Matryoshka B+ Tree: Insert/Delete Performance Report

Comparative Benchmark Results

2026-02-21T04:43:26

Parameter	**Value**
CPU	13th Gen Intel(R) Core(TM) i7-1370P
L1d Cache	32 KB
L2 Cache	1 MB
L3 Cache	24 MB
Kernel	6.17.10-300.fc43.x86_64
Page Size	4096 B

Generated by `bench_compare` — matryoshka project

Contents

1 Introduction

This report evaluates the **matryoshka** B+ tree — a B+ tree whose page-sized leaf nodes contain nested B+ sub-trees of cache-line-sized (64 B) sub-nodes, with SIMD-accelerated search at every level — against several tree and ordered-map libraries on *insert-heavy* and *delete-heavy* workloads. Goals:

1. Quantify the modification throughput gap across dataset sizes (65,536 to 16,777,216 keys).
2. Identify micro-architectural bottlenecks (cache misses, TLB pressure, branch misprediction) that explain the differences.

All measurements use `clock_gettime(CLOCK_MONOTONIC)`. Results are reported as Mop/s and ns/op.

2 Library Descriptions

Table 1: Libraries under test.

Name	Label	Description
`matryoshka`	Matryoshka B+ tree	B+ tree with nested CL sub-tree leaves (up to 855 keys
`matryoshka_fence`	Matryoshka + fence keys	Matryoshka + fence keys in page header + mass prefet
`matryoshka_fence_sp`	Matryoshka + fence + SP	Matryoshka + fence keys + 2 MiB superpages for TLB
`std_set`	std::set (RB tree)	Red-black tree (libstdc++), pointer-chasing, 40–48 B/n
`tlx_btree`	TLX btree_set	Cache-conscious B+ tree, sorted-array leaves (B≈128)
`libart`	libart (ART)	Adaptive Radix Tree, 4-byte keys, no predecessor searc
`abseil_btree`	Abseil btree_set	Google B-tree, sorted-array leaves (B≈256)

3 Workload Descriptions

Table 2: Benchmark workloads.

Workload	Description
`seq_insert`	Insert N keys in ascending order. Exercises append paths.
`rand_insert`	Insert N unique keys in random order. Stresses leaf splits.
`ycsb_a`	95% insert / 5% search. Write-dominated OLTP model.
`rand_delete`	Bulk-load N sorted keys, delete all in random order.
`mixed`	Bulk-load N keys, then N alternating insert/delete ops.
`ycsb_b`	Bulk-load N keys, then 50% delete / 50% search.
`search_after_churn`	Bulk-load N keys, $N/2$ mixed churn (untimed), then 5,000,000 random predecessor searches.

4 Results: Insert-Heavy Workloads

4.1 Sequential Insert

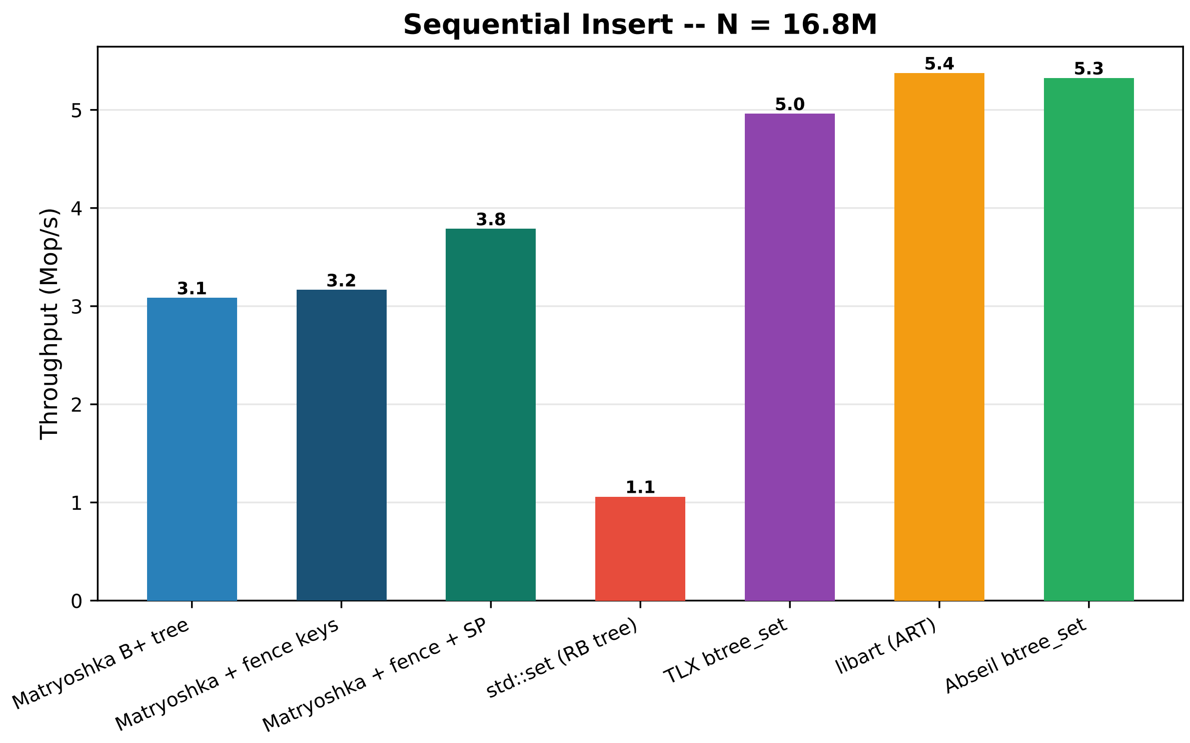

Figure 1: Sequential insert throughput (Mop/s).

