PaC-Trees: Supporting Parallel and Compressed Purely-Functional Collections

Laxman Dhulipala
University of Maryland
USA
laxman@umd.edu

Guy E. Blelloch
Carnegie Mellon University
USA
guyb@cs.cmu.edu

Yan Gu
UC Riverside
USA
ygu@cs.ucr.edu

Yihan Sun
UC Riverside
USA
yihans@cs.ucr.edu

Abstract
Many modern programming languages are shifting toward a functional style for collection interfaces such as sets, maps, and sequences. Functional interfaces offer many advantages, including being safe for parallelism and providing simple and lightweight snapshots. However, existing high-performance functional interfaces such as PAM, which are based on balanced purely-functional trees, incur large space overheads for large-scale data analysis due to storing every element in a separate node in a tree.

This paper presents PaC-trees, a purely-functional data structure supporting functional interfaces for sets, maps, and sequences that provides a significant reduction in space over existing approaches. A PaC-tree is a balanced binary search tree which blocks the leaves and compresses the blocks using arrays. We provide novel techniques for compressing and uncompressing the blocks which yield practical parallel functional algorithms for a broad set of operations on PaC-trees such as union, intersection, filter, reduction, and range queries which are both theoretically and practically efficient.

Using PaC-trees we designed CPAM, a C++ library that implements the full functionality of PAM, while offering significant extra functionality for compression. CPAM consistently matches or outperforms PAM on a set of microbenchmarks on sets, maps, and sequences while using about a quarter of the space. On applications including inverted indices, 2D range queries, and 1D interval queries, CPAM is competitive with or faster than PAM, while using 2.1–7.8x less space. For static and streaming graph processing, CPAM offers 1.6x faster batch updates while using 1.3–2.6x less space than the state-of-the-art graph processing system Aspen.

CCS Concepts:
- Computing methodologies → Shared memory algorithms.

1 Introduction
Almost all modern programming languages include extensive support for collections, such as sets, maps, and sequences either as libraries or built-in data types. Support for such collections has become the cornerstone of large-scale data processing, as exemplified by systems such as Apache Spark [47]. Among the interfaces for collections, there has been a trend towards a functional style, shying away from mutation (e.g., Spark is functional). Functional interfaces have several advantages over mutating ones, including being safe for parallelism, allowing safe composition, permitting flexible implementations (e.g., using copies when helpful), and supporting snapshots. Supporting snapshots is particularly useful in scenarios in which a stream of updates is being made to a collection which is concurrently being analyzed [19, 21, 32, 35].

Recent work [45] has developed a purely functional library, PAM, for representing sequences, ordered sets, ordered maps, and augmented maps (defined in [45]) using balanced trees, called P-trees. P-trees use path copying to perform updates, supporting functional updates at a reasonably low cost (e.g., $O(\log n)$ per point update). However they come at a cost of high space usage—every element requires a node in the tree. This is particularly problematic for large-scale data analysis, since in large-systems memory is often the dominating cost.

In this paper we present Parallel Compressed trees (PaC-trees): a purely-functional data structure for supporting a similar functionality as P-trees but with significant reduction in space—up to an order of magnitude (see Fig. 1). Our approach is based on blocking the leaves and compressing the blocks using arrays (see Fig. 4). We present innovative techniques for compressing and uncompressing the blocks without needing to re-implement the full functionality of P-trees. Importantly, in the paper we analyze the cost of all the operations as a function of the block size B as well as the...
within the blocked leaves. In such an encoding, each element is encoded as a list of integer keys using difference encoding. The C-trees from Aspen [21] also support difference encoding. GBBS is the static compressed graph representation from the Graph Based Benchmark Suite [22] which uses difference encoding, and serves as a baseline for the tree-based graph representations.

collection size. This is analyzed both in terms of the work (runtime sequentially) and span (longest dependent path in parallel). The costs for a sample of the supported functions are given in Table 1. These costs can help the user decide on a block size for their particular application—a parameter that can be specified when creating a collection.

Using PaC-trees we have implemented CPAM: a C++ library which implements the full functionality of PAM, along with significant extra functionality involving compression. By default CPAM supports difference (or delta) encoding [36] within the blocked leaves. In such an encoding, each element is encoded based on the value of the previous element in the collection. This can greatly reduce space when elements that are close in the ordering of the collection are related. For example, if a graph is numbered so that neighboring vertices have similar indices, then the neighbors in a neighbor list will have small differences. These small numbers can then be encoded in a handful of bits each [42]. Similarly in an inverted index where each word points to a sequence of documents it appears in, if the documents are sorted, the differences between adjacent document identifiers can be small. This is especially true for common words, which take up the bulk of the space. In the paper we bound the extra space needed (due to the index using the tree structure) for PaC-trees compared to a static representation of the data (i.e., an array) directly using difference encoding (see Theorem 4.2).

In our default blocked representation, the first element of a block is represented uncompressed, and the rest of the elements are compressed relative to the previous element. In addition to delta-encoding, CPAM also supplies an interface for the user to define their own form of compression for each block. For example, they can quantize values, or use other variable length codes when keys are known to be small. CPAM uses a reference counting garbage collector to manage the memory for both the internal nodes and the compressed leaf nodes, which can be of variable size due to compression.

CPAM supports augmentation in which each tree node is encoded based on the value of the previous element in the collection. This can greatly reduce space when elements that are close in the ordering of the collection are related. For example, if a graph is numbered so that neighboring vertices have similar indices, then the neighbors in a neighbor list will have small differences. These small numbers can then be encoded in a handful of bits each [42]. Similarly in an inverted index where each word points to a sequence of documents it appears in, if the documents are sorted, the differences between adjacent document identifiers can be small. This is especially true for common words, which take up the bulk of the space. In the paper we bound the extra space needed (due to the index using the tree structure) for PaC-trees compared to a static representation of the data (i.e., an array) directly using difference encoding (see Theorem 4.2).

In our default blocked representation, the first element of a block is represented uncompressed, and the rest of the elements are compressed relative to the previous element. In addition to delta-encoding, CPAM also supplies an interface for the user to define their own form of compression for each block. For example, they can quantize values, or use other variable length codes when keys are known to be small. CPAM uses a reference counting garbage collector to manage the memory for both the internal nodes and the compressed leaf nodes, which can be of variable size due to compression.

CPAM supports augmentation in which each tree node is encoded based on the value of the previous element in the collection. This can greatly reduce space when elements that are close in the ordering of the collection are related. For example, if a graph is numbered so that neighboring vertices have similar indices, then the neighbors in a neighbor list will have small differences. These small numbers can then be encoded in a handful of bits each [42]. Similarly in an inverted index where each word points to a sequence of documents it appears in, if the documents are sorted, the differences between adjacent document identifiers can be small. This is especially true for common words, which take up the bulk of the space. In the paper we bound the extra space needed (due to the index using the tree structure) for PaC-trees compared to a static representation of the data (i.e., an array) directly using difference encoding (see Theorem 4.2).

