% higher throughput than the best competitor in the parallel benchmarks. These results confirm the viability of our new methodology to achieve both portable fine-grained data locality and high concurrency for search trees.

Index Terms—Concurrent data abstractions, trees, energy efficiency, performance optimization, data locality, portability.

1 INTRODUCTION

As energy efficiency is emerging as one of the most important metrics in designing computing systems [26], [41], [45], [47], data should be organized and accessed in an energy efficient manner. Unlike conventional locality-aware data structures and algorithms that only concern whether the data is on-chip (e.g., in cache) or not (e.g., in DRAM), new energy-efficient data structures and algorithms must consider data locality in finer-granularity: where on chip the data is. It is estimated that for chips using the 10nm technology, the energy gap between accessing data in nearby on-chip memory (e.g., data in SRAM) and accessing data across the chip (e.g., on-chip data at the distance of 10mm), will be as much as 75x (2pJ versus 150pJ), whereas the energy gap between accessing on-chip data and accessing off-chip data (e.g., data in DRAM) will be only 2x (150pJ versus 300pJ) [18]. Therefore, in order to construct energy efficient software systems, data structures and algorithms should support not only high parallelism but also fine-grained data locality [18].

Concurrent search trees are crucial data structures that are widely used as a back-end in many important systems such as databases (e.g., SQLite [29]), filesystems (e.g., Btrfs [38]), and schedulers (e.g., Linux’s Completely Fair Scheduler (CFS)), among others. However, existing highly concurrent search trees do not consider fine-grained locality. Non-blocking concurrent search trees (e.g., [14], [22], [23], [50]) and Software Transactional Memory (STM)-based search trees (e.g., [2], [12], [17], [21]) among others have been regarded as the state-of-the-art concurrent search trees. The prominent highly concurrent search trees widely used in several benchmark distributions are the concurrent red-black trees developed by Oracle Labs [21] and the concurrent AVL trees developed by Stanford [12]. These highly concurrent search trees, however, do not take into account fine-grained data locality.

Existing fine-grained locality-aware search trees poorly support concurrency and are usually platform-dependent, limiting their portability across different platforms. For example, Intel Fast [30] and Palm [59] are optimized for a specific platform. Concurrent B-trees (e.g., B-link tree [32]) only perform well if their parameter B is chosen correctly. More recent research on system- and database- oriented concurrent search trees [15], [33], [34], [35], [37], [45] has produced some excellent examples of cache-conscious concurrent search trees. Unfortunately, none of the research addresses the issue of portability, because they were mostly developed and evaluated for a specific platform.

Portable fine-grained locality can be theoretically achieved using cache-oblivious (CO) methodology [25]. In the CO methodology, an algorithm is categorized as cache-oblivious for a two-level memory hierarchy if it has no variables that need to be tuned with respect to hardware parameters, such as cache size and cache-line length, in order to optimize its cache complexity, assuming that the optimal off-line cache replacement strategy is used. If a cache-
oblivious algorithm is optimal for an arbitrary two-level memory, the algorithm is also optimal for any adjacent pair of available levels of the memory hierarchy \[10, 20\]. Therefore, cache-oblivious algorithms are expected to be locality-optimized irrespective of variations in memory hierarchies, enabling portable fine-grained locality.

Portable fine-grained data locality for sequential search trees can be theoretically achieved using the van Emde Boas (vEB) layout \[24\], which is analyzed using ideal cache (CO) models \[25\]. The vEB layout has inspired several cache-oblivious sequential search trees such as cache-oblivious B-trees \[7, 8\] and cache-oblivious binary trees \[11\]. The vEB-based trees recursively arrange related data in contiguous memory locations, minimizing data transfer between any two adjacent levels of the memory hierarchy (see Section 2.4 for details).

However, existing vEB-based trees poorly support concurrent update operations and have high overhead and large memory footprints. Inserting or deleting a node may result in relocating a large part of the tree in order to maintain the vEB layout. Bender et al. \[9\] discussed the problem and provided important theoretical designs of concurrent vEB-based B-trees. Nevertheless, we have found that these designs are not very efficient in practice due to the actual overhead of maintaining necessary pointers as well as their large memory footprint (see Section 8.3).

Our practical insight is that it is unnecessary to keep the entire vEB-based tree in a single contiguous block of memory. In fact, allocating a contiguous block of (virtual) memory for a vEB-based tree does not guarantee a contiguous block of physical memory. Modern OSes and systems utilize different sizes of continuous physical memory blocks, for example, in the form of pages and cache lines. A contiguous block in virtual memory can be translated into several pages with gaps in physical memory (e.g., RAM); a page can be cached by several cache lines with gaps at any level of cache. This is one of the motivations for our new bounded ideal cache model (see Section 3.2). The upper bound on the contiguous block size can be obtained easily from any system (e.g., page-size), which is platform-independent. In fact, the memory transfer complexity of our search operation in the new bounded ideal cache model is independent of the upper bound values (see Lemma 3.1 in Section 3.2).

1.1 Our contributions

In this paper, we investigate whether it is practical to achieve both fine-grained data locality and portability in concurrent search trees and if so, whether portable fine-grained data locality actually improves energy efficiency and performance compared with conventional coarse-grained data locality used in B-trees. To the best of our knowledge, there is no previous experimental study of how portable fine-grained data locality actually influences energy efficiency and performance in concurrent search trees. Such a study is necessary nowadays when multilevel memory hierarchies are becoming more prominent in commodity systems. Modern CPUs tend to have at least three levels of caches.

Our contributions are fourfold:

1) We have devised a new bounded ideal cache model (or BCO) (see Section 2) and a novel concurrency-aware vEB layout (or CvEB) that makes the vEB layout suitable for highly-concurrent data structures with update operations (see Section 5).

2) Based on the new concurrency-aware vEB layout, we have devised a new portable fine-grained locality-aware concurrent search tree called GreenBST (see Sections 2 and 5). To the best of our knowledge, GreenBST is the first practical search tree that achieves both portable fine-grained data locality and high concurrency. GreenBST is open source and available at: https://github.com/uit-agc/GreenBST.

3) We have analyzed and compared GreenBST performance (in operations/second) and energy efficiency (in operations/Joule) with seven prominent concurrent search trees (see Table 1) on a high performance computing (HPC) platform (Intel Xeon), an embedded platform (ARM), and an accelerator platform (Intel Xeon Phi) (see Table 2 in Section 7 using parallel micro-benchmarks (Synchrobench \[27\]). Our experimental results show that GreenBST achieved up to 50% more energy efficiency and 60% higher throughput than the best competitor on the commodity HPC, embedded and accelerator platforms. Unlike platform-dependent search trees that are optimized for a specific platform (e.g., Fast \[30\] and Palm \[39\]), GreenBST is platform-independent and performance-portable.

4) We have provided insights into how portable fine-grained data locality actually influences energy efficiency and performance in concurrent search trees (see Section 8). Among our findings are: i) portable fine-grained data locality is able to reduce data movement, resulting in lower energy consumption and higher performance across HPC, embedded and accelerator platforms; ii) reducing the (hidden) overhead (e.g., pointers) in the tree structure allows a larger portion of real data in each memory transfer, resulting in better energy efficiency; and iii) for multicore platforms, efficient concurrency control is necessary for energy-efficient data structures, in addition to locality-awareness.

2 BOUNDED IDEAL CACHE MODEL

In order to devise locality-aware concurrent search trees, we need theoretical execution models that promote both data locality and concurrency. In this section, we present a new execution model called bounded ideal cache model (BCO) that promotes both data locality and concurrency. Unlike previous models such as the I/O model \[3\] and the ideal cache model (CO) \[25\] that promote only data locality, the new BCO model enables new designs of concurrent data structures that achieve both data locality and high concurrency (see concurrency-aware vEB layout (Section 4) and GreenBST (Section 5).

Before presenting the new BCO model, we first provide some background on relevant models, namely the ideal cache model (CO) \[25\]. These models enable the analysis of data transfer between two levels of the memory hierarchy. Lower data transfer complexity implies better data locality and, therefore, higher energy efficiency since energy consumption caused by data transfer dominates the total energy consumption \[18\].

