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

Laxman Dhulipala
University of Maryland
laxman@umd.edu

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

Yan Gu
UC Riverside
ygu@cs.ucr.edu

Yihan Sun
UC Riverside
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.

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 [54]. 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 [22, 25, 37, 42].

Recent work [52] has developed a purely functional library, PAM, for representing sequences, ordered sets, ordered maps, and augmented maps (defined in [52]) 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...
Table 1. Primitives from the Sequence, Set, and Map interfaces in CPAM, including the work and span bounds. Note that primitives marked with \(^\dagger\) 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(^\dagger)</td>
<td>(O(\log n + B))</td>
<td>(O(\log n))</td>
</tr>
<tr>
<td>Reverse(^\dagger)</td>
<td>(O(n))</td>
<td>(O(\log n))</td>
</tr>
<tr>
<td>Set and Map</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Build</td>
<td>(O(n \log n))</td>
<td>(O(\log n))</td>
</tr>
<tr>
<td>Next/Previous</td>
<td>(O(\log n + B))</td>
<td>(O(\log n))</td>
</tr>
<tr>
<td>Range</td>
<td>(O(\log n + B))</td>
<td>(O(\log n))</td>
</tr>
<tr>
<td>Insert</td>
<td>(O(\log n + B))</td>
<td>(O(\log n))</td>
</tr>
<tr>
<td>Union</td>
<td>(O(m \log \frac{m}{n} + \min(mB, n)))</td>
<td>(O(\log n \log m))</td>
</tr>
<tr>
<td>Intersect</td>
<td>(O(m \log \frac{m}{n} + \min(mB, n)))</td>
<td>(O(\log n \log m))</td>
</tr>
<tr>
<td>Difference</td>
<td>(O(m \log \frac{m}{n} + \min(mB, n)))</td>
<td>(O(\log n \log m))</td>
</tr>
</tbody>
</table>

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 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 [43] 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 [49]. 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 maintains an aggregate of the values of its subtree (see more details in Section 3). The aggregation function is declared as part of the type of the tree. Augmentation is useful in many applications, and indeed we use it in all of the applications we describe later. PaC-trees store an augmented value per internal node, and one for each block at the leaves. Storing one value per block significantly reduces space relative to P-trees in PAM, which store a value for every element.

To demonstrate the effectiveness of PaC-trees, and their implementation in CPAM, we measure performance and space usage on (1) a collection of microbenchmarks that directly use some of the functions supported by the library, and (2) a handful of real-world applications.

For the microbenchmarks, we compare the performance of CPAM to PAM, and for sequences to the Intel implementation of the C++17 parallel STL library [33] (ParallelSTL). ParallelSTL is a highly optimized library supporting only sequences based on arrays. A summary of the results for sequences is given Fig. 2, and details including performance of ordered maps, and augmented maps are given in Section 10. Compared to PAM, CPAM achieves significantly better performance due to the reduced memory footprint, and hence reduced number of cache misses, while only requiring about 1/4-th as much space even without compression. Compared to ParallelSTL, CPAM has similar performance on operations.
that visit the whole sequence, like \texttt{reduce}, but is significantly slower on \texttt{nth} since it requires \(O(\log n + B)\) work as opposed to \(O(1)\) for a random array access for \texttt{ParallelSTL}. On \texttt{append} \texttt{CPAM} is significantly faster since it requires \(O(\log n + B)\) work to join to trees instead of \(O(n)\) required by \texttt{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, \texttt{CPAM} achieves competitive performance to \texttt{PAM} while using 2.1x–7.8x less space. For graph processing, we compare to an existing system \texttt{Aspen} [25] that represents graphs using trees. \texttt{CPAM} uses 1.3–2.6x less space compared to \texttt{Aspen}, and is almost always faster than \texttt{Aspen} in all tested graph algorithms.

The main contributions of this paper are:

- A new functional data structure, \texttt{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 \texttt{PaC-trees} as a library, \texttt{CPAM}, supporting the full functionality of \texttt{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 \texttt{P-trees} and their C++ implementation in \texttt{PAM} [52]. 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.

\texttt{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, 16, 30, 38]. 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.\footnote{We also do not know of work that then compresses within the blocks.}

\texttt{Aspen} [25] 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 \texttt{CPAM}. The tree representation in \texttt{Aspen} is also very different. It randomly selects elements from the collection to be \texttt{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. \texttt{PaC-trees} do not require randomization, and have stronger theoretical bounds for primitive operations such as \texttt{union} than the bounds provided by \texttt{C-trees} in \texttt{Aspen}. We use \texttt{CPAM} to implement the full functionality of \texttt{Aspen} and compare to \texttt{Aspen} in Section 10.5.
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 [54] system supports a functional interface for collections. However it has significant differences. Firstly it only supports unordered sets. Secondly 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, 18, 20, 28, 39, 44]. 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 work support range queries [5, 29], or arbitrary queries on a snapshot [53]. On 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 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. 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 [36] and Okasaki [47].

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

Parallelism. Our implementation of PaC-trees is based on nested fork-join parallelism [23, 31, 35]. 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 \(s_1\) \(\parallel s_2\) 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 [15, 19]. We use \(\log n\) to denote \(\log_2(n+1)\) 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 [49].

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 multiversioning with low time and space overhead [7, 51]. 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 multiversioning, functional data structures also allow for multiple histories.