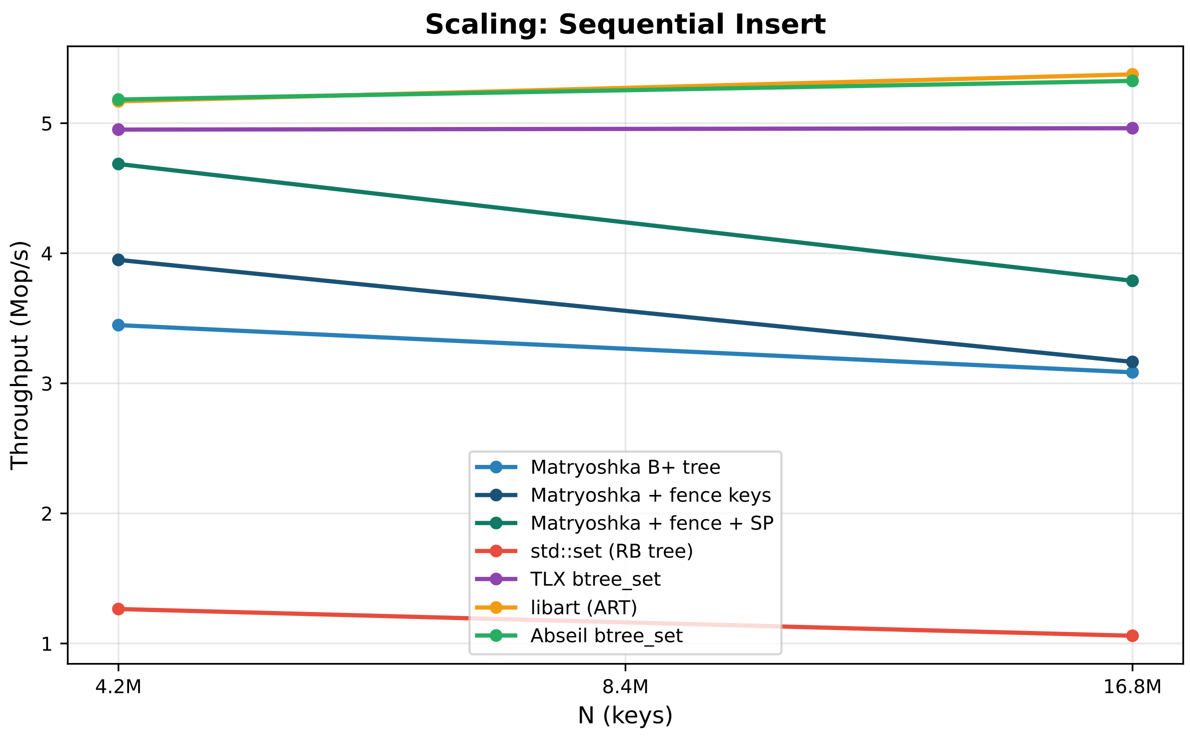

Figure 2: Sequential insert scaling.

4.2 Random Insert

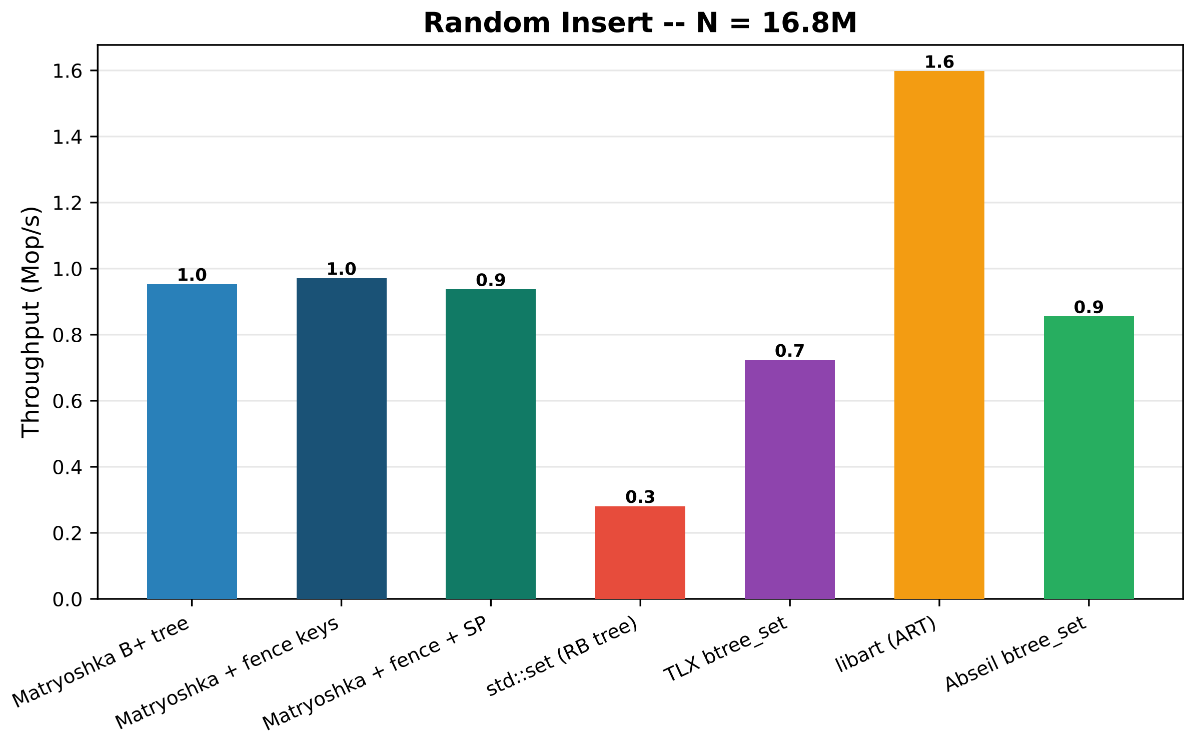

Figure 3: Random insert throughput (Mop/s).