In our default blocked representation, the first element of a block is represented uncompressed, and the rest of the elements are compressed relative to the previous element. In addition to delta-encoding, CPAM also supplies an interface for the user to define their own form of compression for each block. For example, they can quantize values, or use other variable length codes when keys are known to be small. CPAM uses a reference counting garbage collector to manage the memory for both the internal nodes and the compressed leaf nodes, which can be of variable size due to compression.

**Table 1.** Primitives from the Sequence, Set, and Map interfaces in CPAM, including the work and span bounds. Note that primitives marked with † are specific to Sequences, and Set and Map primitives cannot be applied to Sequences. $m, n$ are defined to be the size of the smaller and larger sets, respectively. $B$ is the block size (the size of a blocked leaf in a PaC-tree). We assume a parallelizable encoding for the span bounds.

<table>
<thead>
<tr>
<th>Primitive</th>
<th>Work</th>
<th>Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>Build</td>
<td>$O(n)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>Map</td>
<td>$O(n)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>Filter</td>
<td>$O(n)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>Reduce</td>
<td>$O(n)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>Take</td>
<td>$O(\log n + B)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>n-th</td>
<td>$O(\log n + B)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>FindFirst</td>
<td>$O(k)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>Append†</td>
<td>$O(\log n + B)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>Reverse†</td>
<td>$O(n)$</td>
<td>$O(\log n)$</td>
</tr>
<tr>
<td>Union</td>
<td>$O(m \log \min(n, B) + \min(m, B))$</td>
<td>$O(\log n \log m)$</td>
</tr>
<tr>
<td>Intersect</td>
<td>$O(m \log \min(n, B) + \min(m, B))$</td>
<td>$O(\log n \log m)$</td>
</tr>
<tr>
<td>Difference</td>
<td>$O(m \log n + \min(m, B))$</td>
<td>$O(\log n \log m)$</td>
</tr>
</tbody>
</table>
CPAM has similar performance on operations that visit the whole sequence, like reduce, but is significantly slower on nth since it requires \(O(\log n + B)\) work as opposed to \(O(1)\) for a random array access for ParallelSTL. On append CPAM is significantly faster since it requires \(O(\log n + B)\) work to join two trees instead of \(O(n)\) required by ParallelSTL to copy the input arrays into the output array.

We consider four applications: graphs, inverted indices, 2D range queries and 1D interval queries. For inverted indices, 2D range query and 1D interval query, CPAM achieves competitive performance to PAM while using 2.1x–7.8x less space. For graph processing, we compare to an existing system Aspen [21] that represents graphs using trees. CPAM uses 1.3–2.6x less space compared to Aspen, and is almost always faster than Aspen in all tested graph algorithms.

The main contributions of this paper are:

- A new functional data structure, PaC-trees, and associated parallel algorithms that support compression for sequences, sets, maps and augmented maps.
- Theoretical bounds on the costs (work and span) and the space of the data structure and associated algorithms.
- An implementation of PaC-trees as a library, CPAM, supporting the full functionality of PAM in addition to supporting default and user defined compression schemes.
- An experimental evaluation of the ideas and implementation on microbenchmarks and non-trivial applications.

2 Related Work

Our work extends P-trees and their C++ implementation in PAM [45]. Our key contribution is the ability to compress the trees achieving up to an order-of-magnitude reduction in space. This is achieved while being able to present cost bounds both in terms of time and space. These bounds are a function of a block size the user can select.

B-trees [6] and their variants block not just the leaves but all nodes of a tree, such that internal nodes can have a high fan-out. They are widely used in practice, especially for disk based data structures since nodes are on the scale of a page on disk and can be retrieved efficiently. However they are less relevant in the context of purely functional in-memory trees. In particular, path copying requires that an update copy all nodes on the path from the root to the leaf. If the nodes are large (e.g. 128+ elements each, as in our leaves) this copying would be very expensive both in terms of space and time. Various work has suggested blocking the leaves of a binary tree to represent sequences [1, 8, 15, 26, 33]. The idea is to reduce the cost of operations such as append or subsequence relative to array representations. As far as we know, these ideas have never been applied to ordered sets or ordered maps. We also do not know of work that then compresses within the blocks.

Aspen [21] is a system for graph processing, based on purely functional trees and uses compression for the neighbor lists. At a high-level, our goals are shared with Aspen (e.g., non-mutating updates), but Aspen has several limitations. Importantly it is only designed for graphs, supporting only a small part of the functionality of CPAM. The tree representation in Aspen is also very different. It randomly selects elements from the collection to be heads. It then attaches a block of nodes to each head corresponding to the keys between the head and the next head, and puts the heads into a binary tree. PaC-trees do not require randomization, and have stronger theoretical bounds for primitive operations such as \(\text{union}\) than the bounds provided by C-trees in Aspen. We use CPAM to implement the full functionality of Aspen and compare to Aspen in Section 9.4.

Fig. 3 compares P-trees from PAM, functional B-trees, C-trees from Aspen, and PaC-trees. The comparison illustrates how they differ when inserting a new key.

Like CPAM, the Apache Spark [47] system supports a functional interface for collections. However it has several significant differences. First, it only supports unordered sets. Second, although it has a shared-memory parallel implementation, it is primarily designed for a distributed setting. This means its shared-memory implementation is not ideal.3

There is extensive research on concurrent tree data structures [3, 4, 16, 18, 24, 34, 37]. This work is mostly orthogonal to our work. Such trees support a fraction of the functionality of CPAM, typically just supporting linearizable inserts, deletes, updates and finds. Some recent works support range queries [5, 25], or arbitrary queries on a snapshot [46].

We note that the design of the chunked sequence datatype [1] could in principle be extended to support sets, maps, and augmented maps, although the implementation is specialized for ephemeral sequences.

3Their shared-memory implementation is between 3.2–4.9x slower than CPAM for a map, reduce, and group-by style example taken from their user guide. For primitives such as map and reduce, their implementation performs up to 2 orders of magnitude worse than CPAM (see the full version).

Figure 2. Relative performance of sequence primitives in CPAM (using \(B = 128\)), PAM, and ParallelSTL [29] on a 72-core machine with 2-way hyper-threading enabled. The numbers shown on top of the bars are the parallel (144-thread) running times in milliseconds. Lower is better. All benchmarks are run on sequences of length \(10^6\) containing 8-byte elements. For append, ParallelSTL takes 17.7 milliseconds on average (1594x larger than append in CPAM). CPAM and PAM represent sequences using purely-functional trees, whereas ParallelSTL uses arrays (hence static).
be the tree size, an insertion needs to copy
(compressing all nodes in a BST) and (d) our PaC-tree (compressing
Binary search trees.
Figure 3. An illustration of (a) P-tree in PAM [11, 45] (regular BST), (b) B-tree (multi-way search tree), (c) C-tree [21] in Aspen (compressing all nodes in a BST) and (d) our PaC-tree (compressing all leaves in a BST) in CPAM. The orange nodes show a tree with keys 0-7 and 9. We then consider inserting a key 8. Blue nodes are what we need to create (copy or new) due to path-copying. Round
nodes are tree nodes each storing a single key, and square nodes are organized in blocks of size $O(B)$ (expected for C-trees). Letting
be the tree size, an insertion needs to copy $O(\log n)$ nodes in
P-tree, $O(B \log B n)$ in B-tree, and $O(B + \log(n/B))$ in C-tree (in expectation) on a PaC-tree.

the other hand concurrent trees support asynchronous updates, which PaC-trees do not—such updates are inherently non-functional. To support multiple concurrent updates, PaC-trees would require batching the update and applying them as a batch in parallel (fairly comparing concurrent and batched structures like PaC-tree seems challenging for this reason). We expect the use cases would be quite different.