Join-based algorithms. PaC-trees are implemented using the join-based approach [11, 14, 32, 50–52] first implemented in PAM [52]. In the framework, a variety of tree algorithms are implemented based on two primitives, joint and expose.\(^4\) Given a balancing scheme \(S\), the joint\((T_L, e, T_R)\) function returns a balanced tree \(T\) satisfying \(S\) 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\((T_i)\) function returns a triple \((T_L, e, T_R)\), where \(e \in T\) is an entry, \(T_i\) and \(T_R\) are two binary trees such that both \(T_L\) and \(T_R\) satisfy \(S\), are balanced with each other under \(S\), and \(T_L\) \((T_R)\) contains all keys in

\(^3\)Their 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 Section 10.2)

\(^4\)PAM did not explicitly use expose as a primitive, but only conceptually treated it as a primitive.
$T$ that go before (after) $e$ in $T$’s in-order value. It has been shown that on weight-balance trees with $\alpha \leq 1 - 1/\sqrt{2}$, a join operation can be done in $O(\log \frac{n}{m} \alpha)$ 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, Fig. 6, and Fig. 10 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 $T$, $\alpha \leq \frac{w(v)}{w(v) + \omega(v)} \leq 1 - \alpha$, where $\alpha \leq 1 - \frac{1}{\sqrt{2}}$ is a constant, and $\omega(v)$ is either $le(v)$ or $re(v)$. 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 $v$) to be a regular binary tree (without flat nodes), where all flat nodes in $T$ (or $v$ 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 $PaC(\alpha, B, C_{DE})$ maintaining $S$ is $O(n \log \frac{nm}{n})$ 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
1. **fold(T) {**
2. flatten T into array A
3. (encoding if needed)
4. return A; )
5. **unfold(A) {**
6. /* return a perfectly balanced tree
7. from sorted array A */
8. expose(T) { }
9. if (isflat(T)) { }
10. return T = unfold(T);
11. return (k(T'), k(T'), rc(T'));
12. else return (k(T), k(T), rc(T));
13. **join(T_L, k_T_R) {**
14. if (heavy(T_L, T_R))
15. return join_right(T_L, k_T_R);
16. if (heavy(T_R))
17. return join_left(T_L, k_T_R);
18. return node(T_L, k_T_R); }

9. **join_left is symmetric */
10. **join_right(T_L, k_T_R) {**
11. (l', k', e) = expose(T_L);
12. if (balance(|T_L|, |T'|))
13. return node(T_L, k_T_R);
14. T' = join_right(c, T_R);
15. (l_k, r_1) = expose(T');
16. if (balance(|T'|))
17. return node(l, k', T');
18. if (balanced(|l|, |l|)) and
19. (balanced(|l| + |l'|, |l'|))
20. return rotateleft(node(l, k', T'));
21. else return rotateleft(node(l, k',
22. rotafter(T')));
23. **join2(T_L, T_R) {**
24. if (T_L = nil) return T_R;
25. if (T_L = nil) return T_R;
26. (T_L, m, _) = split(T_L, last(T_L));
27. return join(T_L, m, T_R); }
28. **filter(T, f) {**
29. (n = 0) return nil;
30. if (T == nil) return nil;
31. L = from_sorted(A, n/2) ||
32. (k == (m)) return (L, m, R);
33. R = from_sorted(A+n/2, n-n/2);  
34. if (l < m) return (L, m, R);
35. L = filter(L, f) ||
36. (L, b, LR) = split(L, k);
37. return (L, b, join(LR, m, R));
38. else {  
39. (LR, b, RR) = split(R, k);
40. return (join(L, m, LR), b, RR);  
41. }  

**Figure 5. Primitives on PaC-trees.** All codes are functional (e.g. rotates copy nodes).

1. from_sorted(A, n) {
2. if (n = 0) return nil;
3. if (T == nil) return nil;
4. L = from_sorted(A, n/2) ||
5. (k == (m)) return (L, m, R);
6. R = from_sorted(A+n/2, n-n/2);  
7. return node(L, A[n/2], R);  
8. build(A, n) {
9. parallel_sort(A, n);
10. return from_sorted(A, n);  
11.
12. }

**Figure 6. Examples of parallel algorithms on PaC-trees.** “||” indicates calls that are made in parallel.

**Figure 7. Illustration of primitives on PaC-trees.** For Figures (a)–(d), B = 3. For Figures (e)–(f), B = 2. Fig. (a): the expose function on a regular node and the node function to obtain a regular node when the output tree size is larger than 4B. Fig. (b): the expose function on a flat node and the node function to obtain a flat node when the output tree weight is between B and 2B. Fig. (c): the node function to obtain a flat node when the output size is between 2B and 4B. Fig. (d): fold and unfold functions. Fig. (e): join function on two regular nodes and its corresponding split function. Fig. (f): join function on a regular node and a flat node and its corresponding split function.

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 8). Some of the theoretical results also require special base cases (see Section 6.1).

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 Fig. 10. 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 sets of 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 $\text{join}_\text{right}(T_L, k, T_R)$. This algorithm first checks if $T_L$ and $T_R$ are balanced and connects them if so. Otherwise, it recursively calls $\text{join}_\text{right}$ 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.

**Split.** For a PaC-tree $T$ and key $k$, $\text{split}(T, k)$ returns a triple $(T_L, b, T_R)$, where $T_L$ ($T_R$) 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($L_R, k(T), rc(T)$).

**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 $k(T)$, denoted as $L_L (R_L)$. Then two recursive calls to union are made in parallel. One unions $L(T_L)$ with $L_1$ (all keys smaller than $k(T)$), returning $L_L$, and the other one unions $R(T_R)$ with $R_1$ (all keys larger than $k(T))$, returning $R_R$. Finally the algorithm combines the results with join($T_L, k(T), R_R$).

**Other algorithms.** We present the pseudocode for the other two set algorithms (intersection and difference) in Fig. 10. We also show the code for three other useful functions, multi_insert, map and reduce in Fig. 8. We note that these algorithms are exactly the same as in PAM, by extracting out the semantics of join and expose.