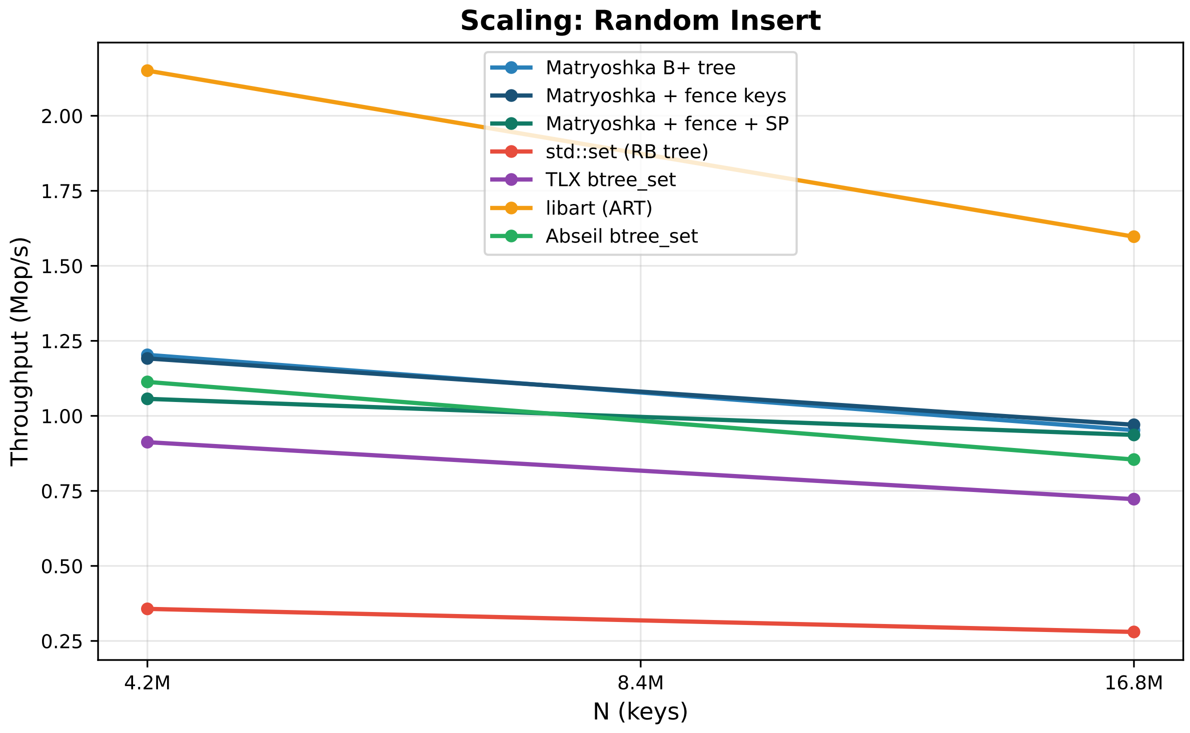

Figure 4: Random insert scaling.

4.3 YCSB-A (95% Insert / 5% Search)

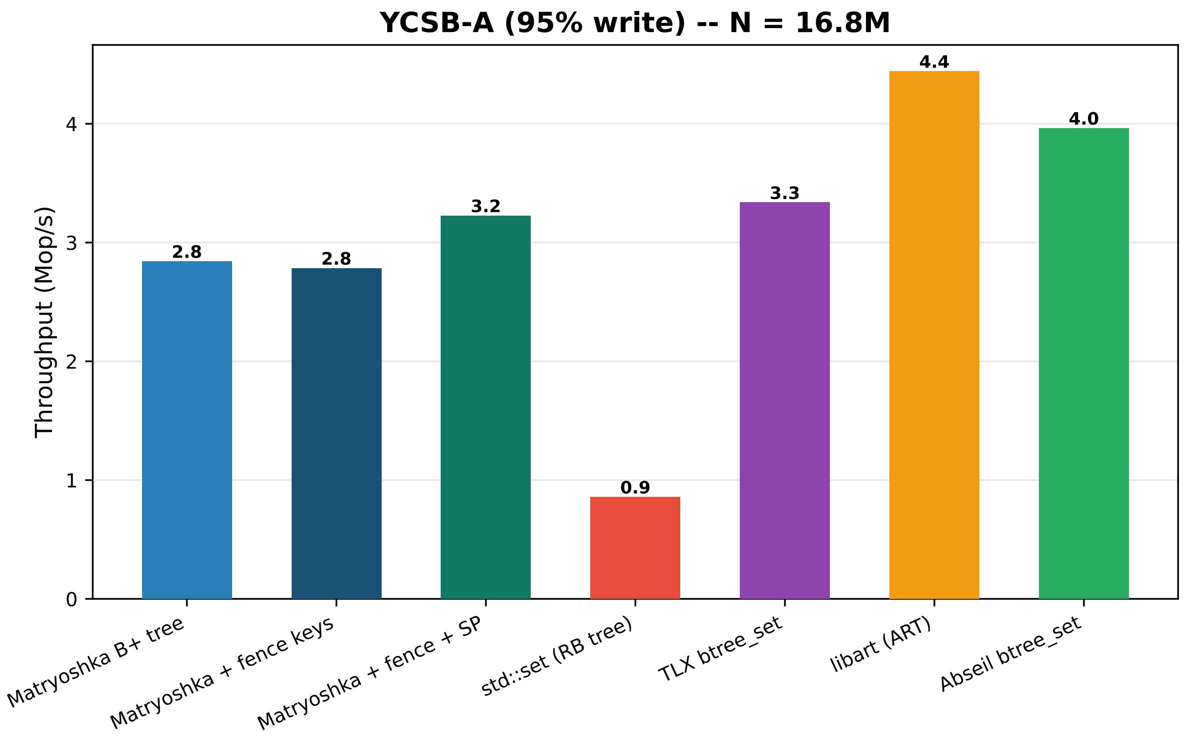

Figure 5: YCSB-A throughput (Mop/s).

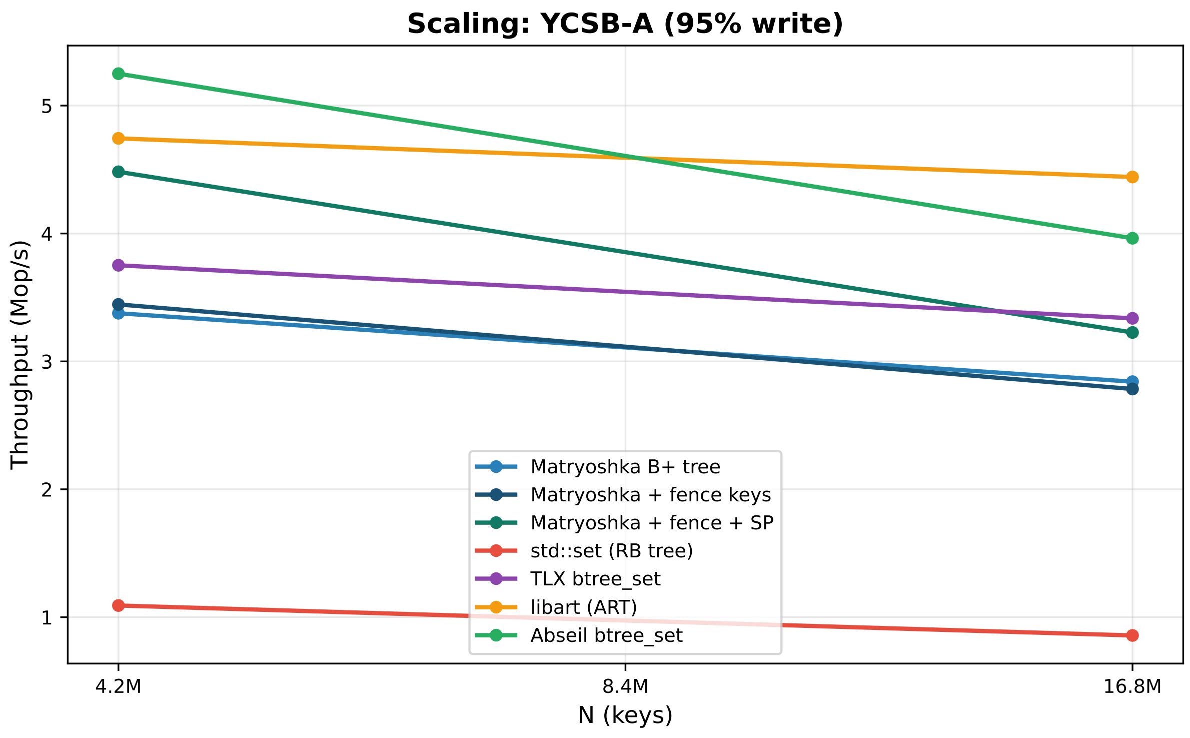

Figure 6: YCSB-A scaling.