Blandford and Blelloch developed tree structures for ordered sets that support compression [9]. They present space bounds that are similar to ours, in terms of relating the space of a difference encoded sequence to the space of the data structure. However they support a small fraction of the functionality described in our work.

Functional trees using path-copying date back to at least the early 1990s [2], and in the sequential setting have been studied by Kaplan and Tarjan [31] and Okasaki [40].

3 Preliminaries

Binary search trees. A binary search tree (BST) is either an empty node, denoted as $nil$, or a node consisting of a left BST $T_L$, a key $k$ (or with an associated value), and a right BST $T_R$.

denoted node$(T_L, k, T_R)$, where $k$ is larger than all keys in $T_L$
and smaller than all keys in $T_R$. We use $lc(T)$ and $rc(T)$ to extract the left and right subtrees of $T$, respectively, and use $k(T)$ to denote the key stored at $T$’s root. The size of a BST $T$, or $|T|$, is the number of nodes in $T$. The weight of a BST $T$, or $w(T)$, is $1 + |T|$. The height of a BST $T$, or $h(T)$, is 0 for nil, and $\max(h(lc(T)), h(rc(T))) + 1$ otherwise. A tree node is a leaf if it has no children, and a regular node otherwise. The left (right) spine of a binary tree is the path of nodes from the root to a nil node, always following the left (right) tree.

A weight-balanced tree, or $BB[\alpha]$ trees [39] is a BST where for every $T = node(T_L, v, T_R)$, $\alpha \leq \frac{w(T_L)}{w(T)} \leq 1 - \alpha$. We omit the parameter $\alpha$ with clear context. A weight-balanced tree $T$ has height at most $\log_{\frac{1}{\alpha}} w(T)$.

Parallelism. Our implementation of PaC-trees is based on nested fork-join parallelism [20, 27, 30]. We analyze our algorithms use work-span model based on binary-forking [12]. The work $W$ of a parallel algorithm is the total number of operations, while the span is the critical path length of its computational DAG. We use $s_1 || s_2$ to indicate that statements $s_1$ and $s_2$ can run in parallel. Almost all algorithms use divide-and-conquer to enable parallelism. Any computation with $W$ work and $S$ span will run in time $T < \frac{W}{P} + S$ on $P$ processors assuming shared memory and a greedy scheduler [14, 17]. We use log $n$ to denote $\log_{\frac{1}{\alpha}} (n+1)$ in the cost bounds.

Encoding schemes. We use Difference Encoding (DE) to encode integer keys. Given a sorted set of keys, $K$, the difference encoding scheme stores the differences between consecutive keys using an integer code, such as byte or \gamma codes. We only consider byte codes in this paper since they are cheap to encode and decode and do not waste much space compared to using \gamma codes [42].

Functional data structures. PaC-trees are purely functional data structures. In functional data structures values are immutable, so updates must be made by copying parts of the structure. For search trees, only the path to the update location needs to be copied. Hence for balanced trees of size $n$, single point updates such as inserts and deletes involve copying $O(\log n)$ nodes (Fig. 3(a)). This also applies to multi-point updates. For example, if a filter ends up removing a single element, only $O(\log n)$ nodes need to be copied. Functional trees can also easily support multversioning with low time and space overhead [7, 44]. Because the data are immutable, any operation accesses the tree in an isolated version. Updates can be applied in batches in parallel and yield a new version. This enables all read-only queries to be performed at the same time without being affected by ongoing (concurrent) updates. In addition to multversioning, functional data structures also allow for multiple histories.

Join-based algorithms. PaC-trees are implemented using the join-based approach [11, 13, 28, 43–45] first implemented in PAM [45]. In the framework, a variety of tree algorithms
are implemented based on two primitives, join and expose.\textsuperscript{4} Given a balancing scheme $\mathcal{S}$, the \texttt{join}(T_L, e, T_R) function returns a balanced tree $T$ satisfying $\mathcal{S}$ which has the same in-order values as node($T_L, e, T_R$). In other words, it concatenates $T_L$ and $T_R$ by an entry $e$ in the middle while preserving the balancing invariants (see Fig. 7 as an example of joining two PaC-trees). The expose($T_L$) function returns a triple ($T_L, e, T_R$) where $e \in T$ is an entry, $T_L$ and $T_R$ are two binary trees such that both $T_L$ and $T_R$ satisfy $\mathcal{S}$, are balanced with each other under $\mathcal{S}$, and $T_L$ ($T_R$) contains all keys in $T$ that go before (after) $e$ in $T$’s in-order value. It has been shown that on weight-balanced trees with $\alpha \leq 1 - 1/\sqrt{2}$, a join operation can be done in $O(\log \frac{n}{m})$ work [11], where $n = \max(|T_L|, |T_R|)$ and $m = \min(|T_L|, |T_R|)$.

Based on join and expose, many parallel tree algorithms can be expressed in a simple and elegant recursive style (see Fig. 5 and Fig. 6 for examples). We adopt the join-based approach in our implementation, and in particular carefully designed join and expose functions for PaC-trees. This greatly simplifies the implementation and correctness arguments of our algorithms. We give more details in Sections 5 and 6.

**Augmentation.** An augmented tree is a search tree where each node maintains an aggregated value (called augmented values) of all entries in its subtree. Typical examples would be a weighted sum, minimum or maximum of values, where we can obtain the augmented value in a node by combining augmented values of the children and itself. This generalizes to all associative operations. PaC-trees support generic user-defined augmentation for any associative operations. An example of PaC-tree with augmentation is shown in Fig. 4.

4 PaC-Trees

In this paper, we propose PaC-trees to support purely functional collections, which support parallelism, determinism, compression, augmentation, strong theoretical bounds, and multi-versioning. PaC-trees are purely functional. The base data structure of a PaC-tree is a weight-balanced BST. The internal nodes remain binary so they are cheap to copy. The leaves in a PaC-tree are organized in blocks of size $B$ to $2B$ for some parameter $B$. An illustration is shown in Fig. 3. If the blocks grow too large, they are split, and if they become too small they are merged with a neighboring node.

**Definition 4.1** (PaC-tree). A PaC-tree $PaC(\alpha, B, C)$, parameterized by the balancing factor $\alpha$, block size $B$, and encoding scheme $C$ satisfies the following invariants:

- **(Weight Balance)** For any tree node $v \in T$, $\alpha - \frac{1}{\sqrt{2}} \leq \frac{w(v)}{w(\pi)} \leq 1 - \alpha$, where $\alpha \leq 1 - \frac{1}{\sqrt{2}}$ is a constant, and $\pi$ is either $\text{lc}(v)$ or $\text{rc}(v)$. Unless mentioned otherwise, we use $\alpha = 0.29$.