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 Section 6.2.

We start with the cost of the join and split algorithms.

**Theorem 6.1.** Consider a join algorithm 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 $O(\log^2 m)$ work and span. If both $T_L$ and $T_R$ are simple trees, the algorithm takes $O(B)$ work and $O(\log B)$ span. Otherwise, the algorithm takes $O(B + n/B)$ work and $O(\log n)$ span.
of tree nodes we need to visit on the right spine is $T_t$ node balanced with right spine of the tree until finding a subtree $t$ decompresses a leaf. Note that the algorithm will follow the right spine of the tree to find a leaf. This proves that the total number of tree nodes we need to visit on the spine is $O(\log \frac{n}{m})$.

For two simplex trees, the work is no more than copying both $T_L$ and $T_R$ and concatenating them, which is $O(B)$ work and $O(\log B)$ span.

If $T_L$ is a complex tree and $T_R$ is a simplex tree, we need to first follow the right spine to find a leaf node $l$ in $T_L$, which takes $O(\log \frac{n}{m})$ time. Then it combines the leaf node with $T_R$, which flattens both $l$ and $T_R$, concatenates them, and rebalance the result. This process takes no more than $O(B)$ work and $O(\log B)$ 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

---

**Figure 8.** Some other algorithms on PaC-trees.

```plaintext
1 1 m_ins_helper(T, A, m) {
2 2 if (T == nil) return from_sorted(A, m); // f is an associative binary operation
3 3 if (n == 0) return T;
4 4 (L, K, R) = expose(T);
5 5 s = binary_search(A, m, k);
6 6 if (A[x] == k) b = 1;
7 7 T_L = m_ins_helper(L, A, s); // f is the identity
8 8 T_R = m_ins_helper(R, A+s+b, m-s-b);
9 9 return join(T_L, k, T_R); }
10 10 multi_insert(T, A) {
11 11 A' = parallel_sort(A);
12 12 return m_ins_helper(T, A', |A|); }
13
14
15
```

**Figure 9.** Some useful primitives for join-based algorithms on PaC-trees for more efficient set algorithms.

```plaintext
16 16 node(l, k, r, expand=false) {
17 17 if (l == nil) {
18 18 let the left child of k be l;
19 19 let the right child of k be r;
20 20 if (expand) return k;
21 21 if (|k| > 4B) return k;
22 22 if (B <= |k| <= 4B) return fold(k);
23 23 else { // 2B < |k| <= 4B
24 24 // redistribute k's both subtrees to be flat nodes with |k|/2 entries
25 25 return k; }
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
```
join respectively.

on two PaC-trees of sizes $m$ and $n \geq m$. The work and span for these algorithms are $O(m \log \frac{n}{m} + B)$ and $O(n \log n \log m)$ 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 16).
2. **join work**: all work done by join (Line 19) or join2 in intersection and difference.
3. **expose work**: all work done by expose (Line 15).

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.

We use $o$ 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 $m = \min(|T_D|, |T_P|)$ and $n = \max(|T_D|, |T_P|)$.

**Lemma 6.4.** For each function call to $o$ 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.

**Proof.** Assume the return value is $R$.

First of all, the work of split is $\Theta(|D| + B)$. Note that the work of join (or join2) can be bounded by $O(|R| + B)$.
Notice that difference returns the keys in \( D \setminus P \). Thus for both intersection and difference we have \( R \subseteq D \). Therefore \(|R| \leq |D|\), which means the work done by \( \text{join} \) or \( \text{join2} \) is no more than the work done by split.

For union, first of all, we always call \( \text{join} \) instead of \( \text{join2} \). If \(|P| \leq |D|\), then \(|R| \leq 2|D|\). \( \text{join} \) costs work \( O(\log |R| + B) = O(\log |P| + B) \), which is no more than \( \Theta(\log |D| + B) \).

Consider \(|P| > |D|\). The subtrees \( lc(P) \) and \( rc(P) \), which are used in the recursive calls, have size at least \( a|P| \) and at most \((1 - \alpha)|P|\). After combining with a subset of elements in \( D \) (which has size smaller than \(|P|\)), the return value of each recursive call should have size at least \( a|P| \) and \((2 - \alpha)|P|\). Denote these two trees from recursive calls as \( t_l \) and \( t_r \), respectively. Note that \( \alpha \) is a constant, so the difference of size between \( t_l \) and \( t_r \) is also no more than a constant. WLOG assume \(|t_l| \geq |t_r|\). In the following, we discuss different cases of whether \( t_l \) and \( t_r \) are complex or simplex trees. We will show that, in all cases, joining \(|t_l|\) and \(|t_r|\) has work \( O(\log |D| + B) \).

1. When both \( t_l \) and \( t_r \) are simplex trees. From Theorem 6.1, \( \text{join} \) costs \( O(B) \) work.
2. When both \( t_l \) and \( t_r \) are complex trees. From Theorem 6.1, \( \text{join} \) costs \( O(|t_l|/\log |t_l|) = O(1) \) work.
3. When \( t_l \) is a complex tree, but \( t_r \) is a simplex tree. This means that \(|t_l| > B \) and \(|t_r| \leq B \). From Theorem 6.1, \( \text{join} \) costs \( O(B + \log |t_l|/B) \) work. Note that, as stated above, \(|t_r| \geq \alpha|P|\). Considering \(|t_r| \leq B \), we know that \(|P| = O(B)\), which also indicates \(|t_l| = O(\log B)\). Plug this into the work of \( \text{join} O(B + \log |t_l|/B) \), we can get the bound \( O(B) \).

In summary, in all cases the work of \( \text{join} \) or \( \text{join2} \) is asymptotically bounded by the corresponding split function. \( \square \)

Next, we prove the bounds for split work and expose work, respectively.