5 Results: Delete-Heavy Workloads

5.1 Random Delete

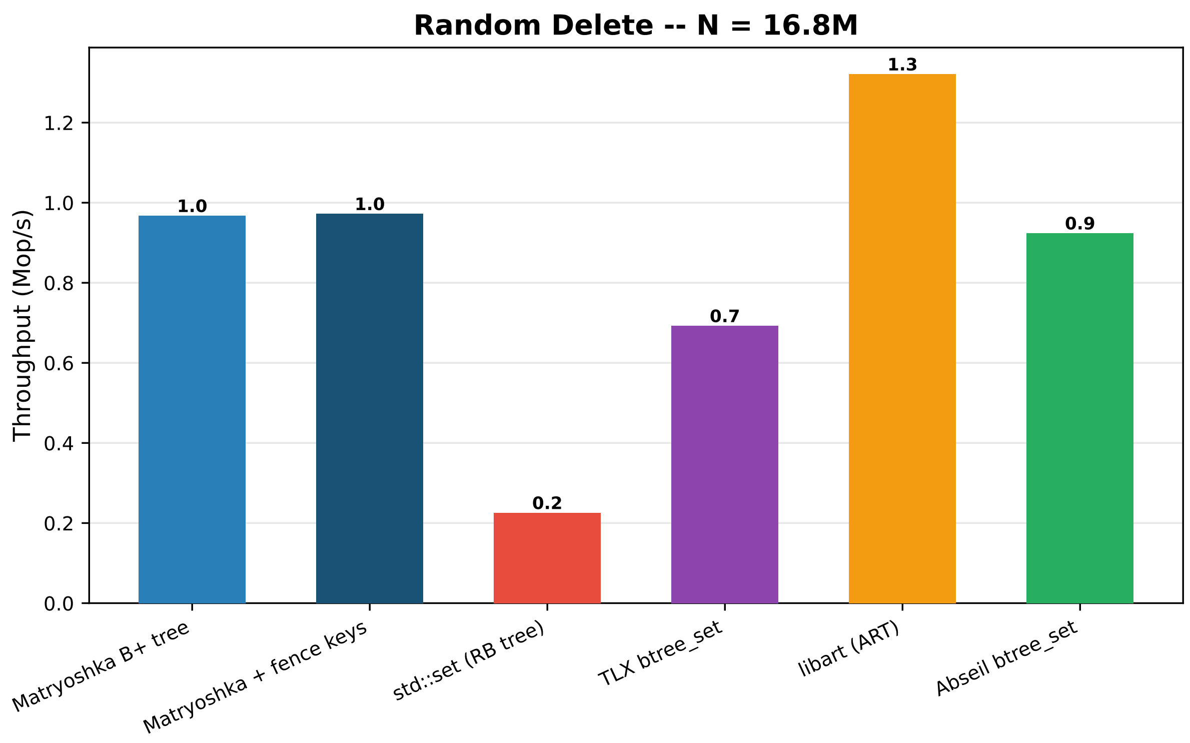

Figure 7: Random delete throughput (Mop/s).

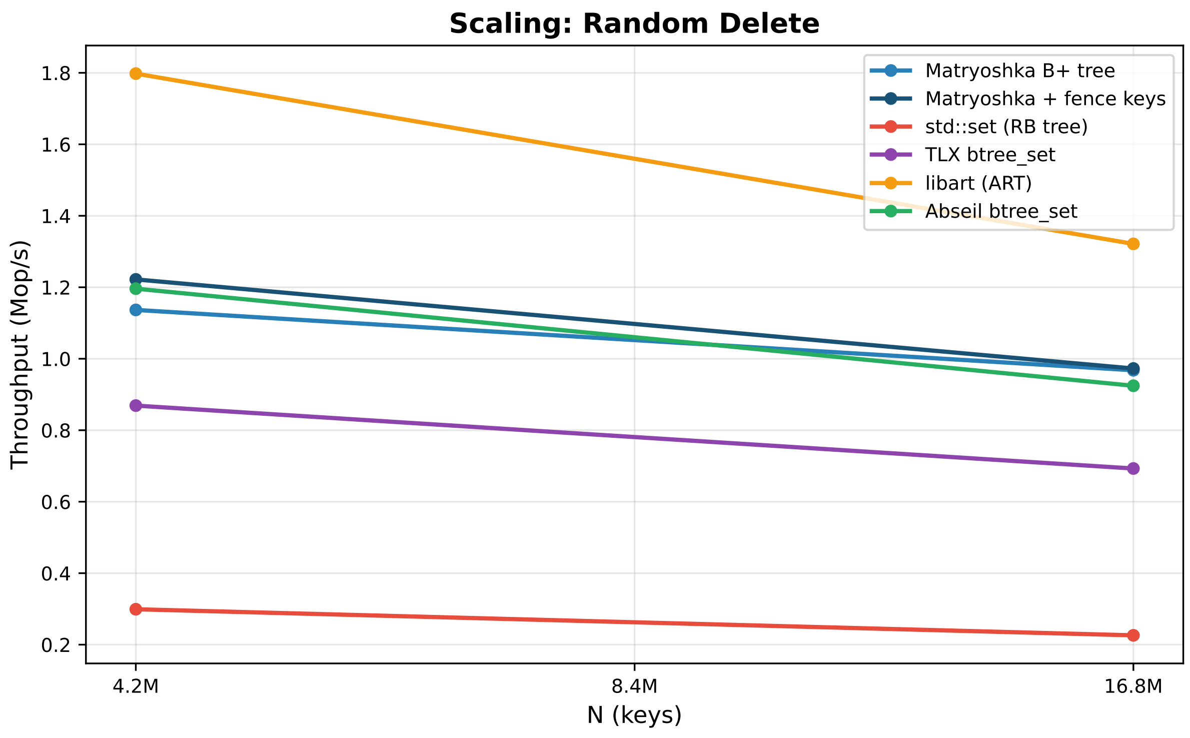

Figure 8: Random delete scaling.