- **(Blocked Leaves)** If $|T| \geq B$, each leaf $u \in T$ maintains $B$ to $2B$ entries in an array (called a block) using the encoding scheme $C$. Unless mentioned otherwise, we assume $C$ is empty, which means the entries are blocked without additional compression of the entries.

When the context is clear, we omit $\alpha, B$ and $C$ in the definition and simply call it a PaC-tree. We call a leaf node containing multiple entries in a PaC-tree a flat node, and a node containing a single entry a regular node. We say a PaC-tree (or a subtree) $T$ is a simplex tree if $|T| < B$, and thus $T$ only contains regular nodes. We say a PaC-tree (or a subtree) $T$ is a complex tree if $T$ contains both regular nodes and flat nodes. We define the expanded version of a PaC-tree $T$ (or a flat node $\pi$) to be a regular binary tree (without flat nodes), where all flat nodes in $T$ (or $\pi$ itself) are fully expanded as perfectly-balanced binary trees. In Fig. 4, we show an example of an expanded tree.

We now present the space bound of a PaC-tree. For integer keys, we can use difference encoding to bound the space.

**Theorem 4.2.** The total space of a PaC-tree $PaC(\alpha, B, C_{DE})$ maintaining a set $E$ of integer keys is $s(E) + O(|E|/B + B)$, where $C_{DE}$ is difference encoding, and $s(E)$ is the size needed for $E$ using difference encoding.

**Proof.** The space needed for a PaC-tree includes the regular nodes and the leaf nodes. First of all, when $|E| < B$, all entries are maintained in a simplex tree, taking $O(B)$ space. When $|E|$ $\geq B$, there are $O(|E|/B)$ regular nodes, each taking $O(1)$ space for meta-data (pointers, size, etc.). The total space used by regular nodes is $O(|E|/B)$. All the leaf nodes are organized in blocks. Let $A$ be an array that stores all keys in $E$ using difference encoding. Comparing the total size of all the blocks and $A$, the only extra space is the first element of each block (which cannot be compressed). There are $O(|E|/B)$ such blocks, and thus the extra space used is $O(|E|/B)$.

We note that this bound is deterministic, as opposed to the bound for C-trees (which only holds in expectation).
Furthermore, using known facts about difference encoding yields the following result, showing that PaC-trees yield a compact parallel representation of ordered sets [9].

**Corollary 4.3.** Given any set from $U = \{0, \ldots, m - 1\}$ with $|S| = n$, the total space of a PaC-tree PaC($a, B, C_{DE}$) maintaining $S$ is $O(n \log \frac{2n + \pi}{\pi})$ bits for $B = \Omega(\log n)$.

5 Algorithms

We now describe join-based algorithms on PaC-trees. To enable a general ordered map interface, we implement PaC-trees based on the PAM interface. PAM supports dozens of operations on sequences, sets, maps, and augmented maps, and it would require significant work to re-implement them all. Instead, we carefully redesigned join and expose such that all the other algorithms can remain the same as in PAM. In particular, none of the other algorithms have to deal with the blocked leaves or compression, which greatly simplifies the algorithm design and correctness arguments. We found that the overhead of this approach is not large, but for many frequently-used operations, we design special base cases for dealing with compressed nodes. These base cases can improve the performance by up to 6x (see Section 7). Some of the theoretical results also require special base cases (see the full version).

At a high-level, when exposing a flat node, the node is automatically expanded (using unfold), and similarly when join obtains a complex tree of size $B$ to $2B$, it is flattened (fold). An illustration of unfold and fold is shown in Fig. 7. We start with the join and expose algorithms. We then present the union algorithm as an example to illustrate join-based algorithms, and give the code for other functions in Fig. 6 and the full version of this paper. We focus on union as it is the core sub-routine used in applications such as inserting or deleting batches of vertices and edges in graphs, combining inner trees when constructing range trees, and updating documents in an inverted index, among others.

**Expose.** This function returns the left subtree, root data and the right subtree of a node $T$. For a regular node, this function just reads the child pointers and the root. For a flat node, this function first unfolds the tree into a perfectly balanced tree and then reads the corresponding data.

**Join.** Recall that the join function takes two trees $T_L$ and $T_R$, and a key $k$ (or a key-value) as input, and returns a balanced tree concatenating entries in $T_L$, $k$, and $T_R$ in order (see Fig. 7). In other words, when trees are used for ordered sets or maps, $k$ should be larger than all keys in $T_L$ and smaller than all keys in $T_R$. Pseudocode for join is shown in Fig. 5.

The algorithm first compares the weights of $T_L$ and $T_R$. When balanced, they are directly connected by $k$. The other two cases are symmetric so WLOG we assume $|T_L| > |T_R|$. In this case, the algorithm must attach $T_R$ in the right spine of $T_L$, which will be handled by join_right($T_L, k, T_R$). This algorithm first checks if $T_L$ and $T_R$ are balanced and connects them if so. Otherwise, it recursively calls join_right on $rc(T_L)$ and $T_R$, getting $T'$. If we re-attach $T'$ as $T_L$’s right child, we will get a “correct” output tree (modulo balance). We then use a single or double rotation to rebalance if necessary. It is known that either a single or double rotation can rebalance a weight-balanced tree in this situation [11]. This guarantees the weight balance invariant of PaC-trees.

To also guarantee the blocked leaves invariant, we add two conditions when calling node to create a new node with its left and right subtrees. Whenever a node with size $B$ to $2B$ is created, we fold the tree into a flat node. Whenever a node with size $2B$ to $4B$ is created, we extract the median of the tree as the root to re-distribute its two subtrees, such that both subtrees are flat nodes with (almost) the same size.

**Lemma 5.1.** The join function maintains the invariants of PaC-trees.

**Split.** For a PaC-tree $T$ and key $k$, split(T, k) returns a triple $(T_L, b, T_R)$, where $T_L$ (or $T_R$) is a tree containing all keys in $T$ that are less (greater) than $k$, and $b$ the entry of key $k$ if $k \in T$ (see Fig. 7). We first use expose(T) to get its left (right) subtrees lc(T) ($rc(T)$) and root key $k(T)$, and compare $k$ with $k(T)$. If $k = k(T)$, we simply return $(lc(T), k, rc(T))$. Otherwise WLOG we assume $k$ is smaller. In that case, the entire right subtree $rc(T)$ and the root $k(T)$ belong to $T_R$. We then split $lc(T)$ by $k$, getting $(L_L, b, L_R)$. By definition, all keys smaller than $k$ should be in $L_L$, and all keys larger than $k$ can be obtained by join($L_R, k(T), roman{rc}(T)$).

**Union.** Using join and split, we can implement set algorithms on two PaC-trees, such as union, intersection and difference. We describe union as an example (the other two are similar). This algorithm uses divide-and-conquer. At each level of recursion, $T_L$ is split by the root of $T_R$, breaking $T_L$ into two subsets with all keys smaller (larger) than $k(T_L)$, denoted as $L_L$ ($R_L$). Then two recursive calls to union are made in parallel. One unions $L(T_L)$ with $L_L$ (all keys smaller than $k(T_L)$), returning $T_{L_L}$, and the other one unions $R(T_L)$ with $R_L$ (all keys larger than $k(T_L))$, returning $T_{R_L}$. Finally the algorithm combines the results with join($T_{L_L}, k(T_2), T_{R_L}$).