2.1 Ideal cache model

The ideal cache model was introduced by Frigo et al. \[25\], which is similar to the I/O model \[3\] except that the block size \(B\) and memory size \(M\) are unknown. Using the same analysis as the
I/O model, an algorithm is categorized as cache-oblivious if it has no variables that need to be tuned with respect to cache size and cache-line length, in order to optimize the data transfer complexity.

An optimal cache-oblivious algorithm for a two-level memory is also optimal for any adjacent pair of a multi-level memory. Therefore, without prior knowledge about a memory hierarchy, a cache-oblivious algorithm can automatically adapt to the memory hierarchy with multiple levels. It has been reported that cache-oblivious algorithms perform better and are more robust than the cache-aware algorithms \cite{10} for a more comprehensive overview of the I/O model and the ideal cache model.

\section{Bounded ideal cache model}

We define \textit{bounded ideal cache model} (BCO) to be the ideal cache model (CO) with an extension that an upper bound $U$ on the unknown memory block size $B$ is known in advance.

This extension is inspired by the fact that although the exact block size at each level of the memory hierarchy is architecture-dependent (e.g., register size, cache line size), obtaining a single upper bound for all the block sizes (e.g., register size, cache line size, and page size) is architecture-independent. For example, the page size obtained from the operating system is such an upper bound.

Unlike the CO model, the new BCO model facilitates designing concurrent data structures and algorithms that support not only fine-grained data locality but also high concurrency. The BCO model inherits fine-grained data locality from the CO model while achieving higher concurrency by the new ability of organizing data in smaller memory chunks of known size $U$. Without knowing an upper bound $U$, we must organize data for unknown block size $B$ that can be extremely large. We will elaborate on the advantage of the BCO model in designing concurrency-aware vEB layout in Section 3.2.

Moreover, the new BCO model maintains the key feature of the original CO model \cite{25}. First, temporal locality (i.e., reuse of the same data located in cache) is exploited perfectly as there are no constraints on cache size $M$ in the BCO model. Since the Least Recently Used (LRU) cache replacement policy with cache size $(1+\epsilon)M$, where $\epsilon > 0$, is almost as good as the optimal offline cache replacement policy with cache size $M$ \cite{20}, an optimal cache replacement policy can be assumed in the BCO model.

Second, analysis for a simple two-level memory are architecture-independent for an unknown multilevel memory (e.g., registers, L1/L2/L3 caches and memory). Namely, an algorithm that is optimal in terms of data transfer for a two-level memory is asymptotically optimal for an unknown multi-level memory. This feature enables designing algorithms that can utilize fine-grained data locality (e.g., at cache levels) in the multi-level memory hierarchy of modern architectures.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|l|c|c|c|}
\hline
\# & Algorithm & Ref & Description & Synchronization & Affiliation & Data structure \\
\hline
1 & SVEB & \cite{11} & Conventional vEB layout search tree & global mutex & U. Aarhus & binary-tree \\
2 & CITRUS & \cite{5} & RCU-based search tree & lock-based & Technion & binary tree \\
3 & LFBST & \cite{36} & Non-blocking binary search tree & lock-free & UT Dallas & binary tree \\
4 & BSTT & \cite{32} & Portable scalable concurrent search tree & lock-based & EPFL & binary tree \\
5 & LYBTREE & \cite{22} & Concurrent B-tree (B-link tree) & lock-based & CMU & b-tree \\
6 & ABTREE & \cite{13} & Lock-free (a,b)-tree & lock-free & Technion & (a,b)-tree \\
7 & BWTREE & \cite{24} & Lock-free Bw-tree & lock-free & CMU & b-tree \\
\textbf{new} & GreenBST & - & Locality aware concurrent search tree & lock-based & \textbf{this paper} & b-tree \\
\hline
\end{tabular}
\caption{List of evaluated concurrent tree algorithms. These algorithms are sorted by synchronization type.}
\end{table}

\section{Concurrency-aware vEB layout}

In this section, we present a new concurrency-aware van Emde Boas layout (CvEB), an improvement to the conventional van Emde Boas layout (vEB). The new CvEB layout is based on the new BCO model presented in Section 2.

We first define the notations that will be used to elaborate on the improvement:

- $b_i$ (unknown): block size (in terms of the number of nodes) at level $i$ of the memory hierarchy (like $B$ in the I/O model \cite{3}), which is unknown as in the ideal cache model \cite{25}. When the specific level $i$ of the memory hierarchy is irrelevant, we use notation $B$ instead of $b_i$ in order to be consistent with the I/O model.

- $U$ (known): the upper bound (in terms of the number of nodes) on the block size $b_i$ of all levels $i$ of the memory hierarchy.

- $GNode$: the largest recursive subtree of a vEB-based search tree that contains at most $U$ nodes (see dashed triangles of height $2^k$ in Figure 1b). $GNode$ is a fixed-size tree-container with the vEB layout.

- "level of detail" $k$, $k \in \mathbb{N}$, is a number representing a partition of a vEB-based tree into recursive subtrees of height at most $2^k$.

- Let $L$ be the level of detail of GNode. Let $H$ be the height of GNode, we have $H = 2^L$. For simplicity, we assume $H = \log_2(U+1)$.

- $N,T$: size and height of the whole tree in terms of basic nodes (not in terms of GNodes).

\subsection{Conventional van Emde Boas (vEB) layout}

The conventional van Emde Boas (vEB) layout has been introduced in \cite{24} and widely used in cache-oblivious data structures \cite{7}, \cite{8}.
where of which is stored in a contiguous block of memory (of size $S$) upper bound $U$. Our key idea is that if we know an upper bound concurrency-aware data structures with update operations, we introduce a novel in order to make the vEB layout suitable for highly concurrent inserting new nodes in the block of memory as follows. The tree is conceptually split between nodes of height $h/2$ and $h/2+1$, resulting in a top subtree $T$ and $m_1 = 2^{h/2}$ bottom subtrees $W_1, W_2, \ldots, W_{m_1}$ of height $h/2$. The $(m_1 + 1)$ top and bottom subtrees are then located in contiguous memory locations where $T$ is located before $W_1, W_2, \ldots, W_{m_1}$. Each of the subtrees of height $h/2$ is then laid out similarly to $(m_1 + 1)$ subtrees of height $h/4$, where $m_2 = 2^{h/4}$. The process continues until each subtree contains only one node, i.e., the finest level of detail, 0.

One of the key features of the vEB layout is that the cost of searching a key located at a leaf in this layout is $O(\log B N)$ memory transfers, where $N$ is the tree size and $B$ is the unknown memory block size in the ideal cache model [25].

The new concurrency-aware vEB layout is the same as in the conventional vEB layout (see Lemma 3.1) while the CvEB layout supports high concurrency with update operations. For example, when subtree $W$ in Figure 2 is full, the CvEB layout allows allocating a new subtree $X$ and linking $X$ to $W$ as in k-ary search trees. By incorporating pointers, the CvEB layout enables highly concurrent (update) operations on subtrees (e.g., by using universal methodologies for implementing highly concurrent data structures [28]).

**Lemma 3.1.** For any upper bound $U$ of the unknown memory block size $B$, a search in a complete binary tree with the new concurrency-aware vEB layout achieves the optimal memory transfer $O(\log B N)$, where $N$ and $B$ are the tree size and the unknown memory block size in the ideal cache model [25], respectively.

**Proof.** (Sketch) Figure 2 illustrates the proof. Let $k$ be the coarsest level of detail such that every recursive subtree contains at most $B$ nodes, where $B$ is unknown. Since $B \leq U$, $k \leq L$, where $L$ is the coarsest level of detail at which every recursive subtree (or GNode) contains at most $U$ nodes. That means there are at most $2^{L+k}$ subtrees of height $2^k$ along the search path in a GNode and no subtree of height $2^k$ is split due to the boundary of GNodes. Namely, in Figure 3, each triangle of height $2^k$ fits within a dashed triangle of height $2^k$.

Because at any level of detail $i \leq L$ in the CvEB layout, a recursive subtree of height $2^k$ is stored in a contiguous block of memory, each subtree of height $2^k$ within a GNode is stored in at most $2^{L+k}$ memory blocks of size $B$ (depending on the starting location of the subtree in memory). Since every subtree of height $2^k$ fits in a GNode (i.e., no subtree is stored across two GNodes), every subtree of height $2^k$ in the tree is stored in at most $2$ memory blocks of size $B$.