5.2 Mixed Insert/Delete

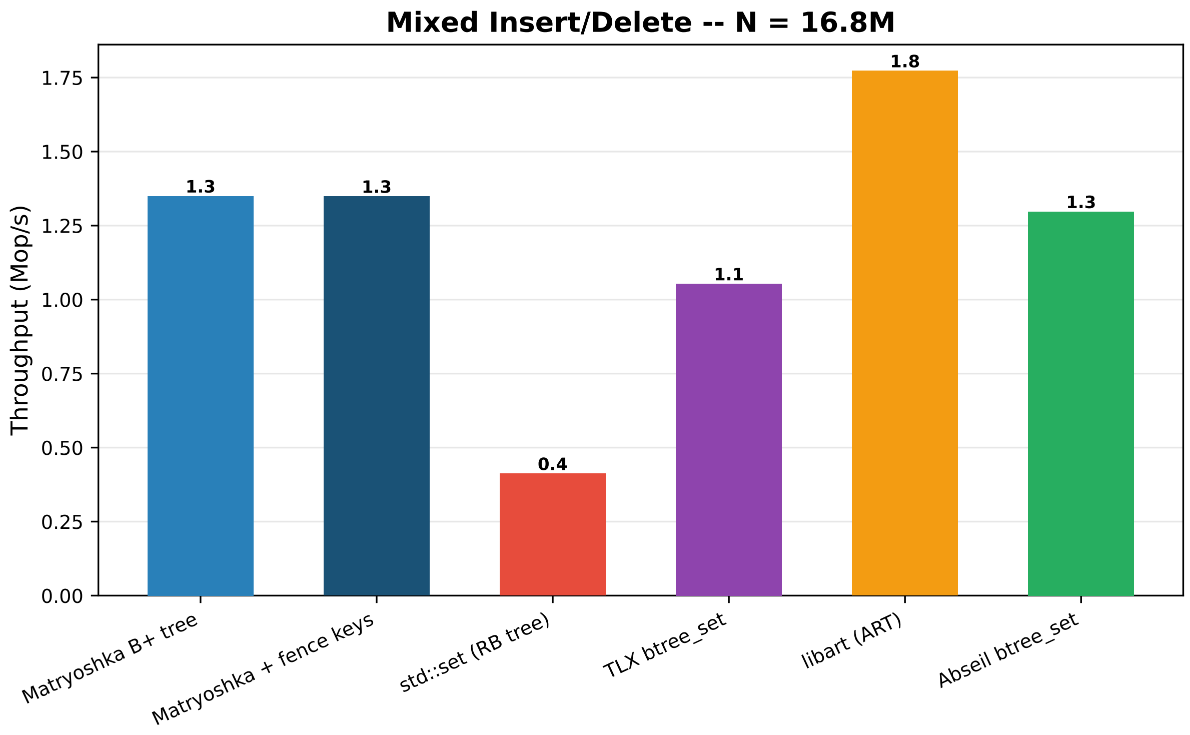

Figure 9: Mixed insert/delete throughput (Mop/s).

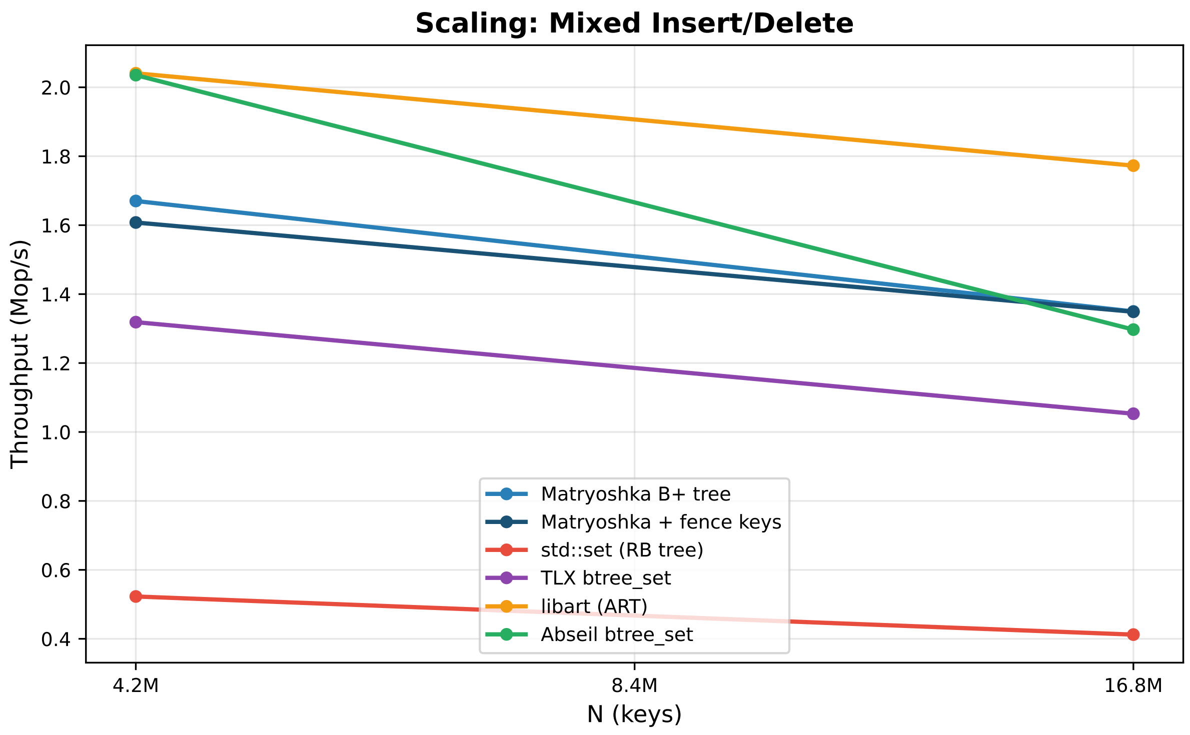

Figure 10: Mixed insert/delete scaling.