**Lemma 6.6.** 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 at most one expose, after which the flat node will be fully expanded. In summary the cost is \( O(\min(mB, n)) \). \( \square \)

**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 done 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 cannot 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 \( O(m \log \frac{n}{m} + mB) \). \( \square \)

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 \( O(\log |T_p|/B) \) 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 \( \text{nil} \) 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 \( \text{split} \) and \( \text{join} \). In total the span is \( O(\log m \log n) \). \( \square \)

### 6.1 Set Algorithms with Better Work Bound

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, which we present next. We note that in our implementation, we use the version in Fig. 5, which has good performance in practice. The main result in this section is the theorem below, based on the algorithm shown in Fig. 10 as union_.

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

**Algorithm.** The general idea is to avoid folding and unfolding simplex trees during the union algorithm. In particular, we hope each flat node is folded and unfolded \( O(1) \) times during the entire union algorithm. To ensure this, we implemented a special base case for the set algorithms. We show the code for union in Fig. 10, the other two are similar. The base case will explicitly determine if the current input is a flat node. If any of them is, it will be expanded directly. In the subsequent \( \text{join} \), we will pass an extra parameter to indicate that the tree is already expanded, and thus there is no need to fold or unfold them again in this \( \text{join} \) algorithm. This parameter is the last parameter of the \( \text{join} \) algorithm in Fig. 9. It is set to \( \text{false} \) by default, which makes it exactly the same as the version in Fig. 5. When it is set to \( \text{true} \), the \( \text{join} \) algorithm will never unfold any node. Instead, at the end of the base case, the entire result tree will be fixed using \( \text{refold} \) (Line 41), which traverses the tree and folds any subtree of size \( B \) to \( 2B \) back to blocks.

**Theoretical Cost.** As the span bound is not affected, we will only show the new proof for the work here. First, note that
total work for each of these algorithms can be considered as several parts:

1. all work done by unfold operations in the base cases, including possibly those in split and join (or join2) function calls, denoted as unfold work.
2. all work done by refold operations, denoted as refold work.
3. all work done by split operations except for the unfold work in base cases (already charged in (1), denoted as split work.
4. all work done by join or join2 operations, denoted as join work.

We note that all the rest of the cost can be asymptotically bounded by the above four categories of work.

One observation is that the split work and unfold work are identical among the three set algorithms. This is because the three algorithms behave the same on the way down the recursion when they join back.

We start with some notation. We follow some the notation we use op to denote the algorithm or function call on union, intersection or difference, and use op_base to denote the corresponding base case algorithm or function call (using the unfolded version). In these three algorithms, the first tree (T1) is split by the keys in the second tree (T2). We call T1 the deconstructed tree and T2 the pivot tree, denoted as Td and Tp respectively. The returned tree of the algorithms is denoted as Tr. We use m = min(|Tp|, |Td|) and n = max(|Tp|, |Td|). We denote the subtree rooted at v ∈ Tp as Tp(v), and the tree of keys from Td such that v is operated with as Td(v) (i.e., op(v, Td(v)) or op_base(v, Td(v))) is called at some point in the algorithm. This essentially means that v’s subtree in Tp is processed with the tree Td(v) in a recursive call. Note that Td(v) may not be a subtree in Tp, but is a tree of keys as a subset of Td.

**Lemma 6.8.** The refold work can be asymptotically bounded by the unfold work.

**Proof:** We note that during the process of tracking down the tree, we will refold the subtree if and only at least a subset of it was previously unfolded at some point in this algorithm. Since refold costs O(B) work, it can be asymptotically bounded by the corresponding unfold function invoked previously.

**Theorem 6.9.** For each function call to op on trees Tp(v) and Td(v), the work done by join (or join2) is asymptotically bounded by the work done by split.

**Proof:** In the following, we use P and D to denote Tp(v) and Td(v), respectively, for simplicity. Assume the return value is R. First of all, the work of split is Θ(log |D| + B).

For intersection or difference, the work of join (or join2) is O(log |R| + B). Notice that difference returns the keys in D\P. Thus for both intersection and difference we have |R| ≤ |D|. Therefore |R| ≤ |D|, which means the work done by join or join2 is no more than the work done by split.

For union, first of all, we always call join instead of join2. If |P| ≤ |D|, then |R| ≤ 2|D|. join costs work O(log |R| + B) = O(log |P| + B), which is no more than Θ(log |D| + B).

Consider |P| > |D|. The subtrees lc(P) and rc(P), which are used in the recursive calls, have size at least a|P| and at most (1−a)|P|. After combining with a subset of elements in D (which has size smaller than |P|), the return value of each recursive call should have size at least a|P| and (2−a)|P|. Denote these two trees from recursive calls as t1 and t2 respectively. Note that α is a constant, so the difference of size between t1 and t2 is also a constant. WLOG assume |t1| ≥ |t2|. In the following, we discuss different cases of whether t1 and t2 are complex or simplex trees. We will show that, in all cases, joining |t1| and |t2| has work O(log |D| + B).

1. When both t1 and t2 are simplex trees. From Theorem 6.1, join costs O(B) work.
2. When both t1 and t2 are complex trees. From Theorem 6.1, join costs O(log |t2|/t1) = O(1) work.
3. When t1 is a complex tree, but t2 is a simplex tree. This means that |t1| > B and |t2| ≤ B. From Theorem 6.1, join costs O(B + log |t2|/B) work. Note that, as stated above, |t2| ≥ a|P|. Considering |t2| ≤ B, we know that |P| = O(B), which also indicates |t2| = O(B). Plug this into the work of join O(B + log |t2|/B), we can get the bound O(B).