Let $T$ be the tree height. A search will traverse $[T/2^k]$ subtrees of height $2^k$ and thereby at most $2[T/2^k]$ memory blocks are transferred.

Since a subtree of height $2^{k+1}$ contains more than $B$ nodes, $2^{k+1} \geq \log_2 (B+1)$, or $2^k \geq \frac{1}{2} \log_2 (B+1)$.

We have $2^{L-1} \leq N \leq 2^L$ since the tree is a complete binary tree. This implies $\log_2 N \leq T \leq \log_2 N + 1$.

Therefore, the number of memory blocks transferred in a search is $2[T/2^k] \leq 4[\log_2 N + 1] = 4(\log_2(B+1) + \frac{1}{2}) = O(\log_2 N)$, where $N \geq 2$. 

Figure 2 illustrates the new concurrency-aware vEB (CvEB) layout. Let $L$ be the coarsest level of detail such that every recursive subtree contains at most $U$ nodes. Namely, let $H$ and $S$ be the height and size of such a subtree then $H = 2^L$ and $S = 2^H - 1 \leq U$. The tree is recursively partitioned into level of detail $L$ where each subtree represented by a triangle in Figure 3, is stored in a contiguous memory block of size $U$. Unlike the conventional vEB, the subtrees at level of detail $L$ are linked to each other using pointers, namely each subtree at level of detail $k > L$ is not stored in a contiguous block of memory. Intuitively, since $U$ is an upper bound on the unknown memory block size $B$, storing a subtree at level of detail $k > L$ in a contiguous memory block of size greater than $U$, does not reduce the number of memory transfers, provided there is perfect alignment. For example, in Figure 2, traversing from a subtree $W$ at level of detail $L$, which is stored in a contiguous memory block of size $U$, to its child subtree $X$ at the same level of detail will result in at least two memory transfers: one for $W$ and one for $X$. Therefore, it is unnecessary to store both $W$ and $X$ in a contiguous memory block of size $2U$. As a result, the memory transfer cost of search operations in the new CvEB layout is the same as in the conventional vEB layout (see Lemma 3.1) while the CvEB layout supports high concurrency with update operations. For example, when subtree $W$ in Figure 2 is full, the CvEB layout allows allocating a new subtree $X$ and linking $X$ to $W$ as in k-ary search trees. By incorporating pointers, the CvEB layout enables highly concurrent (update) operations on subtrees (e.g., by using universal methodologies for implementing highly concurrent data structures [28]).
4 GreenBST Overview

To prove that the new conceptual CvEB layout (see Section 3) is useful for concurrent search trees to achieve both portable fine-grained data locality and high concurrency in practice, we devise a novel locality-aware concurrent search tree based on the new CvEB layout, which is called GreenBST.

A GreenBST $G_T$ consists of $|G_T|$ GNodes of fixed size $U$ (see Figure 3). Each of the GNodes contains a pointer-less binary search tree (BST) $T_i, i = 1, ..., |G_T|$. Nodes of tree $T_i$ are basic nodes which should be distinguished from GNodes. GreenBST provides the following operations: UPDATE($v, T_i$), which adds or removes value $v$ from GreenBST $T_i$, and SEARCH($v, T_i$), which determines whether value $v$ exists in $T_i$. We use term update operation for either insert or delete operation. We assume that duplicate values are not allowed inside the tree and a special value, for example 0, is reserved as the EMPTY value.

4.1 Data structures

The topmost level of GreenBST is represented by struct UNIVERSE-TYPE (line 1 in Figure 4) that contains pointer root to the root node of the root GNode. GNodes are represented by struct GNODE-TYPE (line 3 in Figure 4). A GNode $G$ includes a pointer root pointing to $G$’s root, a nodes group of size $U$ that hold values/keys, and a link array of size $U$ that links $G$’s basic leaf-nodes to the roots of $G$’s child GNodes. GNode metadata contains: i) pointer nextRight pointing to the GNode’s right sibling; ii) field highKey containing the lowest value/key of the right sibling GNode; iii) counter rev used for optimistic concurrency [31]; and iv) a lock locked protecting the GNode for update operations. Fields locked, rev, highKey and nextRight are used for GreenBST concurrency control.

Each NODE-TYPE structure (line 11 in Figure 4) contains i) field value holding a value for guiding the search, or a data value (or key) if it is a leaf-node; and ii) field mark indicating a logically deleted node (if the field is set to true).

Struct MAP-TYPE (line 14 in Figure 4), is used to remove pointers from the NODE structure (see Section 5.1 for details).

4.2 Balanced and concurrent tree

GreenBST high-level structure is inspired by the B+tree structure and GreenBST concurrency is inspired by the B-link tree concurrency which provides lock-less search operations [32]. However, unlike the B-link tree, GreenBST is an in-memory tree and uses the new CvEB layout for its GNodes. Moreover, GreenBST uses optimistic concurrency to handle lock-less concurrent search operations even in the occurrences of concurrent update operations.

To keep GreenBST balanced while supporting concurrent updates, the whole tree is built in a bottom-up manner, which is handled by the UPDATE function (see Figure 6). Meanwhile, the search operation traverses GreenBST in a top-down, left-to-right manner using a combination of the GNODE-SEARCH function to find the relevant leaf GNode (see Figure 5) and a binary search to find the relevant node within the found GNode.

Function UPDATE inserts a given key at a leaf of GreenBST $G_T$, provided key does not exist in the tree (see Figure 6, line 11). It first finds the appropriate leaf GNode to store the key (line 2) and locks the GNode (or its siblings if the GNode has been split) using the MOVE_RIGHT function (line 4). The MOVE_RIGHT function serializes concurrent updates by combining right-scanning (line 55) and lock coupling (see lines 56–59). If key is inserted to a node at the last level $H$ of the found GNode (i.e., at the boundary of the GNode), the UPDATE function will either rebalance the GNode’s embedded tree to reduce its height (line 10) or split the GNode into two GNodes (line 21).

Function REBALANCE(key, $T_i$) is responsible for rebalancing a GNode $T_i$ after an insertion (see Figure 6, line 41). GreenBST rebalance is incremental, significantly reducing the rebalance overhead. GreenBST incremental rebalance is described in detail in Section 5.3.

GreenBST split operation on a GNode $T_v$ distributes the member keys of $T_v$ between $T_r$ and a new GNode $T_s$ as evenly as possible (see Figure 5, line 21). The split operation also updates $T_v$.nextRight to point to $T_r$ (line 28) and fills $T_r$.highKey with $T_s$’s minimum key (line 27). Lastly, $T_v$’s minimum key is inserted into the parent GNode of both $T_r$ and $T_s$ (line 23). If $T_v$ is currently the root GNode, a new root GNode is created and it becomes the new parent of $T_r$ and $T_s$. Note that for a GNode $T_r$ with minimum key $K_{min}$, a new key less than $K_{min}$ will be forwarded to one of $T_r$’s left sibling GNodes. The leftmost sibling GNode will host the minimum key that has ever been inserted to GreenBST.

The delete operation in GreenBST is handled by the UPDATE function which marks the leaf node containing key as deleted (see Figure 6, line 5). Deleted nodes are reclaimed by the rebalance and split operations. The offline memory reclamation techniques used in the B-link tree can be deployed to merge nearly empty GNodes in the case where a large part of the workload is delete operations. GreenBST is, however, designed for workloads
function SEARCH(key, GreenBST, maxDepth)
  GNode ← GNODERESEARCH(key, GreenBST, maxDepth)
  return GNode

function UPDATE(key, GreenBST, maxDepth)
  GNode ← GNODERESEARCH(key, GreenBST, maxDepth)
  Push to stack the last visited GNodes at each level along the search path from GreenBST root to GNode
  GNode ← MOVE_RIGHT(key, GNode)
  Lock and move right, if necessary
  if (type = delete) then
    delete (GNode)
  else
    if (type = insert) then
      Insert key to an appropriate basic node L and update GNode.link]
  end if
  if (key is put at the last level H of GNode) then
    if (GNode isleaf and rebalance is possible) then
      Rebalance, see Section 5.5 for details
    end if
  end if
  insert to parent GNode at level
  Lock parent GNode
end function

function INSERT(key, GreenBST)
  GNode ← new GNode
  insert to parent GNode at level
  Lock parent GNode
end function

function DELETE(key, GreenBST)
  GNode ← new GNode
  insert to parent GNode at level
  Lock parent GNode
end function

Figure 5: Function SEARCH in GreenBST. The GNODERESEARCH function will return the leaf GNode for finding the searched key. From there, a simple binary search using LEFT and RIGHT functions (see Figure 7) is used to find the key location in the GNode or its right siblings. The SEARCH function utilizes counter rev, pointer nextRight and value highKey to ensure correct results even when executing concurrently with update operations.