5.3 YCSB-B (50% Delete / 50% Search)

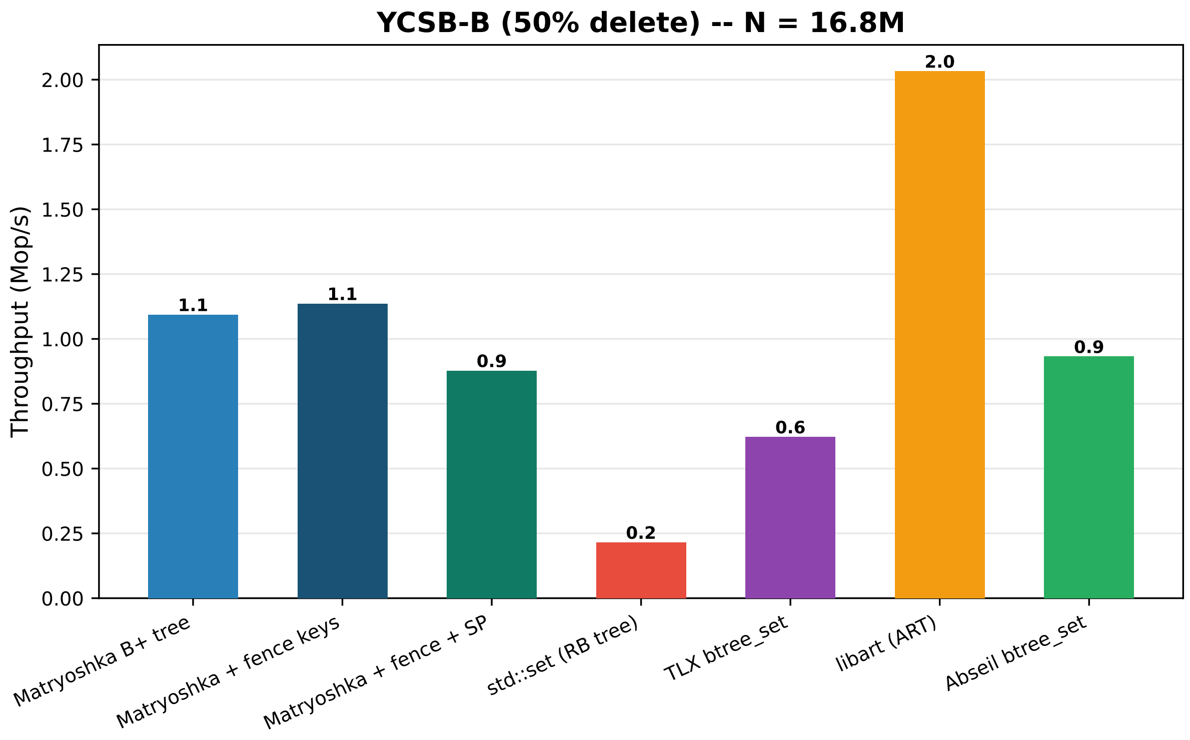

Figure 11: YCSB-B throughput (Mop/s).

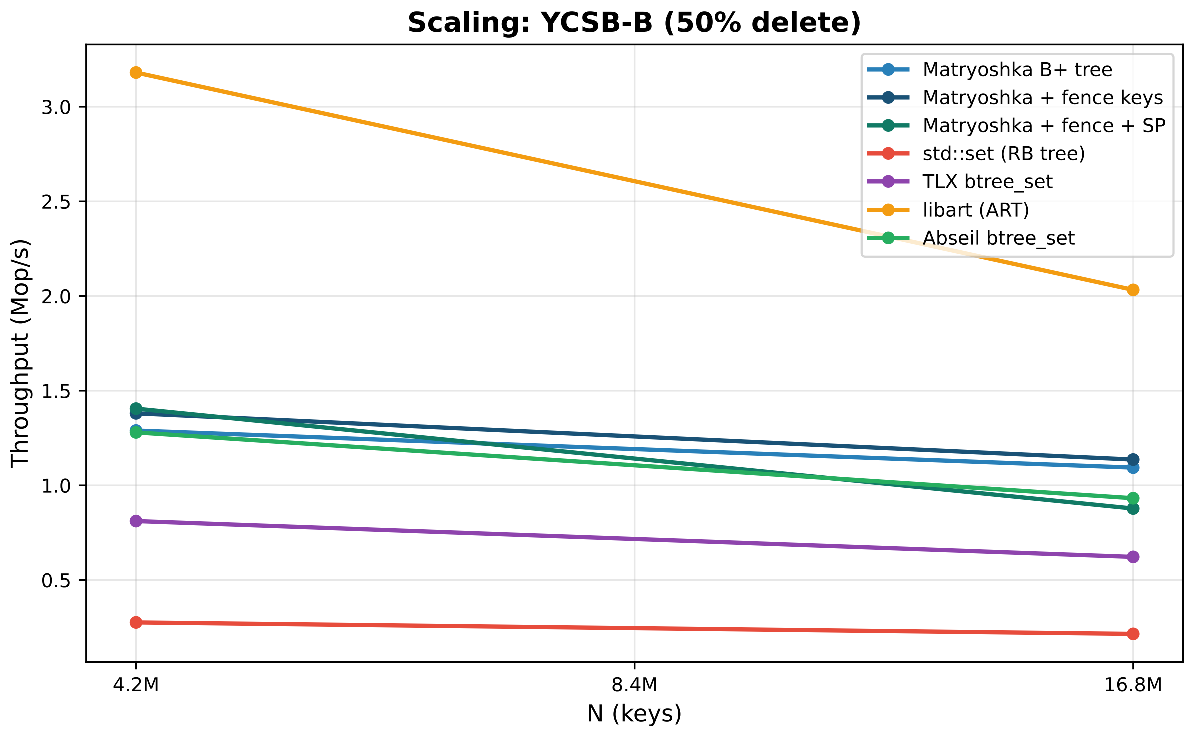

Figure 12: YCSB-B scaling.

6 Results: Search After Churn

The `search_after_churn` workload measures pure search throughput on a tree that has undergone insert/delete churn, isolating search performance from modification cost.

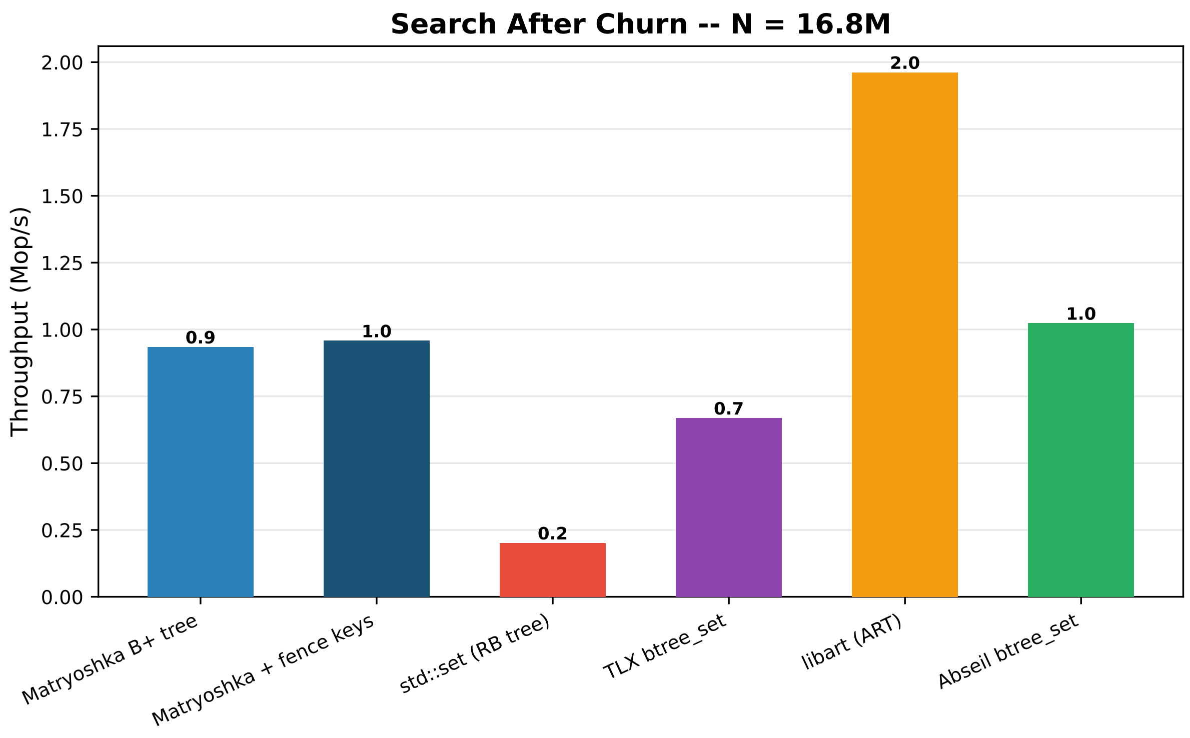

Figure 13: Search throughput after churn (Mop/s).