In summary, in all cases, the work of join or join2 is asymptotically bounded by the corresponding split function.

**Lemma 6.10.** The unfold work is O(min(mB, n)).

**Proof:** First, note that in our unfolded version of base cases, any block needs to be unfold at most once. Each time the algorithm hits a flat nodes, it unfold the entire subtree in O(B) time.

If |Tp| = m, we note that there are at most m/B flat nodes in Tp that needs to be unfolded, so the total work to unfold Tp is O(m). Each of the O(m) entries in Tp will cause an unfold on at most one block in Td. Therefore, the total work to unfold Td is O(mB). On the other hand, note that each
block in \(T_d\) can be unfolded at most once, which also means that the work of unfolding \(T_d\) is \(O(n)\). In summary, the work of all unfold functions is \(O(\min(mB, n))\).

If \(|T_d| = m\) and \(|T_p| = n\), there will be at most \(O(m)\) nodes in \(T_p\) used to process a subset tree in \(T_d\). Since \(|T_d| = m\), the total work of unfolding \(T_d\) is at most \(O(m)\). Based on the same argument as above, the total work of unfolding \(T_p\) is \(O(mB)\) because at most \(O(m)\) unfold functions are invoked, and is also \(O(n)\) because there are at most \(O(n)\) entries in \(T_p\). The total work is also \(O(\min(mB, n))\).

**Lemma 6.11** (Split work on expanded trees). The total split work done on two expanded weight-balanced trees of sizes \(n\) and \(m \leq n\) is \(O(m \log \frac{n}{m})\).

This directly follows [11].

**Lemma 6.12.** The total split work is \(O(m \log \frac{n}{m} + \min(mB, n))\).

**Proof.** The total work for split functions 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.

We first note that in the base cases, the split(\(\cdot\), \(\cdot\), \text{true}) function must be working on an expanded tree. As a result, the total work to traverse and split all non-flat nodes can be asymptotically bounded by the split work considering if both \(T_p\) and \(T_d\) are fully expanded. This cost can be computed by Lemma 6.11, which is \(O(m \log \frac{n}{m})\).

We then consider all work done by the split functions on flat nodes in the non-base cases. We will show it is \(O(\min(mB, n))\).

Note that in each split, this happens at most once, costing \(O(B)\) work. If \(|T_p| = m\), there can be \(O(m/B)\) such split function calls in the non-base cases, and thus the total non-base case split work on flat nodes is \(O(m)\).

If \(|T_p| = n\), we discuss in two cases. If \(n/B \leq m\), there are \(O(n/B)\) regular nodes in \(T_p\), and thus there can be at most \(O(n/B)\) split calls. Therefore the total work in this case is \(O(n)\), which is also \(O(\min(mB, n))\). If \(n/B > m\), there are at most \(O(m)\) such split function calls, since there are only \(m\) nodes in \(T_d\). In this case, the total work of this part is \(O(mB)\), which is also \(O(\min(mB, n))\).

**6.2 Non-parallelizable Encoding Schemes**

As mentioned at the beginning of this section, if the encoding scheme is not parallelizable, when we deal with a flat node, we have to deal with it sequentially and this can affect the span bound of our algorithms. Again we will use the set algorithms as examples.

**Theorem 6.13.** Consider the union algorithm (and similar intersection and difference algorithms) in Fig. 5 on two PaC-trees of sizes \(m\) and \(n \geq m\) using encoding scheme \(C\). When \(C\) takes \(O(B)\) work and span to compress and decompress a block of size \(B\), the span for these set algorithms is \(O(B + \log n'(B + \log m'/B))\), where \(n'\) is the pivot tree size, and \(m'\) is the decomposed tree size.

**Proof.** The algorithms need \(O(\log |T_p|/B)\) rounds to reach a flat node, where the flat node will be expanded, taking \(O(B)\) span. Note that this \(O(B)\) additional span is taken only once for each flat node, and they are all at the leaf level of \(T_p\). As a result, they do not add up and can be charged only once in the span. Then the algorithm keeps recursing until a nil node is reached, which is \(O(\log B)\) rounds. The total number of rounds of recursive calls is still \(O(\log n')\). In each of the recursive calls, we need \(O(B + \log m'/B)\) span to deal with split and join. In total the span is \(O(B + \log n'(B + \log m'/B))\).
split and merge work for the updates. If \( c = 0 \), we can have a tree with all leaf nodes containing \( B \) entries. If we delete any entry, the associated leaf nodes will contain \( B - 1 \) entries, need to merge with a neighbor leaf and end up with having \( 2B - 1 \) nodes. Then if we insert two entries in this leaf, we need to split again with \( O(B) \) work. To avoid this, we set a padded region of size \( cb \) on both sides of the range—once resized, the new leaf node contains \((1 + c)B \) to \((2 + 2c)B \) entries. As such, we need to remove or insert another \( cb \) records to trigger the next resizing, so the amortized work is \( O(B/cB) = O(1) \) per update.

**Theorem 7.1.** A batch of \( m \) insertions or deletions can be processed using \( O(m \log(n/B)) \) amortized work if the batch is unsorted, or \( O(m \log(n/Bm)) \) amortized work if the batch is sorted. The span is \( O(\log(n/B) \log m + \log B) \) for both cases.

**Proof:** We first assume the update batch is unsorted. We use the tree root to partition the batch, which takes \( O(m) \) work and \( O(\log m) \) span. Then we can recursively and in parallel update the left part of the batch and the left subtree, the right part of the batch and the right subtree. After they both finish, we join the two trees with the tree root. The base case is when the corresponding batch for the subtree is empty, or the tree goes to a leaf node. We terminate for the first case. For the second case, we update the leaf with work proportional to the update array size (concatenation for insertions, marking tombs for deletions). The update may trigger a clean-up for the leaf array (split if the array size is larger than \((2 + 3c)B \), merge if the size is smaller than \( B \), or pack if the tombs occupy over a constant fraction of the array). Once a clean-up is triggered, the work is \( O(B) \) and the span is \( O(\log B) \), and as explained, the work is constant amortized to each previous update to this leaf node.