**Other algorithms.** We show the pseudocode of other parallel algorithms in Fig. 6 and more in the full version of the paper. We omit the details as they are self-explanatory and all of them are exactly the same as in PAM, just by plugging in the new version of join and expose functions for PaC-tree. Almost all of them use divide-and-conquer to enable parallelism. We refer the reader to [45] for more details.

Importantly, all of our PaC-tree algorithms are theoretically efficient. We present the work-span bound in Table 1 and give a proof for union as an example in Section 6. Note that Lemma 5.1 ensures the correctness of the other algorithms, as their return values are always obtained by a join.

**Theorem 5.2.** All join-based algorithms on PaC-tree maintain the invariants of PaC-trees.
6  Theoretical Guarantees

In the following section we show work and span bounds for operations on PaC-trees. We assume the encoding scheme is empty, which means that to flatten or expand a block of size \( n \) costs \( O(n) \) work and \( O(\log n) \) span. If the encoding scheme is not parallelizable (e.g., for difference encoding), the span bound of the algorithms will be affected. We present more details in the full version of the paper.
We start with the cost of the join and split algorithms.

**Theorem 6.1.** Consider join on two PaC-trees $T_L, T_R$ and an key $k$. Let $n = \max(|T_L|, |T_R|)$ and $m = \min(|T_L|, |T_R|)$. If both $T_L$ and $T_R$ are complex trees, the algorithm takes $O(\log \frac{n}{m})$ work and span. If both $T_L$ and $T_R$ are simplex trees, the algorithm takes $O(n)$ work and $O(\log(n))$ span. Otherwise, the algorithm takes $O(B + n/B)$ work and $O(\log(n))$ span.

**Theorem 6.2.** Consider a split algorithm on a PaC-tree $T$. If $T$ is a complex tree, the work and span of split are $O(\log \frac{|T|}{m} + B)$ and $O(\log(|T|))$, respectively. If $T$ is a simplex tree, the work and span of split is $O(\log(|T|))$.

Due to page limit, we provide the proofs of Theorems 6.1 and 6.2 in the full version of this paper. Based on these results, we now analyze the cost of the set operations.

**Theorem 6.3.** Consider the union algorithm (and the closely related intersection and difference algorithms described in the full version) in Fig. 5 on two PaC-trees of sizes $m$ and $n \geq m$. The work and span for these algorithms are $O(m \log \frac{n}{m} + mB)$ and $O(\log(n \log m))$ respectively.

To prove the theorem, we first present some definitions and lemmas. First, note that all the work can be asymptotically bounded by the three categories below:

(1). **split work**: all work done by split (Line 14),
(2). **join work**: all work done by join (Line 17) or join2 in intersection and difference,
(3). **expose work**: all work done by expose (Line 13).

One observation is that the split work is identical among the three set algorithms. This is because the three algorithms behave the same on the way down the recursion when doing splits, and only differ in what they do at the base case and on the way up the recursion when building the output tree (see the other two set algorithms described in the full version).

We use $op$ to denote the set operation (one of union, intersection or difference). In these algorithms, the tree $T_1$ is split by the keys in $T_2$. We call $T_1$ the **decomposed tree** and $T_2$ the **pivot tree**, denoted as $T_d$ and $T_p$ respectively. Let $m = \min(|T_p|, |T_d|)$ and $n = \max(|T_p|, |T_d|)$.

**Lemma 6.4.** For each function call to op on trees $P \subseteq T_p$ and $D \subseteq T_d$, the work done by join (or join2) is asymptotically bounded by the work done by split.

We present the proof in the full version of this paper. Next, we prove the bounds for split work and expose work, respectively.

**Lemma 6.5.** The expose work is $O(\min(mB, n))$.

Proof. Expose costs $\Theta(B)$ when the subtree is a flat node, and $O(1)$ otherwise. At most $O(m)$ nodes in $T_p$ will split $T_d$, so the total cost is $O(mB)$. The cost is also no more than $O(n)$ since each node is involved in at most one expose, after which the flat node will be fully expanded. In summary the cost is $O(\min(mB, n))$. \(\Box\)

**Lemma 6.6.** The total split work is $O(m \log \frac{n}{m} + mB)$.

Proof. The total split work can be viewed as two parts: the total work to do by split functions to traverse and split non-flat nodes, and the work to expose and split the flat nodes. Note that here “non-flat nodes” include both regular nodes in complex trees, and all the nodes in expanded trees.

First of all, the total work to traverse and split all non-flat nodes can be asymptotically bounded by the split work when both $T_p$ and $T_d$ are considered to be fully expanded. This cost is $O(m \log \frac{n}{m})$ from the result for P-trees [11].

We then consider all work done by split functions on flat nodes. The only extra cost is the cost of unfold. Every node in $T_p$ will be used at most once to split $T_d$, which involves at most one unfold function with cost $O(B)$. There can be at most $O(m)$ nodes in $T_p$ used to split $T_d$. Thus the total unfold work in split is $O(mB)$.

Therefore in total the split work is $O(m \log \frac{n}{m} + mB)$. \(\Box\)

We can now prove Theorem 6.3.

**Proof.** (Theorem 6.3) Combining Lemmas 6.4 to 6.6 proves the work bound in Theorem 6.3. For the span, note that the algorithms need $O(\log \frac{|T_p|}{m} + B)$ rounds to reach a flat node, where the flat node will be expanded, taking $O(\log(B))$ span. Then the algorithm keeps recursing until a nil node is reached, which takes $O(\log B)$ rounds. In each of the recursive calls, we need $O(\log(|T_d|))$ span to deal with split and join. In total the span is $O(\log(m \log n))$. \(\Box\)

Note that the $O(mB)$ term can be expensive when $m$ is large. In fact, we can show a tighter bound using a more efficient (but more complicated) base case. We show the bound in Theorem 6.7, and defer the algorithm and proof to the full version. In our implementation, we use the version in Fig. 5, which has good performance in practice.

**Theorem 6.7.** There are algorithms for union, intersection and difference on two PaC-trees of sizes $m$ and $n (n \geq m)$ with work $O(m \log \frac{n}{m} + \min(n, mB))$ and span $O(\log n \log m)$.

## 7 Implementation

In this section, we describe CPAM, our implementation of PaC-trees. CPAM is built in C++, based on the PAM framework [45]. Our implementation of sequence and map primitives are mostly unchanged. Most of the changes are to introduce flat nodes, to handle folding and unfolding in join, to express the recursive functions using the expose primitive, and in some cases to add optimized base cases.