4.3 Customized concurrency control

GreenBST uses locks and variables nextRight and highKey to coordinate concurrent search and update operations, while counter rev is used for the search optimistic concurrency [31]. When a GNode G needs to be updated, G’s rev counter is increased by one before the update operation starts (lines 6 and 13 in Figure 6). The counter is increased by one again after the update operation finishes (see lines 8 and 37). Note that all update operations happen when the lock is still held (see line 8) and therefore, only one update operation may increase G’s rev counter and update G at a time. The rev counter prevents the search operation from returning a wrong key because of a concurrent update operation (lines 8 and 36 in Figure 5).

GreenBST search uses optimistic concurrency [31] to ensure that the operation always returns the correct answer even if it arrives at a GNode that is undergoing update operation (e.g., insert and delete). First, before starting to traverse a GNode G, a search operation records G’s rev counter (line 14 in Figure 5). Before following a link to a child GNode or returning a link, the search operation first re-checks the GNode’s rev counter and moves nextRight (or highKey) if necessary (lines 10 and 14). This ensures that a concurrent update has occurred (see lines 10 and 14), and therefore, new update operation may lock and proceed (see lines 11 and 14). This helps prevent race conditions and ensures that the operation always returns the correct answer even if it executes concurrently with update operations.
as GNode’s cached map (Section 5.1), efficient inter-GNode connection (Section 5.2) and incremental rebalancing (Section 5.3).

### 5.1 Cached map instead of pointers

Although removing pointers connecting basic nodes in GNode reduces data transfer between memory levels, it raises several challenges. Two of the key challenges are how to properly connect child- and parent-nodes within a GNode and how to establish child-parent paths between GNodes (i.e., inter-GNode links). We address the former in this subsection and the latter in subsection 5.2.

We replace GNode basic (left and right) pointers with functions LEFT and RIGHT that utilize a cached map (see line 14 in Figure 4) and Figure 7. The LEFT and RIGHT functions, given an arbitrary node \( p \) and the memory address of its GNode base, return the addresses of \( p \)'s left and right child nodes, respectively, and 0 if \( p \) has no children (i.e., \( p \) is a leaf node of the GNode). The LEFT and RIGHT functions throughout GreenBST share a common cached map instance (Figure 7, line 1). As all GNodes use the same fixed-size vEB layout, only one map instance with size \( U \) is needed for all traversing operations. This approach makes GreenBST memory footprint small and keeps the frequently-used map instance in cache.

For example, assume that a GNode has 127 basic nodes. Without using the new cached map, the set of basic (left and right) pointers occupy 2032 bytes (127 \( \times \) 16 bytes) of memory in a 64-bit operating system, four times the amount of memory used by the actual data (i.e., 127 \( \times \) 4 bytes = 508 bytes, assuming node’s value is a 4-byte integer). Even if the node value is multiple words in length, pointers are still going to occupy a large part of GNode memory (e.g., even for 64-byte node value, pointers still occupy 20% of GNode memory). Therefore, using pointers is inefficient, because every block transfer between levels of memory carries a significant portion of pointers (or overhead) instead of actual data. The new cached map completely removes the pointer overhead.

Our new mapping approach addresses the drawbacks of previous approaches (e.g., pointers and arithmetic-based implicit addresses) and is unique to the concurrency-aware vEB layout as it exploits the fixed-size feature of GNodes. Previous approaches such as pointers and arithmetic-based implicit addresses in cache-oblivious (CO) trees are found to have weaknesses. Pointer-based approaches induce high overhead in term of data transferred between memory levels: the inclusion of pointers reduces the amount of real data (e.g., keys) in each block transferred. The implicit addresses, which demand arithmetic calculation for every node traversal, induce high computation overhead, especially when the tree is big. Our new mapping approach eliminates both pointer overhead (i.e., no pointers) and computation overhead (i.e., addresses are pre-computed and stored in the cached map). The LEFT(\( p, \) base) (resp. RIGHT(\( p, \) base)) function only finds \( p \)'s index and gets the address of \( p \)'s left child (resp. right child) from the cached map (e.g., \( id_{x} \) at line 11 and map\( [idx], \) left at line 13 in Figure 7).

Note that the mapping approach does not induce memory fragmentation because the approach applies only for an individual GNode. Namely, the cached map is the access structure of an individual GNode, which is used to find the left/ right child of an basic node in the GNode, without need of left/ right pointers. As an GNode is a fixed tree-container of size \( U \) and with the well-defined vEB layout, the cached map is fixed. The update operations change only the values of GNode basic nodes (e.g., from EMPTY to an input value in the case of insertion), but do not change GNode structure.

### 5.2 Inter-GNode connection

To enable traversing from a GNode to its child GNodes, we develop a new inter-GNode connection mechanism. Figure 5 explains how the inter-GNode connection works in a pointer-less search operation. We logically assign binary values to basic edges within GNode so that each path from GNode root to a basic leaf node is represented by a unique bit-sequence. As basic nodes have only left and right conceptual edges, we assign 0 and 1 to the left and right edges, respectively (lines 25 and 30). The bit-sequence is then used as an index in the link array containing pointers to child GNodes (line 23). The maximum size of the bit representation is GNode height or \( \log_{2}(U) \) bits.

Particularly, each basic leaf node \( L \) of a non-leaf GNode has two child GNodes pointed by link[\( L_{left} \)] and link[\( L_{right} \)] where \( L_{left} \) and \( L_{right} \) are computed when the GNodeSearch function traverses GreenBST (lines 25, 29 and 32). For example, if \( L \)'s depth is 1 (i.e., GNode has only one basic node) and GNode height maxDepth is 3, \( L_{left} = 0 \) and \( L_{right} = (1 << (3 - 1)) = 4 \) (line 32). Namely, link[0] and link[4] point to \( L \)'s left child GNode and \( L \)'s right child GNode, respectively.

### 5.3 Incremental rebalancing

GreenBST is inspired by the incremental rebalance proposed in [11] which significantly reduces the rebalance overhead. However, unlike the approach in [11] that occasionally has to rebalance the whole tree, we only apply the incremental rebalance on the basic tree embedded in each GNode (line 41 in Figure 6). Note that for workloads dominated by search and insert operations, the high-level tree of GNodes is already balanced because of the bottom-up construction (see Section 4.2). For workloads dominated by delete operations, the high-level tree of GNodes may be unbalanced due to empty GNodes when the merge operation is not implemented.

To briefly explain the idea, we denote density(\( w \)) as the ratio of number of keys inside a subtree \( T \) rooted at node \( w \) to the maximum number of keys that subtree \( T \) can hold (line 45 in see Figure 6). For example, a subtree with root \( w \) at level \( H - 3 \), where \( H \) is GNode height, can hold at most \( 2^{H - 1} - 1 \) keys. If the subtree only contains 3 keys, then density(\( w \)) = \( \frac{3}{2^{H - 1} - 1} = 0.42 \). Let \( \Gamma \) be the density threshold at level \( i \), we have \( 0 < \Gamma_{1} < \Gamma_{2} < \ldots < \Gamma_{H} = 1 \), where \( H \) is GNode height. After a new key is inserted at a basic leaf \( v \), we find the nearest ancestor node \( w \) of \( v \) so that density(\( w \)) \( \leq \Gamma_{\text{depth}(w)} \) where \text{depth}(\( w \)) is the level where \( w \) resides, counted from the GNode root (e.g., depth(GNode.root) = 1) (see line 47). If \( w \) is found, we rebalance the subtree rooted at \( w \).
Density thresholds $\Gamma_i, i < H$, are tuning parameters and are set manually. In our GreenBST implementation, $\Gamma_i, i < H$, are set to 0.5.