Similar to the previous analysis, we can split the work into the split work, the join work, and the base case work. The base case work is constant per update. The split work is \( O(m \log(n/B)) \)—each tree level will partition the update array and the total cost is \( O(m) \) per tree level. Since the tree has size \( n \) and \( O(\log(n/B)) \) levels, the total split work is \( O(m \log(n/B)) \). The join work is logarithmic in the subtree size, and at most \( m \) leaf nodes are modified, so the overall work for join is \( O(m \log(n/Bm)) \), bounded by split work. Putting all pieces together, the work is \( O(m \log(n/B)) \) amortized. The span is \( O(\log(n/B) \log m + \log B) = O(\log m) \) for split and join for \( O(\log(n/B)) \) levels, and \( O(\log B) \) for the base case.

If the update batch is sorted, the split becomes a binary search with cost logarithmic in the current update array size. In this case, the split work is bounded by the join work, so the total work becomes \( O(m \log(n/Bm)) \) amortized. The span remains unchanged. \( \square \)

This new version of the PaC-tree may of interest when updates are more frequent than queries, or the queries are more costly. For instance, if the query is “reporting the top-\( k \) elements”, where \( k \) can be large, the work for this query is \( \Omega(k) \). In this case, we can use this alternative version of the PaC-tree to reduce the update cost. Assume \( k \) is static throughout the algorithm, we can set \( B = k \) for the CPAM tree. Based on Theorem 7.1, each update costs \( O(\log(n/k)) \) work. For a query, we only need to look into the first leaf array which contains at least \( k \) entries and at most \((2 + 3c)k \) entries. Then we can use the classic algorithm [34] to find the \( k \)-th element from this array, and pack those no larger than it. Hence, this query takes \( O(k) \) work and \( O(\log k \log log k) \) span, and the span can be optimized to \( O(\log k) \) using the deterministic sampling technique in [13]. This is much more efficient than directly running the classic algorithm [34] with \( O(n) \) work, and better than keeping the entire search tree (e.g., a P-tree) which has \( O(\log n) \) work per update.

**8 Implementation**

In this section, we describe CPAM, our implementation of PaC-trees. CPAM is built in C++, based on the PAM framework [52]. 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 \( filter,\ union \).

**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 in-place and functional updates separately, this is not attractive. Instead, we designed a simple approach to handle both cases using the same code.
Our approach is to store an additional bit indicating whether the supplied node is visible solely to the current function, or whether the node has some external observer, and should therefore be copied. We refer to these special pointers to tree nodes with an additional bit for visibility as *extra pointers*. When an extra pointer is copied, e.g., an algorithm like union wishes to use it as part of the resulting tree, we copy this node if the visibility bit is set or if the node has a reference count more than 1, and otherwise we simply return the same node. Similarly, when we expose an extra pointer pointing to a regular node \( v \), we set the visibility bits on the children either if the \( v \)’s visibility bit is set, or if \( v \) has a reference count greater than 1. If \( v \) was visible only to the caller, as an optimization we return it as an additional result, allowing the caller to potentially reuse this node. Our approach lets us write simple algorithms which modify the tree in-place when possible, and begin copying once it reaches subtree that is visible to other observers.

**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 [49].

9 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 Aspen [52]. Our graph processing application is based on Aspen [25]. 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 [48], including breadth-first search, maximal independent set, and single-source betweenness centrality. Our implementations are based on the ones in Aspen and GBBS [26, 27]. We design parallel batch-updates for our representation, which are applicable in graph-streaming and batch-dynamic graph algorithms.

10 Experiments

**Experimental Setup.** We run experiments on a 72-core Dell PowerEdge R930 (with two-way hyper-threading) with 4 x 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.

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

- PaC-trees are competitive with PAM for microbenchmarks (Section 10.1) and applications including inverted indices (Section 10.3) and 2D range queries and 1D interval queries (Section 10.4) while using 2.1x–7.8x less space.
- Varying the block size \( B \) for an PaC-tree trades off off performance for space efficiency (Section 10.1). 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 10.5).

10.1 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 \( \text{union} \) 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 \( \text{union} \) only depends on the smaller size. In this case, since the cost of \( \text{union} \) 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. 12 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 \( \text{find} \) and \( \text{range} \), we see a steady increase in the running time for \( B > 16 \) and see a similar trend for \( \text{union} \)-Imbal, which combines the union of trees with \( 10^5 \) and \( 10^6 \) elements. This slowdown with increasing \( B \) is due to the extra \( O(mB) \) term in the work of \( \text{union} \). For the smallest block size \( (B = 1) \), our running time matches that of P-trees on this operation.

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 unaugmented maps using PaC-trees as a function of the block size \( B \) (Fig. 13). 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. 13). 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.

10.2 Comparison with Collections in Spark
We compared CPAM with the shared-memory parallel implementation of Apache Spark on a simple benchmark drawn from the Apache Spark tutorial. The benchmark first loads the same Wikipedia corpus that we use for our inverted index application (1,943,575,146 words in 8,125,326 documents). The first example then tokenizes the dataset into words, and computes the longest word length. The second example computes the most frequently occurring word by using the \( \text{reduceByKey} \) primitive in Spark to group common words and compute the mostly frequently occurring word using a reduce. We use in-memory caching for the intermediate mapped dataset and report the fastest (cached) time.