**Optimized Base Cases.** We first implemented union as in Fig. 5, which recursively calls expose to access the left and right subtrees. Although simple and theoretically efficient, in
practice unfolding flat nodes into expanded trees and recursing on these trees requires additional memory allocations, and potentially more cache-misses. We therefore designed a new sequential base-case for union when \( |T_L| + |T_R| < \kappa \), where \( \kappa \) is a configurable base-case granularity. Our base-case works by writing both \( T_L \) and \( T_R \) into a pre-allocated array \( A \) of size \( \kappa \) and merging them in-place to perform the union. It then constructs a PaC-tree from the result in A. Compared to the original version of union that only uses expose, using the special base-case with \( \kappa = 4B \) is 4.4x faster, and using \( \kappa = 8B \) is 6.7x faster (\( B = 128 \)). We observed similar improvements for some other commonly-used primitives such as filter, map_reduce, multi_insert, multi_delete, and intersection. We use \( \kappa = 8B \) in our experiments. We use a parallel granularity of \( 4B \), which is the threshold for forking parallel tasks in algorithms such as filter and union.

**Persistence and Memory Management.** CPAM uses a reference counting garbage collector for memory management. CPAM provides functional ordered maps, and thus by default does not modify the input trees. However, in certain cases an application may wish to modify a tree in-place to save memory, e.g., when updates and queries are separated. Although one could deal with in-place and functional updates separately, this is not attractive. Instead, we designed a simple approach to handle both cases using the same code, which we describe in the full version of the paper.

**Compression on Blocks.** CPAM makes it easy to apply user-specified encoding schemes. Our data structure is templated over a type representing a block encoding scheme (no encoding by default). To add a new encoding scheme, users provide a structure with methods that calculate the encoded size for a block, encode the elements into a buffer, and decode elements from an encoded buffer. This design allows users to specify encoding schemes based on the underlying data type or application, such as text compression. For example, it is easy to add new types of difference coding, e.g., using \( y \)-coding, which would obtain better space usage at the expense of worse running time [42].

8 Applications

In this section we describe four applications that we implement using CPAM. Our inverted index, and range and interval tree applications are based on the implementations from PAM [45]. Our graph processing application is based on Aspen [21]. We focus on the key features of the applications in the context of PaC-trees here.

**Inverted Index.** We implement a weighted inverted index, similar to those used in search engines. The inverted index maintains a top-level map from words to document lists (\( B = 128 \)). Each document list is a map from document id to an importance score (\( B = 128 \)). The document lists are augmented to maintain the highest importance score. The inverted index supports standard AND/OR queries over words, returning results by rank, and top-\( k \) (based on importance) queries. The document ids are compressed using difference encoding, requiring less than two bytes per document.

**2D Range Tree.** The two-dimensional range tree is a top-level map from \( x \)-coordinate to \( y \)-coordinate (\( B = 128 \)). The tree is augmented so that every internal node stores all \( y \)-coordinates in its subtree (this is itself a set represented as a PaC-tree with \( B = 16 \)). Updates can add and delete points, and queries can list of or count the points in a given rectangular range. The range tree supports count queries in \( O(\log^2 n) \) time, which can be batched to run in parallel.

**Interval Tree.** The interval tree maintains intervals over the number line, for example, representing the time of a TCP connection, or the time a user is logged into some service. A stabbing query can report all or any intervals that cross a given point. The intervals are represented as an augmented tree from left-coordinate to right-coordinate with \( B = 32 \). The augmentation maintains the maximum right-coordinate in the subtree. This allows stabbing queries in time \( O(k \log n) \) where \( k \) is the number of intervals requested or returned (whichever is less). Intervals can be inserted or deleted in \( O(\log n) \) time and can be batched to run in parallel.

**Graph Processing.** Graphs are represented as a two-level structure similar to the inverted index, with a top-level augmented tree (the vertex tree) from vertices to edge lists (\( B = 64 \)). Each edge list is a map from neighbor-id to an edge-weight (or empty when unweighted) called an edge tree (\( B = 64 \)). The augmentation on the vertex tree maintains the total number of edges in the graph. We focus on unweighted graphs in this paper but note that our implementation also supports weights. As with inverted indices, using difference encoding allows us to store an edge using just 2–3 bytes on average including the bytes used for regular nodes.

On top of this representation, we implement graph algorithms using the Ligra interface [41], including breadth-first search, maximal independent set, and single-source betweenness centrality. Our implementations are based on the ones in Aspen and GBBS [22, 23]. We design parallel batch-updates for our representation, which are applicable in graph-streaming and batch-dynamic graph algorithms.

9 Experiments

**Experimental Setup.** We run experiments on a 72-core Dell PowerEdge R930 (with two-way hyper-threading) with 4 \( \times \) 2.4GHz Intel 18-core E7-8867 v4 Xeon processors (with a 4800MHz bus and 45MB L3 cache) and 1TB of main memory. Our programs use a work-stealing scheduler for parallelism [10]. We use numactl -i all to balance the memory allocations across the sockets for parallel executions. Unless otherwise mentioned, all of the reported numbers are run on 72 cores with hyper-threading.
Table 2. Microbenchmark results. We fix \( B = 128 \) for PaC-trees. \( n \) is the tree size. For set functions and multi-insert, \( m \leq n \) is the size of the other set (batch). For other functions, \( m \) is the number of queries tested. \( T_i \) is the sequential running time. \( T_{144} \) is parallel running time using 72 cores (144 hyperthreads). Diff means difference encoding. We highlight the best parallel running time (or size) per experiment in green and underlined.

<table>
<thead>
<tr>
<th>( n ) ( m )</th>
<th>PaC-tree</th>
<th>PaC-tree (Diff)</th>
<th>P-tree (PAM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No augmentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size (GB)</td>
<td>10^8</td>
<td>1.61</td>
<td>0.926</td>
</tr>
<tr>
<td>Build</td>
<td>10^8</td>
<td>0.186</td>
<td>29.8</td>
</tr>
<tr>
<td>Union</td>
<td>10^8 ( B )</td>
<td>5.33</td>
<td>0.088</td>
</tr>
<tr>
<td>Union</td>
<td>10^8 ( B )</td>
<td>1.09</td>
<td>0.021</td>
</tr>
<tr>
<td>Intersect</td>
<td>10^8 ( B )</td>
<td>4.35</td>
<td>0.065</td>
</tr>
<tr>
<td>Difference</td>
<td>10^8 ( B )</td>
<td>3.00</td>
<td>0.055</td>
</tr>
<tr>
<td>Map</td>
<td>10^8 ( B )</td>
<td>0.859</td>
<td>0.037</td>
</tr>
<tr>
<td>Reduce</td>
<td>10^8</td>
<td>0.306</td>
<td>0.018</td>
</tr>
<tr>
<td>Filter</td>
<td>10^8</td>
<td>0.997</td>
<td>0.028</td>
</tr>
<tr>
<td>Find</td>
<td>10^8</td>
<td>1.05</td>
<td>1.17</td>
</tr>
<tr>
<td>Insert</td>
<td>10^8 ( B )</td>
<td>0.829</td>
<td>0.036</td>
</tr>
<tr>
<td>Multi-Insert</td>
<td>10^8 ( B )</td>
<td>18.8</td>
<td>0.332</td>
</tr>
<tr>
<td>Range</td>
<td>10^8 ( B )</td>
<td>11.5</td>
<td>0.318</td>
</tr>
</tbody>
</table>

With augmentation

<table>
<thead>
<tr>
<th>( n ) ( m )</th>
<th>PaC-tree</th>
<th>PaC-tree (Diff)</th>
<th>P-tree (PAM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (GB)</td>
<td>10^8</td>
<td>1.63</td>
<td>0.936</td>
</tr>
<tr>
<td>Build</td>
<td>10^8</td>
<td>0.197</td>
<td>28.7</td>
</tr>
<tr>
<td>Union</td>
<td>10^8 ( B )</td>
<td>5.52</td>
<td>0.098</td>
</tr>
<tr>
<td>AugRange</td>
<td>10^8 ( B )</td>
<td>12.3</td>
<td>0.331</td>
</tr>
<tr>
<td>AugFilter</td>
<td>10^8</td>
<td>12.2</td>
<td>0.333</td>
</tr>
</tbody>
</table>

Overview of Results We show the following experimental results in this section.