6 Correctness Proof

Definition 6.1. A GreenBST is well-formed if the following structural properties hold to all (concurrent) functions:

- The binary search tree (BST) embedded in each GNode is well-formed. Namely, BST has no key duplicates, the left (resp. right) sub-tree of a basic node BN contains keys less than (resp. greater than or equal to) BN.value.
- Links between GNodes are well-formed. Namely, if a GNode $G_i$ has a nextRight pointer to a sibling GNode $G_j$, $G_i$ contains keys less than $G_j$.highKey and $G_i$ contains keys greater than or equal to $G_j$.highKey. If a basic leaf node BN in $G_i$ has a left (resp. right) pointer to a child GNode $G_j$, $G_j$ contains keys less than (resp. greater than or equal to) BN.value.
- There are no key duplicates among GNodes.

Lemma 6.1. Let $G$ be a GNode in a well-formed GreenBST GT. If the Update function appears atomic to the Search and GNODESEARCH functions, an Update function on $G$ makes $G$ continue being well-formed.

Proof. (Sketch) As the Update function appears atomic to the Search and GNODESEARCH functions (the hypothesis), we only need to prove that an Update function on $G$ makes $G$ continue being well-formed regardless of whether the function is interfered by another Update function.

Case 1: no interference on $G$ from another Update function: If the Update function GU deletes key (line 5 in Figure 6), it sequentially searches $G$’s BST for the basic node containing key as in conventional sequential BST and then marks the node, if found, as deleted (line 7), making $G$’s BST continue being well-formed. Note that deleted nodes are kept in $G$’s BST to maintain the BST structure until the maintenance operations (e.g., REBALANCE and SPLIT) rebuild the whole subtree.

If GU inserts key without triggering maintenance operations (e.g., REBALANCE or SPLIT), it performs the conventional sequential BST insert operation on $G$’s BST, making $G$’s BST continue being well-formed (line 14).

If GU invokes the REBALANCE function on $G$ (line 16), it searches $G$’s BST for an appropriate basic node w and sequentially rebuilds the subtree rooted at w as a balanced BST using the same keys (lines 41-51), maintaining the well-formed properties of $G$’s BST. Moreover, since the BST of inner GNodes $G$ is leaf-oriented, the REBALANCE function does not change the links between $G$ and its child GNodes, making the links continue being well-formed.

If GU triggers the split operation on $G$, splitting $G$ into two GNode $G_1$ (or GNode) and $G_2$ (or newGNode) (lines 21-35), we will prove that links between $G_1$, $G_2$ and $G$’s parent GNode $G_p$ are well-formed. Note that since $G_1$’s BST (resp. $G_2$’s BST) is newly created from the lower half (resp. higher half) of $G$’s sorted keys, $G_1$’s BST and $G_2$’s BST are well-formed and there are no key duplicates between $G_1$ and $G_2$. Indeed, as $G_2$.highKey $== G$.highKey and $G_2$.nextRight $== G$.nextRight (lines 25-26), the $G_2$.nextRight link between $G_2$ and $G$’s right sibling is kept well-formed as before splitting. Similarly, as $G_1$.highKey contains $G_2$’s lowest key $y$ and $G_1$.nextRight points to $G_2$ (lines 27-28), the $G_1$.nextRight link between $G_1$ and $G_2$ is well-formed. Note that since no other function has a reference to $G_2$ until $G_2$ is inserted to parent $G_p$ (line 33), the modifications on $G_1$ and $G_2$ during the split (lines 23-28) are atomic to other Update functions.

Regarding the links between parent $G_p$ and $G_2$, as $G_2$’s lowest key $y$ and pointer link$\text{Prev}$ to $G_2$ are assigned to $G_2$’s basic leaf node $L$ and $L$’s right pointer $G_p$.link[1] respectively (lines 31-34 and 14), the link between $G_p$ and $G_2$ is well-formed (i.e., $G_2$’s keys are greater than or equal to $L$.value). Regarding the link between $G_p$ and $G_1$, because $G_1$ reuses the memory allocated to $G$ and the link between $G_p$ and $G$ is well-formed before the split, the link between $G_p$ and $G_1$ becomes $L$’s left pointer $G_p$.link[1] and therefore is well-formed (i.e., $G_1$’s keys are less than $L$.value, or $G_2$’s lowest key $y$). Since GU successfully locks $G_p$ (line 34) before invoking the Update function on $G_p$ (lines 12-38), the same well-formed proof on $G$ applies for $G_p$.

Note that since $G$’s memory and the link $K$ from $G_p$ to $G$ are re-used for $G_1$, a function that has read $K$ before $G$ is split, will able to access $G_1$ via $K$ and $G_2$ via $G_1$.nextRight, finding all $G$’s keys.

Case 2: interference on $G$ from other Update functions: We will prove that an Update function GU on $G$ appears atomic to other concurrent Update functions and therefore this case becomes Case 1. Indeed, as GU that is modifying $G$, has successfully locked $G$ (line 4 in Figure 6), other concurrent Update functions on $G$ must wait for GU to finish its modification and unlock $G$ (line 9). The linearization point of the Update function in this case is the time point when Lock($G$) in the MOVE_RIGHT function (line 4 in Figure 6) returns (lines 54 or 57).

As GreenBST is initiated as an empty GNode and thus well-formed, Lemma 6.1 implies that GreenBST is always well-formed if we can prove that the Update function appears atomic to the GNODESEARCH and Search functions (cf. Lemmas 6.3 and 6.6).

Lemma 6.2. Let to be the time at which a reference to the leaf GNode $L$ that is returned by the GNODESEARCH(key,GreenBST.maxDepth) function (invoked by the Search or Update function), is made (line 72, 76 or 77 in Figure 5). Any change to $L$ at time $t > t_0$ will be observed (by the Search and Update functions) in either $L$ or one of $L$’s right siblings that are reachable via the nextRight pointers.

Proof. (Sketch) Changes to $L$ (e.g., inserting or deleting key) at time $t > t_0$ that cannot be found in $L$, occur when $L$ has been split into two leaf GNodes, $L$ and $L’$, and the changes are located in $L’$ (lines 21-28 in Figure 6). Note that $L’$ is reachable from $L$ via the $L$.nextRight pointer (line 28). The splitting can occur repeatedly, creating a linked list of sibling GNodes originated from $L$.

To prove that following the $L$.nextRight pointers will eventually find any change made to $L$ at time $t > t_0$, we need to prove that the split operation on a GNode $G$ appears atomic to other concurrent Update and Search functions on $G$. Indeed, as the Update function GU that performs the split operation on $G$, has successfully locked $G$ (line 4) and only releases the lock after finishing the split operation (line 35), the split operation appears atomic to other concurrent Update functions. The linearization point of the Update function in this case is the time point when Lock($G$) in the MOVE_RIGHT function (line 4 in Figure 6) returns (lines 54 or 57).

Moreover, as function GU makes counter $G$.rev odd number before the split (line 13) and makes $G$.rev even number again only after finishing the split (line 28), concurrent Search functions GS
will wait for the update/split to finish before actually accessing GNode G (lines [3] and [8] in Figure 5). Since the UPDATE function must successfully acquire G’s lock (line [4] in Figure 5) before increasing G.rev (line [13]), only one UPDATE function can increase G.rev and therefore odd G.rev indicates an on-going update. Note that if the split function interferes with GS and makes G.rev even number again between lines [3] and [8] in Figure 5 GS will discover that G.rev has been changed (i.e., G.rev ≠ rev) and will then wait for the split operation to finish (line [8] in Figure 5). In this case, the linearization point of the split operation is the time point when increment(G.rev) at line [13] in Figure 5 returns. The linearization point of the SEARCH function is the time point when the SEARCH function observes that G.rev is unchanged and even at line [8] in Figure 5.

**Definition 6.2.** The subtree rooted at a GNode G in a well-formed GreenBST includes both the subtrees of G’s child GNodes linked by G.link and the subtree of G’s sibling GNode linked by G.nextRight (see Figure 3 for illustration).