For the first example, Spark takes 46.9 seconds for the first (uncached) run, and the subsequent average (cached) time is 21.5 seconds. Our CPAM implementation, where the dominant cost is the memory-bound parsing step, requires 6.57 seconds on average (3.2x faster than the cached time). On the second example, Spark takes 96.3 seconds for the first run, and 72.5 seconds for the average cached time. For the second example, the dominant cost for CPAM is the parallel sort (we use a parallel sample-sort). The end-to-end time is 14.6 seconds, which is 4.9x faster than the Spark cached time.

We also tried evaluating the same set of sequence benchmarks shown in Fig. 2 using Spark, but observed significantly worse running times for all of the sequence primitives (up to 2 orders of magnitude worse performance; e.g., reduce on a 100M element sequence with 8-byte elements takes 2.07 seconds, whereas CPAM takes 0.00865 seconds). The slowdown could be due to fixed parallelization overheads in Spark, although their word counting example which we studied above performs reasonably well despite working over a significantly larger dataset (nearly 2 billion words).

10.3 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 [52]. 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).

### 10.4 Interval and Two-Dimensional Range Trees

We benchmark our interval and two-dimensional range trees as in PAM [50]. We build our interval tree on 10^8 intervals, and for queries run stabbing queries over 10^5 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

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. $T_{144}$ is the sequential running time. $T_{44}$ 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>Library</th>
<th>Space</th>
<th>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>8.29</td>
<td>Build</td>
<td>10^8</td>
<td>—</td>
<td>746</td>
<td>9.73</td>
<td>76.6</td>
</tr>
<tr>
<td>PaC-tree (D)</td>
<td>4.07</td>
<td>Build</td>
<td>10^8</td>
<td>—</td>
<td>754</td>
<td>9.81</td>
<td>76.8</td>
</tr>
<tr>
<td>P-tree (PAM)</td>
<td>3.19</td>
<td>Build</td>
<td>10^8</td>
<td>—</td>
<td>575</td>
<td>8.86</td>
<td>64.9</td>
</tr>
<tr>
<td>P-tree (PAM)</td>
<td>3.54</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>PaC-tree</td>
<td>40.3</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>PaC-tree</td>
<td>40.3</td>
<td>Q-All</td>
<td>10^8</td>
<td>10^8</td>
<td>54.2</td>
<td>0.629</td>
<td>86.1</td>
</tr>
<tr>
<td>PaC-tree</td>
<td>89.6</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>PaC-tree</td>
<td>89.6</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>

Figure 11. Relative space usage of different graph representations. GBBS (Diff) is our static baseline compressed graph representation. PaC-tree uses PaC-trees for vertex and edge trees, and PaC-tree (Diff) difference encodes both trees. Aspen uses P-trees for the vertex tree and C-trees with difference encoding for edge trees. P-tree (PAM) uses P-trees for the vertex and edge trees. The values on top of each bar are the memory usage in GiB.

![Figure 11](image-url)
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 $10^5$ 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 [45] and was shown to outperform it [50].

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). 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.

### 10.5 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.** DBLP is co-authorship network based on research papers in computer science. YouTube is a social network graph based on YouTube. USA-Road (RO) is an undirected road network from the DIMACS challenge [24]. LiveJournal (LJ) is a directed graph of the LiveJournal social network [17]. com-Orkut (CO) is an undirected graph of the Orkut social network. Twitter (TW) is a directed graph of the Twitter network [40]. Friendster (FR) is an undirected graph describing friendships from a gaming network. The DBLP, YouTube, and Friendster graphs are obtained from the SNAP dataset [41]. We note that some of our inputs (like the LiveJournal graph) are originally directed, and we symmetricize them before applying our algorithms to maintain consistency with prior work on Aspen [25] and GBBS [26] that symmetrize graphs in their evaluations. Table 4 shows information about our graph inputs, 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 static arrays using difference encoding, which serves as our baseline of graph representation. Fig. 11 shows the relative size of each graph format. We see that the smallest format in all cases is PaC-tree (Diff), which applies PaC-trees with difference encoding for both vertex and edge trees. Using this format yields a space improvement of between 4–9.7x over just using P-trees. For the graphs with high average-degree, most of the savings come from using PaC-trees for the edge trees. Adding difference encoding to both trees yields between 1.05–1.32x space improvement. PaC-trees are also 1.3–2.6x more space-efficient than Aspen. Note that C-trees in Aspen are also difference encoded, so the main difference between the two representations is that PaC-tree (Diff) also uses PaC-trees to chunk the vertex tree, and that PaC-trees employ a deterministic strategy for chunking. PaC-trees with difference encoding achieves consistently lower space compared with Aspen, ranging between 1.3x for Friendster, our largest graph, to a maximum space improvement for 2.62x on USA-Road, our sparsest graph. The space savings come from chunking the vertex trees, which is not possible in Aspen, since the C-tree implementation is specialized for edge trees.

**Graph Algorithm Performance.** We study the performance of three fundamental graph kernels: breadth-first search (BFS), single-source betweenness centrality (BC), and maximal independent set (MIS). Our implementations are based on those in Aspen. We study performance using our most space-efficient version (PaC-tree (Diff)). Following Aspen, our implementation also supports the flat snapshot object, which is an array storing all vertices in the current graph. The

Table 4. Statistics about tested graphs and memory usage of PaC-tree and Aspen in GiB.