- PaC-trees are competitive with PAM for microbenchmarks (Section 9.1) and applications including inverted indices (Section 9.2) and 2D range queries and 1D interval queries (Section 9.3) while using 2.1x–7.8x less space.

- Varying the block size \( B \) for an PaC-tree trades off performance for space efficiency (Section 9.1). For even a modest value of \( B = 128 \), PaC-trees use only 1% more space than a (static) compressed array.

- For graph processing and streaming, CPAM uses 1.3–2.6x less space compared to Aspen, and is almost always faster than Aspen in all tested graph algorithms (Section 9.4).

9.1 PaC-Tree Performance

We begin by studying the performance and space of PaC-trees on a set of microbenchmarks and compare with P-trees from PAM. All experiments in this section use maps and augmented maps where the keys and values are both 64-bit integers. Unless otherwise mentioned PaC-trees use \( B = 128 \).

Microbenchmark Performance. Table 2 shows the results on PaC-trees, PaC-trees with difference-encoding (DE), and P-trees for a representative subset of the map and sequence primitives. The speedups for both types of PaC-trees range from 28.7–101x and are largest for the version using DE due to additional work for difference encoding. In absolute running time, PaC-trees with DE are usually slower than PaC-trees due to compression and decompression costs, but the overhead is mostly within 10%.

In most of the primitives tested, PaC-trees are faster than P-trees while also using 2.5x less space. For example, PaC-trees are 1.68x faster than P-trees in union on two trees of sizes \( 10^8 \). We note that in this case, the union processes the entirety of both input trees, and so the more cache-friendly processing of blocks in PaC-trees results in lower time. However, if sizes of the two trees are different, the work for union only depends on the smaller size. In this case, since the cost of union using PaC-trees has an additional \( O(mB) \) term compared with P-trees, PaC-trees are 5.5x slower than P-trees. However, we expect better performance for smaller block sizes \( B < 128 \), which we discuss next.

Effect of Varying \( B \) on Performance. Fig. 9 shows the results of varying the block size \( B \), on the performance of various operations. Most operations obtain speedups as \( B \) is increased up until \( B = 16 \). For the sequential operations, such as find and range, we see a steady increase in the running time for \( B > 16 \) and see a similar trend for Union-imbalance, which takes the union of trees with \( 10^8 \) and \( 10^5 \) elements. This slowdown with increasing \( B \) is due to the extra \( O(mB) \) term in the work of union. For the smallest block size \( B = 1 \), our running time matches that of P-trees on this operation.
Table 3. Build and query times and space usage in GiB for inverted index, interval tree, and range tree applications. \(T_1\) is the single-thread time, \(T_{144}\) is the 72-core time using hyper-threading, and Spd. is the parallel speedup. The best parallel running time (or size) is highlighted in green and underlined per experiment.

<table>
<thead>
<tr>
<th>Library</th>
<th>Space Method</th>
<th>(n)</th>
<th>(m)</th>
<th>(T_1)</th>
<th>(T_{144})</th>
<th>Spd.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaC-tree</td>
<td>Build</td>
<td>10^8</td>
<td>-</td>
<td>746</td>
<td>9.73</td>
<td>76.6</td>
</tr>
<tr>
<td></td>
<td>Query</td>
<td>10^8</td>
<td>10^8</td>
<td>341</td>
<td>4.46</td>
<td>76.4</td>
</tr>
<tr>
<td>PaC-tree (D)</td>
<td>Build</td>
<td>10^8</td>
<td>-</td>
<td>754</td>
<td>9.81</td>
<td>76.8</td>
</tr>
<tr>
<td></td>
<td>Query</td>
<td>10^8</td>
<td>10^8</td>
<td>367</td>
<td>5.32</td>
<td>68.9</td>
</tr>
<tr>
<td>P-tree (PAM)</td>
<td>Build</td>
<td>10^8</td>
<td>-</td>
<td>575</td>
<td>8.86</td>
<td>64.9</td>
</tr>
<tr>
<td></td>
<td>Query</td>
<td>10^8</td>
<td>10^8</td>
<td>313</td>
<td>4.48</td>
<td>69.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Library</th>
<th>Space Method</th>
<th>(n)</th>
<th>(m)</th>
<th>(T_1)</th>
<th>(T_{144})</th>
<th>Spd.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaC-tree</td>
<td>Build</td>
<td>10^8</td>
<td>-</td>
<td>10.9</td>
<td>0.179</td>
<td>60.8</td>
</tr>
<tr>
<td></td>
<td>Query</td>
<td>10^8</td>
<td>10^8</td>
<td>60.8</td>
<td>0.525</td>
<td>115.8</td>
</tr>
<tr>
<td>P-tree (PAM)</td>
<td>Build</td>
<td>10^8</td>
<td>-</td>
<td>11.6</td>
<td>0.271</td>
<td>42.8</td>
</tr>
<tr>
<td></td>
<td>Query</td>
<td>10^8</td>
<td>10^8</td>
<td>54.3</td>
<td>0.628</td>
<td>86.4</td>
</tr>
<tr>
<td>PaC-tree</td>
<td>Q-Sum</td>
<td>10^8</td>
<td>10^8</td>
<td>54.2</td>
<td>0.629</td>
<td>86.1</td>
</tr>
<tr>
<td></td>
<td>Q-All</td>
<td>10^8</td>
<td>10^8</td>
<td>7.20</td>
<td>0.266</td>
<td>27.0</td>
</tr>
<tr>
<td>P-tree (PAM)</td>
<td>Build</td>
<td>10^8</td>
<td>-</td>
<td>169</td>
<td>2.84</td>
<td>59.6</td>
</tr>
<tr>
<td></td>
<td>Q-Sum</td>
<td>10^8</td>
<td>10^8</td>
<td>60.7</td>
<td>0.735</td>
<td>82.5</td>
</tr>
<tr>
<td></td>
<td>Q-All</td>
<td>10^8</td>
<td>10^8</td>
<td>21.6</td>
<td>0.552</td>
<td>39.1</td>
</tr>
</tbody>
</table>

9.2 Inverted Index

Next, we study our performance on the inverted index application. We run the application on documents derived from a large Wikipedia dataset also used by PAM for a fair comparison. The dataset is processed by removing all markup, converting characters that are not alphanumeric to whitespace and making all words case insensitive [45]. The processed dataset contains 1.94 billion words over 8.13 million documents. Like PAM, our evaluation measures the performance of (1) building an index over (words, doc_id, weight) triples and (2) running queries that fetch the posting lists for two words, compute the intersection of the lists, and select the top 10 documents by weight.