**Lemma 6.6.** For each GNode VG visited during GNODESEARCH(key,GreenBST, maxDepth) where GreenBST is well-formed, the next GNode G made at time t (line [12] or [13]) in Figure 5 from the last optimistic transaction for VG (i.e., last execution of lines [12] - [13] with GNODE == VG) satisfies the following claim: if key in the tree, then it is in the subtree rooted at G at time t.

**Proof.** (Sketch) We will prove the lemma inductively on G. Let G₁, G₂, · · · , Gₙ be the path from the root GNode G₁ to a leaf GNode Gₙ visited by a GNODESEARCH function GS.

The lemma holds when G is the root GNode, i.e., G = G₁ (line [12]).

Assume that the lemma holds for G = Gₖ, we will prove that the lemma holds for G = Gₖ₊₁.

Indeed, as the lemma holds for G = Gₖ, GS will visit Gₖ in the next iteration to find key (line [14] - [16]). We will prove that when visiting Gₖ, GS locates a correct child/ sibling GNode Gₖ₊₁ whose subtree contains key (if key exists).

**Case 1: no interference on Gₖ from UPDATE functions (i.e., no interference between line [14] and line [16] when GNODE = Gₖ).**

If key < Gₖ.highKey, GS traverses Gₖ’s internal binary search tree (BST) using key, reaching the appropriate basic leaf node LN (lines [23] - [30]) and its associated (left or right) pointer link[bits] to a child GNode Gₖ₊₁ at the next GNODE level (line [30]). As in the conventional binary search tree, key, if existing, is located in the subtree rooted at Gₖ₊₁ (see Figure 5 for illustration).

If key ≥ Gₖ.highKey, key, if in the tree, was moved to one of subtrees rooted at Gₖ’s sibling GNodes between the time Gₖ got a reference to Gₖ (line [17] or [35]) and the time GS started to access Gₖ (line [13]). In this case, the next GNode Gₖ₊₁ is Gₖ’s first sibling GNode (line [16]) whose subtree contains key (if key is in the tree) (see Definition 6.2).

**Case 2: interference on Gₖ from an UPDATE function GU.**

In this case, we will prove that function GU appears atomic to function GS and therefore this case becomes Case 1.

Indeed, similar to the proof of Lemma 6.2, as GU makes counter Gₖ.rev odd number during its update (Figure 6, lines [6] - [8] for deletion, lines [13] - [18] for insertion without maintenance, lines [13] - [18] for insertion with rebalance, and lines [13] - [29] for insertion with split), GS will discover GU’s interference during its search via checking counter Gₖ.rev (lines [14] - [14] in Figure 5) and will wait for GU to finish before actually accessing Gₖ (line [35]). The linearization point of the GNODESEARCH function GS is the time point when the function observes that Gₖ.rev is unchanged and even at line [35] in Figure 5. The linearization point of the UPDATE function GU is the time point when increment(Gₖ.rev) returns (Figure 6, line [6] in the case of deletion and line [13] in the case of insertion).

**Lemma 6.4.** Let t₀ be the time at which a reference to a leaf GNode LG returned by the GNODESEARCH(key,GreenBST, maxDepth) function is made. If key in the tree, it is in LG at time t₀.

**Proof.** (Sketch) According to Lemma 6.3, if key is in the tree, it is in the subtree rooted at LG at time t₀. As LG is a leaf GNode (line [13]), key, if existing, is in LG at time t₀.

**Lemma 6.5.** Let L be the GNode that is returned by the GNODESEARCH(key,GreenBST, maxDepth) function. The key, if existing, is located in the first GNode fG with fG.highKey > key in the linked list of L’s siblings originated from L (including L).

**Proof.** (Sketch) Let t₀ be the time at which a reference to the GNode L returned by the GNODESEARCH(key,GreenBST, maxDepth) function is made. According to Lemma 6.4, key, if existing, is located in L at time t₀ and the GNODESEARCH(key,GreenBST, maxDepth) function is correct. According to Lemma 6.3, if key is, existing, is located in LG at time t₀.

Let cG be the current leaf GNode that S is visiting. Initially, cG is LG.

**Case 1: no interference on cG from the UPDATE function since time t₀.** In this case, S returns the same result as does conventional sequential binary search (line [7]). In this case, the linearization point of the SEARCH function is the linearization point of the GNODESEARCH function invoked at line [2] (cf. Lemma 6.3).

**Case 2: interference without splitting on cG from an UPDATE function U since time t₀.** In this case, similar to the atomicity proof of Lemma 6.3 S will discover the interference and wait for the update to finish before actually accessing cG, thanks to counter cG.rev (lines [8] and [9]). As S’s eventual access to cG is not interfered with any update, S returns the same result as does conventional sequential binary search (line [7]). In this case, the linearization point of the SEARCH function is the time point when the function observes that cG.rev is unchanged and even at line [9] in Figure 5. The linearization point of the UPDATE function U is the time point when increment(cG.rev) returns (Figure 6, line [6] in the case of deletion and line [13] in the case of insertion).

**Case 3: interference with splitting on cG from an UPDATE function U since time t₀.** In this case, according to Lemmas 6.2...
and \([6, 5]\) \(s\) will find the correct right sibling \(sG\) of \(cG\) where \(key\), if existing, is located, by following nextRight pointers and checking if \(key < sG\).\(highKey\) \(s\). Note that \(s\) will always find a \(sG\) satisfying \(key < sG\).\(highKey\) since \(cG\)'s last (rightmost) sibling \(IG\) has \(IG\).\(highKey\) = \(\infty\) \(s\). This case becomes Case 2 for the sibling GNode \(sG\).

**Lemma 6.7.** The Update function is correct.

**Proof.** (Sketch) We will prove that the Update(key, GreenBST, maxDepth) function in Figure 6 has the same effect as does the update operation of the conventional sequential binary search trees, even when other Update functions are interfering.

Let \(t_0\) be the time at which the leaf GNode \(IG\) returned by the GNODE\Search function \(\text{line 2}\) in Figure 6 is made. According to Lemma 6.4, \(key\), if existing, is located in \(IG\) at time \(t_0\).

Let \(cG\) be the current leaf GNode that \(U\) is visiting. Initially, \(cG\) is \(IG\).

**Case 1:** no interference on \(cG\) from another Update function since time \(t_0\). In this case, \(U\) performs either the conventional insert operation on \(cG\)'s internal binary search tree in the case of insertion (\(\text{line 14}\) in Figure 6) or marks as deleted the basic node containing \(key\) in the case of deletion (\(\text{line 7}\)), which has a similar effect as does the insert or delete operation of the conventional binary search tree. In addition, if the maintenance conditions (e.g., rebalance or split) are satisfied, the maintenance is performed sequentially without any effect on the correctness (cf. Lemma 6.1). In this case, the linearization point of the Update function is the linearization point of the GNODE\Search function invoked at \(\text{line 2}\) (cf. Lemma 6.3).

**Case 2:** interference without splitting on \(cG\) from other Update functions since time \(t_0\). As \(U\) has successfully locked \(cG\) (\(\text{line 3}\)) before performing any insertion or deletion, other concurrent update functions on \(cG\) must wait for \(U\) to finish its modification and unlock \(cG\) (\(\text{line 9}\)). As \(U\)'s update on \(cG\) is not interfered by other update functions because of locking, \(U\)'s update has a similar effect as does the insert or delete operation of the conventional sequential binary search tree. In this case, the linearization point of the Update function is the time point when lock() in the MOVE\_RIGHT function invoked at \(\text{line 4}\) returns (\(\text{line 54}\) or 57).

**Case 3:** interference with splitting on \(cG\) from other Update functions since time \(t_0\). In this case, according to Lemmas 6.2 and 6.5, \(U\) will find the correct right sibling \(sG\) of \(cG\) where \(key\) is located, by following nextRight pointers and checking if \(key < sG\).\(highKey\) in the MOVE\_RIGHT function invoked at \(\text{line 4}\) (\(\text{line 55} - 59\)). This case becomes Case 2 for the sibling GNode \(sG\).