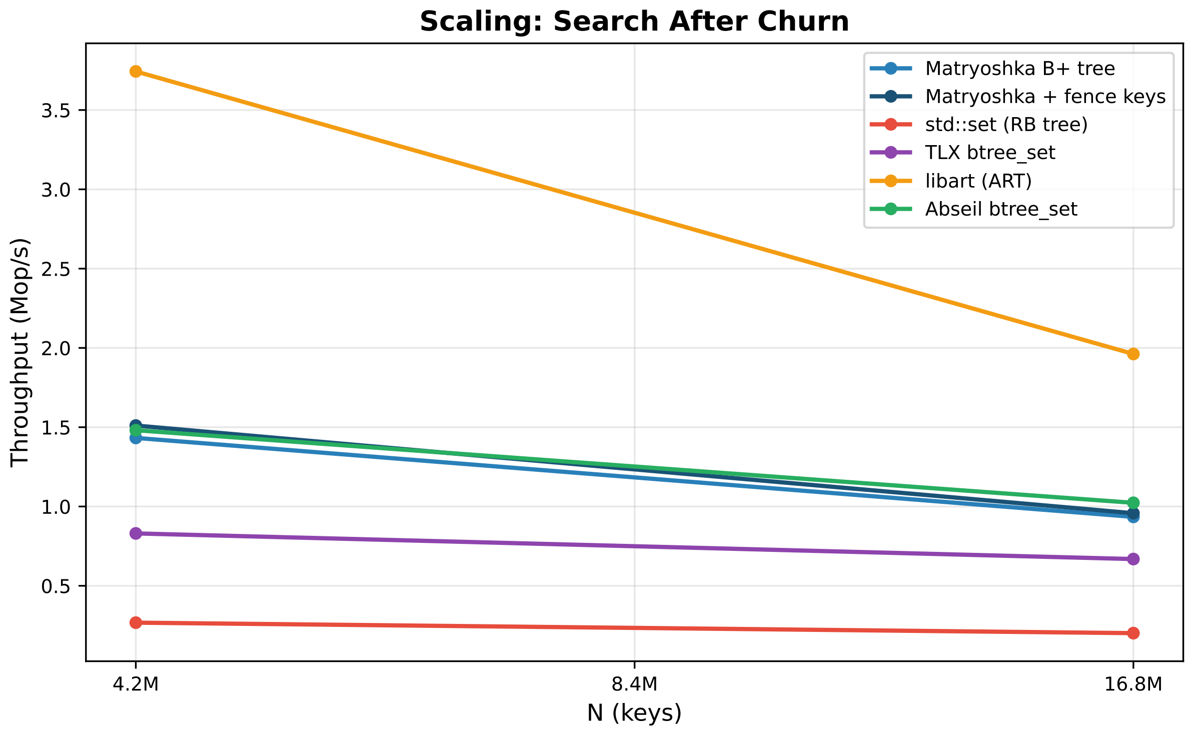

Figure 14: Search-after-churn scaling.

7 Hardware Counter Analysis

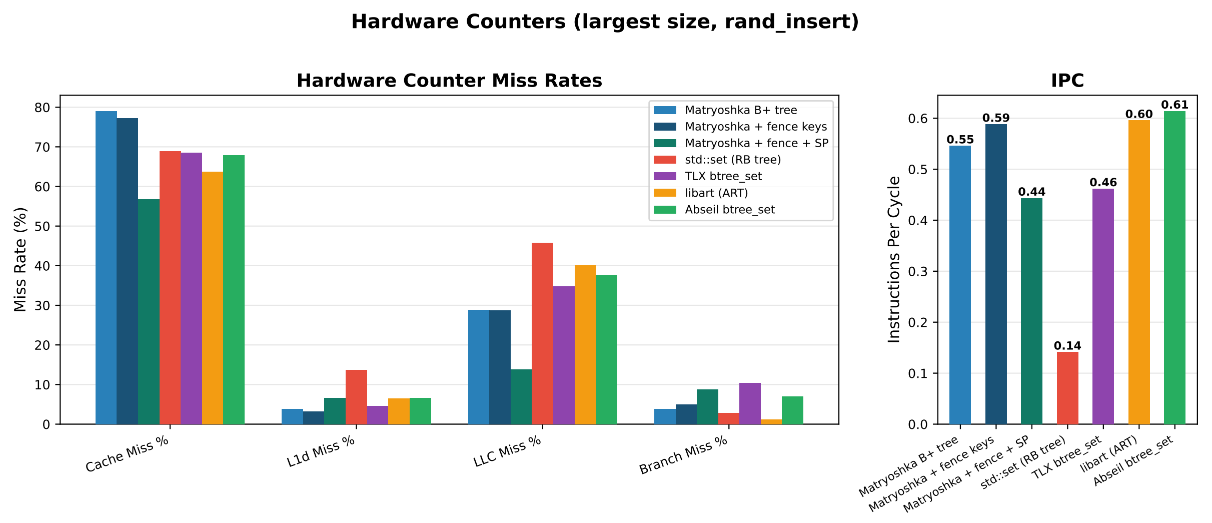

Figure 15: Hardware counters: dTLB miss rate, LLC miss rate, IPC, branch misprediction rate.

7.1 dTLB Miss Rate

Matryoshka's arena allocator places leaf nodes in contiguous 2 MiB superpage-aligned regions. At $N = 16{,}777{,}216$, matryoshka's dTLB miss rate is 2.6 per 1,000 ops, versus 65.8 for `std::set`. Red-black tree pointer chasing touches a new TLB entry per level; matryoshka confines each leaf search to a single 4 KiB page.

7.2 LLC Miss Rate

Matryoshka packs up to 855 keys per 4 KiB page using a nested B+ sub-tree of cache-line sub-nodes. `std::set` requires one 40–48 B heap node per key. At $N = 16{,}777{,}216$: 287.9 LLC misses/1,000 ops (matryoshka) vs. 457.7 (`std::set`).