Table 3 shows the results of the experiment. For building the index, our implementation achieves 76x speedup and our parallel running times are comparable with those of PAM (at most 1.1x slower). For the queries, we observe that the unencoded trees achieve essentially the same parallel time as PAM, whereas the difference encoded trees are 1.18x slower due to the higher cost of intersection operations in our difference encoded implementation. The space usage using PaC-trees is much smaller than that of PAM, being 3.84x smaller without encoding and 7.81x smaller using a custom encoder that combines difference encoding for the keys with byte-encoding for the integer values (weights).

9.3 Interval and Two-Dimensional Range Trees

We benchmark our interval and two-dimensional range trees as in PAM [43]. We build our interval tree on \(10^8\) intervals, and for queries run stabbing queries over \(10^8\) points in parallel. We observe that both building and querying the trees achieves good parallel speedup (60–115x). PaC-trees are 1.51x faster than PAM in construction, and is 1.19x faster for queries. Overall we find that PaC-trees enable better performance than PAM while using 4.37x less space.

We build our range trees on \(10^8\) uniformly random points in the plane between \((0,0)\) and \((1e8, 1e8)\). We run two types of queries: the first count the number of points in the range (Q-Sum), and the second returns all points in the range. We tuned the window sizes used in our queries to match the settings evaluated by PAM (around \(10^6\) points returned per query). Both PaC-trees and P-trees build the data structure in a similar amount of time. PaC-trees achieve better performance than P-trees for both queries, being 1.16x faster for Q-Sum and 1.96x faster for Q-All queries, likely due to requiring fewer cache-misses when processing the tree to output the points within a given range. The range tree application using PAM has previously been compared with range trees in CGAL [38] and was shown to outperform it [43].

For space usage, PaC-trees result in 2.18x less space compared to PAM. We note that 95% of the space used in PAM is for the P-trees stored as augmented values in each node (representing the union of the \(y\)-coordinates in the subtree).

Space Usage. For \(B = 128\), PaC-trees obtain a 2.48x reduction in space usage compared to using P-trees, and a further 1.73x reduction in space usage by using difference encoding. The \(10^8\) pairs stored in the experiments require 1.6GB of memory to represent as a single flat array, which is also a lower bound for the space usage of a search tree structure. To understand how close PaC-trees come to this lower bound, we study the space usage of unencoded maps using PaC-trees as a function of the block size \(B\) (Fig. 10). Using \(B = 32\), PaC-trees are only 1.05x larger than the lower bound and using \(B = 128\), it is just 1.01x larger than the lower bound. For \(B = 128\), just 1.1% of the allocated memory is used for regular nodes and metadata in the flat nodes. These savings are obtained without using any additional encoding. Applying difference encoding improves the space by 1.77x over the unencoded trees and the array lower bound, and is only 1.03x larger than the space used to difference encode all of the keys in a single array, leaving the values uncompressed, which is a lower bound for a search tree structure using difference encoding for such input.

Using PaC-trees requires much lower space overhead for augmentation compared to P-trees (Fig. 10). For P-trees, adding 8 byte augmented values increases the size of the maps by 20%, whereas PaC-trees (both with and without difference encoding) using \(B = 128\) incurs only a 1% increase in space for the augmented values. The savings comes from only storing a single augmented value per flat node, which only uses extra space proportional to \(n/B\) augmented values.
The majority of our savings come from compressing the augmented trees using PaC-trees which results in a 2.53x less space for the inner trees, and 2.18x less space overall.

9.4 Graph Processing and Graph Streaming

Our last set of experiments study the performance of PaC-trees for a set of standard benchmarks from the graph processing and graph streaming literature. Our evaluation roughly follows Aspen’s and we compare our performance and space usage with that of Aspen and its C-tree implementation.

Graph Data and Space Usage. Most of the graphs we study are Web graphs and social networks which are low-diameter graphs that are frequently used in practice. To also test on high-diameter graphs, we ran our implementations on a road network. Complete details about our inputs are in the full version of the paper. Table 4 shows information about our graphs, including the number of vertices, edges, and space used.

We evaluate five graph representations including using PAM, Aspen, PaC-tree with or without difference encoding, and GBBS. Aspen uses C-trees as edge trees and leaves vertex trees uncompressed using P-trees. GBBS is a state-of-the-art static graph processing library which represents graphs as an array storing all vertices in the current graph. The idea is that instead of accessing edges for a vertex through the vertex tree (performing tree traversal), algorithms directly access edge trees through the flat snapshot.

Table 5 shows performance results for three of our graph datasets. Across all three kernels our implementations are 1.12x faster than Aspen’s implementations on average. We observe that flat snapshots can be generated 2.09–3.02x faster in CPAM due to PaC-trees requiring fewer cache-misses to traverse than P-trees when creating flat snapshot array. We note that the implementation of edgeMap and other primitives from Ligra (including constants and other tuning parameters) are exactly the same in both CPAM and Aspen. Aspen also difference encodes in its edge trees (represented using C-trees). The performance improvements that we observe are therefore a result of PaC-trees providing faster flat snapshots, and having better balance in chunk sizes compared to the randomized approach used in C-trees.

Concurrent Updates and Queries. Our last experiment concurrent updates and queries on graphs. The experiment performs $n$ undirected edge insertions drawn from the rMAT generator (details provided in the full version). We use a batch size of 5 in the updates (10 directed edges are inserted per batch). We then spawn two parallel jobs, one performing the updates one batch after the other, and the other performing BFS queries, one after the other. Both the updates and queries are parallel (i.e., they internally make use of parallelism).
10 Conclusion

We have presented PaC-trees, a deterministic compressed ordered map data structure and an implementation of the structure in a library CPAM. The important features of PaC-trees and its implementation in CPAM include the following.

- It is purely functional allowing for persistent snapshots while updates are being made, and safe for parallelism.
- It supports sequences, ordered sets, ordered maps, and augmented maps, with a wide variety of functions on them.
- It provides theoretical bounds on work, span, and space.
- It achieves fast sequential time and gets up to 100x speedup on 72 cores with 144 hyperthreads.
- It achieves memory usage that is close to a compressed array and up to an order of magnitude smaller than PAM.
- It is internally memory managed using reference counting.
- It is backward compatible with PAM.
- It has been used to implement the full functionality of Aspen while improving runtime and/or space.

For future work, we are interested in extending PaC-trees to support higher-fanout internal nodes, similar to B-trees, which would allow users to improve query latency at the expense of increased work when performing updates. Other future work includes applying PaC-trees to improve space utilization in databases, and to improve the performance of collection-based applications using non-volatile memory.

Acknowledgement

This work was supported by the National Science Foundation grants CCF-1901381, CCF-1910030, CCF-1919223, CCF-2103483, and CCF-2119352. We are grateful to the PLDI reviewers for their helpful feedback and suggestions.

References


[34] Laxman Dhulipala, Guy E. Blelloch, and Julian Shun. 2020. The Graph Based Benchmark Suite (GBBS). In ACM Symposium on Principles and Practice of Parallel Programming (PPoPP).