**Lemma 6.8.** GreenBST is deadlock-free.

**Proof.** (Sketch) As the Search function is lock-less, we only need to prove that concurrent instances of the Update function lock GNodes in a well-defined order.

Indeed, for GNodes located at different tree levels, the Update function locks them from a GNode with lower level (i.e., child GNode) to a GNode with higher level (i.e., parent GNode) in the case of splitting (\(\text{lines 33 - 35}\) in Figure 5). For GNodes located at the same tree level, the Update function locks them from left to right in the MOVE\_RIGHT function by following the nextRight pointers (\(\text{lines 54 - 57}\)).

### Table 2: Testing platforms specifications.

<table>
<thead>
<tr>
<th>Name</th>
<th>HPC</th>
<th>ARM</th>
<th>MIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>Intel Haswell-EF Octa</td>
<td>Samsung Exynos Octa</td>
<td>Intel Knights Corner</td>
</tr>
<tr>
<td>Processors</td>
<td>Intel Xeon E5-2699 v3</td>
<td>1x Samsung Exynos 5410</td>
<td>1x Xeon Phi 31S1P</td>
</tr>
<tr>
<td># Cores</td>
<td>18 (36 with hyperthreading)</td>
<td>4x Cortex A7</td>
<td>57 (without hyperthreading)</td>
</tr>
<tr>
<td>Core clock</td>
<td>2.30 GHz</td>
<td>1.66 GHz (A15)</td>
<td>1.17 GHz</td>
</tr>
<tr>
<td>L1 cache (per core)</td>
<td>32/32 KB</td>
<td>32/32 KB</td>
<td>32/32 KB</td>
</tr>
<tr>
<td>L2 cache (per processor)</td>
<td>256 KB × 18 (approx. 4.6 MB)</td>
<td>2 MB (shared A15)</td>
<td>512 KB × 37 (approx. 29.1 MB)</td>
</tr>
<tr>
<td>L3 cache</td>
<td>45 MB (shared)</td>
<td>512 GB DDR4</td>
<td>512 GB LPDDR3</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Quick Path Interconnect (QPI)</td>
<td>Coherent Interconnect (CC) 400</td>
<td>5 GT/s Ring Bus Interconnect</td>
</tr>
<tr>
<td>Memory</td>
<td>32 GB DDR4</td>
<td>2 GB LPDDR3</td>
<td>6 GB DDR5</td>
</tr>
<tr>
<td>Compiler</td>
<td>GNU GCC 7.2.0 (using -O3)</td>
<td>GNU GCC 4.8.2 (using -O3)</td>
<td>Intel C Compiler (ver. 15.0.2) (-O3)</td>
</tr>
</tbody>
</table>

### 7 Experimental evaluation

We ran several different benchmarks to evaluate GreenBST performance (operations per second) and energy efficiency (operations per joule). We combined the benchmark results with the last level cache (LLC) and memory profiles to draw a conclusion of whether C\(\)EB-based trees such as GreenBST provide portable energy efficiency and performance across different platforms.

We evaluated GreenBST against the prominent non-blocking and lock-based search trees in the literature (see Table 1), using parallel micro-benchmark suite Synchrobench [27]. Their LLC and memory profiles (see Figures 3D and 3H) were collected to evaluate the impact of the locality-aware approaches on energy efficiency and performance (see Section 7.1).

The experimental benchmarks were conducted on an Intel high performance computing (HPC) platform, an ARM embedded platform, and an accelerator platform based on the Intel Xeon Phi architecture (MIC platform) (see Table 2). Scalable memory allocator jemalloc 5.1.0 was used for all the benchmarks on the HPC, ARM and MIC platforms. These benchmarks were repeated at least 5 times to guarantee consistent results.

GreenBST is open source and available at: https://github.com/uit-agc/GreenBST.

#### 7.1 Benchmark setup

We compared GreenBST with seven concurrent search trees (see Table 1) using parallel micro-benchmark suite Synchrobench [27]. GreenBST’s U is set to 4095 and LYBTree’s order (or the maximum number of keys within a node) [32] was set to 64 (or \(log_2\)order) = 6, the best configuration (cf. Figure 8). All running threads were pinned to the available logical cores using pthread_setaffinity_np. If the number of threads was less than the number of logical cores in the first CPU, all threads were pinned to the first CPU. Otherwise, threads were distributed evenly among available CPUs. The first thread to finish its work set a global flag, which caused each thread to terminate after its next operation. The experiments were performed in C/C++.

In order to evaluate the efficiency of GreenBST design, we implemented SVEB, a lock-based version of the conventional cache-oblivious search tree using global mutex. Namely, concurrent accesses to the conventional sequential C\(\)EB-based tree [11] were controlled by a global mutex.
All tree operations (i.e., search, insert, delete) used random values \( v \in (0, \text{init} \times 2^4), v \in \mathbb{N} \) where \( \text{init} \) was the initial size of trees. The \( \text{init} \) values were chosen to make the trees partially fit into the last level cache (LLC). The \( \text{init} \) value for HPC and MIC platforms was \( 2^{24} \) (i.e., 64MB of keys, approximately 2 times as much as the last level cache (LLC) of the HPC and MIC platforms) and for ARM platform was \( 2^{20} \) (i.e., 4MB of keys, 2 times as much as the ARM platform’s LLC). Multiple threads performed insertions and deletions until the data structure reached \( \text{init} \) keys. Due to the space constraints, we present only two cases: i) 90% search and 10% update and ii) 50% search and 50% update. The benchmarks were run with different numbers of cores between the minimum and maximum available cores on the HPC and ARM and MIC platforms.

**Energy efficiency metrics** (operations/joule) were the number of operations divided by the energy consumption. The ARM platform was equipped with a built-in on-board power measurement system that was able to measure the energy consumption for the A15 cores, A7 cores, and memory continuously in real-time. For the Intel HPC platform, the Intel PCM [1], using built-in CPU counters was used to measure the CPU and DRAM energy consumption. Energy consumption on MIC platform was measured by polling the `/sys/class/micras/power` interface every 50 milliseconds. The total energy consumed by CPUs and memory system was measured. The measurements started after the tree initialization.

Performance metrics (operations/second) were the number of operations \( \text{(rep} = 5,000,000) \) divided by the maximum time for the threads to finish the whole operations.

### 7.2 Energy efficiency evaluation

On the **Intel HPC platform**, GreenBST was more energy efficient than the other trees in all cases except the case of 1 thread (see Figure [8a] top bar-charts). In the case of 1 thread, SVEB was slightly more energy efficient than GreenBST because of SVEB simple concurrency control - global mutex. The global mutex, however, prevented SVEB from scaling with the number of cores while GreenBST scaled well with the number of cores. GreenBST was 40% more energy efficient than ABTREE, the best competitor, in the experiment running the 50%-search benchmark with 36 cores (see the right bar-chart). GreenBST energy-efficiency advantage over the other trees comes from the new concurrency-aware vEB layout that reduces data movement between memory levels (e.g., between LLC and DRAM as shown in Figure [8d]) while supporting high concurrency. For example, as LYBTREE node is a contiguous array of sorted keys to maximize spatial locality for search operations, insert operations may need to shift many keys in order to have room for a new key, causing high data movement (cf. Figure [8c] right chart). The amount of data transferred between CPU last level cache (i.e., L3-cache) and memory in GreenBST is much less than that in the other trees except SVEB (cf. Figure [8d]).

On the **ARM embedded platform** where the A15 processor with 4 cores was used, GreenBST energy efficiency scaled well and was significantly better than those of the other trees (see Figure [8b] top bar charts). Note that SVEB energy efficiency decreased significantly and became worse in the case of 4 cores. Contrarily, GreenBST scaled well with the number of cores and was 35% more energy efficient than ABTREE, the best competitor, in the experiment running the 50%-search benchmark with 4 cores (see the right bar-chart).

Note that LFBST update operations did not work on the ARM platform since it required 64-bit pointers while the ARM platform was 32-bit. As a result, LFBST was excluded from the experiments on the ARM platform.

On the **Intel MIC accelerator platform**, GreenBST was significantly more energy efficient than the other trees (see Figure [8c] top bar-chart). GreenBST was 50% more energy efficient than LYBTREE, the best competitor, in the case of 90%-search benchmark with 57 cores.