<table>
<thead>
<tr>
<th>Graph</th>
<th>Vertices</th>
<th>Edges</th>
<th>Ours</th>
<th>Aspen</th>
<th>Aspen (Diff)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP (DB)</td>
<td>425,957</td>
<td>2,099,732</td>
<td>0.0130</td>
<td>0.03409</td>
<td>2.62x</td>
</tr>
<tr>
<td>YouTube (YT)</td>
<td>1,138,499</td>
<td>5,980,886</td>
<td>0.0412</td>
<td>0.0934</td>
<td>2.26x</td>
</tr>
<tr>
<td>USA-Road (RU)</td>
<td>23,947,348</td>
<td>57,708,624</td>
<td>0.683</td>
<td>1.843</td>
<td>2.69x</td>
</tr>
<tr>
<td>LiveJournal (LJ)</td>
<td>4,847,571</td>
<td>85,702,474</td>
<td>0.346</td>
<td>0.527</td>
<td>1.52x</td>
</tr>
<tr>
<td>com-Orkut (CO)</td>
<td>3,072,627</td>
<td>234,370,166</td>
<td>0.727</td>
<td>0.893</td>
<td>1.22x</td>
</tr>
<tr>
<td>Twitter (TW)</td>
<td>41,652,231</td>
<td>2,405,026,092</td>
<td>7.59</td>
<td>9.42</td>
<td>1.23x</td>
</tr>
<tr>
<td>Friendster (FS)</td>
<td>65,608,366</td>
<td>3,612,134,270</td>
<td>14.6</td>
<td>19.1</td>
<td>1.30x</td>
</tr>
</tbody>
</table>

Table 5. Parallel running times (in milliseconds) for Aspen and our implementation.

<table>
<thead>
<tr>
<th>Graph</th>
<th>BFS Time</th>
<th>No-FS</th>
<th>FS Time</th>
<th>Ours</th>
<th>Aspen</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiveJournal</td>
<td>19.8</td>
<td>17.5</td>
<td>1.13x</td>
<td>1.38</td>
<td>1.24x</td>
</tr>
<tr>
<td>com-Orkut</td>
<td>14.5</td>
<td>12.4</td>
<td>1.16x</td>
<td>1.12</td>
<td>1.23x</td>
</tr>
<tr>
<td>Twitter</td>
<td>110</td>
<td>97</td>
<td>1.11x</td>
<td>12.5</td>
<td>1.11x</td>
</tr>
<tr>
<td>Friendster</td>
<td>82.1</td>
<td>72.3</td>
<td>1.13x</td>
<td>1.38</td>
<td>1.03x</td>
</tr>
<tr>
<td>com-Orkut</td>
<td>96.9</td>
<td>69.2</td>
<td>1.40x</td>
<td>1.12</td>
<td>1.01x</td>
</tr>
<tr>
<td>Twitter</td>
<td>2735</td>
<td>1030</td>
<td>2.65x</td>
<td>12.5</td>
<td>1.11x</td>
</tr>
</tbody>
</table>
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.

Graph Batch-Update Throughput In this section we study the performance of our graph representation using PaC-trees when performing updates in varying size batches. We focus on algorithms for inserting and deleting batches of edges, since inserting vertices (along with incident edges) can easily be done using the edge insert primitive, and vertex deletions simply use multi_delete on PaC-trees. Our experiment follows the methodology used in Aspen. To generate updates, we sample directed edges from an rMAT generator [21] with \(a = 0.5, b = c = 0.1, d = 0.3\). We use the same generator for our concurrent update and query experiment in Section 10.5. For a batch of size \(K\), we generate \(K\) directed edge updates from the stream (note that there can be duplicates) and repeatedly insert the edges and delete the edges from the batch, reporting the median of three trials. We note that the times we report include the time to sort and remove duplicates from the batch.

Fig. 15 shows the throughput of batch edge insertions (insertions per second) as a function of the batch size. We note that the throughput for deletions are close to that of insertions (within 10% across all graphs). To remove clutter, we show results on the five largest graphs in our datasets. We observe that the throughput of our graph representation improves with increasing batch size; for the largest batch size, the algorithm achieves a maximum throughput of between 719M edge insertions per second for the com-Orkut graph, and a minimum of 527M edge insertions per second for the Twitter graph. We compared these results with those of Aspen on the same machine and find that we obtain 1.62x higher throughput across the three graphs both systems consider in this experiment, and an average throughput increase of 1.65x across these graphs.

Concurrent Updates and Queries. Our last experiment concurrent updates and queries on graphs. The experiment

**Figure 12.** Primitive running times for PaC-trees vs. block size \(B\). We use \(10^8\) key-value pairs (8 bytes each). Union, Intersection and Difference all work on two trees with \(10^8\) elements. Union-lmbal takes the union of trees with \(10^8\) and \(10^9\) elements.

**Figure 13.** Size of PaC-trees (with or without DE) as a function of block size \(B\). We use \(10^8\) key-value pairs (8 bytes each). For augmented maps (-Aug), augmented values are 8 bytes each. The grey line shows the number of bytes to store the \(10^8\) elements in an array and the purple line shows the bytes used to store the difference encoded keys in a single array using byte encoding.

**Figure 14.** Performance of concurrent updates and queries. The time series plot illustrates running times when running BFS queries with batch-insertions of edges concurrently (Concurrent), and when queries and updates are run individually (Solo) on the LiveJournal graph.
performs \( n \) undirected edge insertions drawn from the rMAT generator described above. 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).

Fig. 14 shows the result of the experiment. We find that the concurrent queries are 1.85x slower on average than the queries in isolation, and that the concurrent updates are 1.07x slower on average than updates in isolation. In the concurrent setting, the average latency to make one of the update batches visible is 100 microseconds, and the updates achieve a throughput of 94,000 undirected edge updates per second. We leave further optimizations and a more in depth study of the graph setting for future work with our system.

11 Conclusion
We have presented PaC-tree, 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.

References
http://www.dis.uniroma1.it/challenge9/.