The results on **data movement** between processor last-level cache (LLC) and memory on the HPC and MIC platforms provide insights into why GreenBST was more energy efficient than all the other trees in most cases (see Figures [8d]). The LLC-DRAM data movements on the HPC platform and the MIC platform were collected using Intel PCM and PAPI library, respectively. GreenBST data transferred between LLC and memory was significantly less than those of the other trees across the platforms, thanks to GreenBST concurrent locality-aware layout CvEB (see Section 5.2). For example, on the HPC platform, GreenBST data transferred between LLC and memory was only half of ABTREE, the best competitor, in the 50%-search benchmark using 36 cores (see Figures [8c]). Moreover, GreenBST’s **memory footprint** is smaller than those of LYBTREE, CITRUS, LFBST and BSTTK, four of the six non-vEB trees (see Figure [8c]).

### 7.3 Performance evaluation

On the **Intel HPC platform**, GreenBST outperformed all the other trees in all the cases from 9 cores to 36 cores with both the 90%-search and 50%-search benchmarks (see Figure [8a] bottom line-charts). GreenBST throughput was 40% higher than that of LYBTREE, the best competitor, in the case of the 50%-search benchmark using 36 cores (see the right line-chart).

On the **ARM embedded platform**, GreenBST significantly outperformed all the other trees in all the experiments (see Figure [8b] bottom line-charts). GreenBST throughput was 60% higher than that of ABTREE, the best competitor, in the experiment running 50%-search benchmark with 4 cores (see the right chart).

On the **Intel MIC accelerator platform**, GreenBST outperformed all the other trees in all the experiments (see Figure [8c] bottom line-chart). GreenBST throughput was 40% higher than that of LFBST, the best competitor, in the 50%-search benchmark using 57 cores (see Figure [8c] bottom right line-chart).

### 8 DISCUSSIONS

#### 8.1 Locality-awareness and overhead minimization

The usage of concurrency-aware vEB layouts (or CvEB) is able to reduce GreenBST energy consumption and at the same time increase GreenBST performance across the HPC, embedded and accelerator platforms. Section 7.2 has highlighted how CvEB-based search trees such as GreenBST manage to outperform their counterparts in terms of energy efficiency and performance for synthetic benchmarks on the different platforms.

One of the interesting findings is that minimizing **hidden overhead** (e.g., pointers in search trees) in locality-aware data structures can significantly reduce the energy consumption and increase the runtime performance. Our new optimization techniques such as removing space-overhead of pointers (see Section 5.1) significantly reduce GreenBST space overhead, thereby reducing the data transfer between memory levels (cf. Figures [8d]). Moreover, the smaller embedded trees reduce the maintenance overhead for leaf GNodes because less data needs to be arranged in rebalance
Figure 8: (a,b,c) Energy efficiency and throughput comparison of the trees on the HPC, ARM and MIC platforms. (d) LLC-DRAM data movement on the HPC platform, collected from the CPU counters using Intel PCM, and on the MIC platform, collected using PAPI library. (e) Energy efficiency and data movement on the HPC platform for varying node size. (f) The tree memory footprint, collected using Linux’s PMAP command.

(a) **HPC platform.** GreenBST is up to 40% more energy efficient than ABTREE, the best competitor, and has up to 40% higher throughput than LYBTREE, the best competitor, in the 50% search benchmark using 36 hyper-threaded cores.

(b) **ARM platform.** GreenBST is up to 35% more energy efficient than ABTREE, the best competitor, and has up to 60% higher throughput than ABTREE in the 50% search benchmark using 4 cores.

(c) **MIC platform.** GreenBST is up to 50% more energy efficient than LYBTREE, the best competitor, in the 90% search benchmark using 57 cores and has up to 40% higher throughput than LFBST, the best competitor, in the 50% search benchmark using 57 cores.

(d) **Data movement** in the 50% search benchmark between CPU last level cache (LLC) and main memory (DRAM) on the HPC platform and the MIC platform.

(e) Energy efficiency and data movement between LLC and DRAM on the HPC platform for LYBTREE and GreenBST with varying node size.

(f) The tree memory footprint (in GB) on the HPC platform. Tested using random keys in the range \(1, r\), with \(r = 2^{25}\).
or split operations. The other optimization techniques such as incremental rebalancing (see Section 5.3) significantly reduce GreenBST maintenance overhead (e.g., rebalancing overhead) (see Section 3 in [2] for performance comparison between GreenBST and its variation without the optimization called DeltaTree).

### 8.2 Concurrency control

Some of the benchmark results show that besides data movements, efficient concurrency control is also necessary in order to devise energy-efficient data structures on multicore platforms. For example, in sequential executions (i.e., 1 core), the conventional vEB tree (SVEB) had the smallest amount of data transferred between memory and the last level cache (cf. Figure 8a) and thereby achieved the best energy efficiency (cf. Figure 8a). However, when using 2 or more cores, its energy efficiency failed to scale (cf. Figures 8a, 8b and 8c). SVEB is not designed for concurrent operations and therefore an inefficient concurrency control (i.e., a global mutex) had to be incorporated in order to include SVEB in this study. Note that we were unable to use a more fine-grained concurrency control without significantly changing SVEB data structure because SVEB uses a recursive layout fitted in a contiguous memory block (see Section 3.1). Therefore, although SVEB had the smallest amount of data transfer in sequential executions, in parallel executions the concurrent cores had to spend a lot of time waiting and competing for a lock. This is inefficient as waiting cores still consume power (e.g., static power).

### 8.3 Comparison with previous concurrent cache-oblivious trees

Based on experimental insights, GreenBST, a concurrent CvEB-based tree, is more efficient than previous theoretical concurrent cache-oblivious (CO) trees such as the concurrent packed-memory CO tree and concurrent exponential CO tree [9]. The concurrent packed-memory CO tree has a good amortized memory transfer cost of $\Theta(\log B N + \log^2 N/B)$ for tree updates, assuming that operations occur sequentially. However, the proposed data structure requires each node to have the parent-child pointers. Besides the complication in re-arranging those pointers, we have found that removing pointers from nodes to minimize memory footprint is significantly beneficial for cache-oblivious trees in practice (see Section 5.1).

In the concurrent exponential CO tree [9], expected memory transfer cost for search and update operations is $\mathcal{O}(\log B N + (\log \alpha \log N))$, assuming that all processors are synchronous. Cormen et al. [16, pp. 212], however, wrote that although exponential search tree algorithms [4] are an important theoretical breakthrough, they are fairly complicated in practice. Indeed, nodes in the concurrent exponential CO tree grow exponentially in size, which not only complicates the maintenance of inter-node pointers but also exponentially increases the tree memory footprint in practice. In contrast, the memory footprint of GreenBST with fixed-size GNodes gradually expands on-demand when the tree grows. Thanks to the fixed-size GNodes, GreenBST exploits further spatial locality by utilizing a cached map to eliminate the basic pointers overhead (see Section 5.1 for details).

### 9 CONCLUSIONS

The results presented in this paper provide a starting point to investigate further energy-efficient data structures and algorithms that exploit fine-grained data locality provided by ideal cache models. The results not only show that GreenBST is an energy-efficient concurrent search tree, but also provide important insights into how to develop energy efficient data structures in general. On single core systems, locality-aware data structures that can lower data movement, have been shown to be able to increase energy-efficiency. However, on multicore systems, locality-awareness alone is not enough, and good concurrency control and cache strategy are needed. Otherwise, the energy overhead of either waiting cores or interconnect-based cache coherence mechanisms can exceed the energy saving obtained by less data movement.

The CvEB-based search trees such as GreenBST are composed of tree-containers (i.e., GNodes), thereby being highly decomposable. Therefore, it is possible to extend the CvEB search trees to work with heterogeneous cores and memory systems, for example by utilizing Cosh OS abstractions [6]. Also, devising in-memory key-value stores on top of GreenBST is among our future works.

### ACKNOWLEDGMENTS

This work was supported by the Research Council of Norway under projects PREAPP (grant no. 231746/F20) and eX3 (grant no. 270053). The evaluation was partly performed on resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway (grant no. N9342K).

### REFERENCES