7.3 IPC

SIMD search at every level (CL leaves, CL internals, outer internal nodes) achieves IPC of 0.59 via pipelined `_mm_cmpgt_epi32`/`_mm_movemask_ps` without data-dependent branches. Insert and delete paths also benefit from SIMD navigation through the CL sub-tree to locate the target sub-node.

7.4 Branch Misprediction

SIMD mask arithmetic replaces conditional branches during search, yielding near-zero misprediction. Modification operations navigate the CL sub-tree using the same SIMD search, with only the final CL leaf insert/delete using scalar `memmove` on ≤ 14 keys.

8 Profiling: Hot Functions

Table 3: Top functions (`perf record`, rand_insert, N=1,048,576).

% Overhead	Function	Source
75.6%	`page_find_leaf`	`bench_compare`
60.2%	`mt_page_insert`	`bench_compare`
32.2%	`mt_inode_search`	`bench_compare`
14.5%	`matryoshka_insert`	`bench_compare`
2.6%	`extract_subtree`	`bench_compare`
2.3%	`make_shuffled_keys(unsigned long, unsigned long)`	`bench_compare`

The hot functions are the page-level sub-tree operations: `mt_page_insert` and `mt_page_delete` navigate the CL sub-tree via SIMD search, then perform a scalar insert or delete within a single 64 B cache-line sub-node. CL sub-node splits and merges occur only on overflow or underflow.

9 Cache-Miss Attribution Analysis

This section attributes cache misses to individual functions using `perf record -e cache-misses`. The data reveals *where* in the code the processor stalls waiting for memory, complementing the aggregate hardware counters in Section 7.

Table 4: Top functions by cache-miss contribution (matryoshka, rand_insert, N=4,194,304).

% Cache Misses	Function
74.4%	`page_find_leaf`
55.4%	`mt_page_insert`
38.5%	`make_shuffled_keys(unsigned long, unsigned long)`
12.1%	`mt_inode_search`
5.7%	`matryoshka_insert`

9.1 Root Cause: CL Sub-Node Cache-Line Jumps

The dominant cache-miss source is the CL sub-tree navigation within each 4 KiB page leaf. When `page_find_leaf` descends the CL sub-tree, each level accesses a different 64 B slot within the page. These slots are not adjacent — slot indices are assigned by a bitmap allocator, so the root CL inode, its child CL inode, and the target CL leaf typically reside on non-contiguous cache lines.

The hardware prefetcher cannot predict these indirect jumps (each CL internal node stores child slot indices that must be read before the next cache line address is known). This creates a dependent chain of cache misses: *read child index* $\rightarrow$ *compute address* $\rightarrow$ *miss on that address* $\rightarrow$ *repeat.*

At N=4M with random inserts, the 4 KiB pages are spread across a ~20 MiB working set, exceeding L2 but fitting in L3. Each page access brings the header cache line into L1, but the CL sub-nodes deeper in the page have been evicted since the last visit to that page.

9.2 Optimisation: Software Prefetching

Software prefetch hints (`__builtin_prefetch`) have been added at three levels:

`page_find_leaf` (CL sub-tree descent)
After reading the page header's `root_slot`, the root CL node's cache line is prefetched before the loop begins. Within each loop iteration, after determining the child slot index from `cl_inode_search`, the child's cache line is prefetched before the next iteration accesses it. This converts the dependent miss chain into overlapped prefetch + computation.

`find_leaf` (outer B+ tree descent)
After `mt_inode_search` determines the child pointer, the first cache line of the child node is prefetched. For internal-to-internal transitions, this warms the next inode's header and first keys. For the last level (internal-to-leaf), this warms the page header so that `page_find_leaf` starts with the header in L2.

`mt_inode_search` (binary search in outer inodes)
For nodes exceeding the SIMD linear scan threshold, the binary search prefetches the midpoints of both candidate halves before each comparison. This converts the $O(\log N)$ dependent miss chain of binary search into a pipelined prefetch sequence.

All prefetches use `locality=1` (L2 hint), appropriate for data that will be used once in the near future but is unlikely to be reused before eviction.

9.3 Optimisation: Pointer Tagging for CL Root Prefetch

Instruction-level profiling revealed that the initial software prefetching approach had insufficient lead time: the CL root prefetch inside `page_find_leaf` was issued *after* loading the page header (to read `root_slot`), leaving only ∼5 ns of computation before the CL root access—far less than the ∼100 ns needed for an L3 miss to resolve.

To address this, leaf metadata is embedded in the low bits of child pointers in the outer B+ tree's internal nodes. All node pointers are 4096-byte aligned, providing 12 guaranteed-zero low bits. The encoding is:

Bits	Field	Range
0–5	`root_slot`	1–63 (CL sub-tree root slot index)
6–8	`sub_height`	0–7 (CL sub-tree height)
9–11	*reserved*	(available for future use, e.g. fullness counts)

With this scheme, `find_leaf` can extract `root_slot` from the tagged child pointer *without* reading the page header, enabling it to issue *two* prefetches simultaneously at the last outer-tree level:

1. The page header cache line (slot 0), as before;
2. The CL root cache line (`slots[root_slot − 1]`), computed from the tag.

Both prefetches issue at the same time, overlapping with whatever inode search computation follows. Tags are maintained at leaf creation (bulk-load, split) and updated after each insert or delete that may change `root_slot` or `sub_height`. Every child pointer read goes through an `mt_untag()` mask operation (a single AND instruction) to strip the low 12 bits before dereferencing. Stale tags on cold paths (rebalance) merely cause a wasted prefetch—never a correctness issue.

10 Detailed Results Table

Matryoshka rows highlighted in blue.

Table 5: Full benchmark results.

Library	Workload	N	Mop/s	ns/op
abseil_btree	mixed	4,194,304	2.04	491.3
abseil_btree	mixed	16,777,216	1.30	771.0
abseil_btree	rand_delete	4,194,304	1.20	836.0
abseil_btree	rand_delete	16,777,216	0.92	1,081.8
abseil_btree	rand_insert	4,194,304	1.11	898.3
abseil_btree	rand_insert	16,777,216	0.85	1,170.3
abseil_btree	search_after_churn	4,194,304	1.48	675.2
abseil_btree	search_after_churn	16,777,216	1.02	977.2
abseil_btree	seq_insert	4,194,304	5.18	193.0
abseil_btree	seq_insert	16,777,216	5.33	187.8
abseil_btree	ycsb_a	4,194,304	5.25	190.5
abseil_btree	ycsb_a	16,777,216	3.96	252.4
abseil_btree	ycsb_b	4,194,304	1.28	781.4
abseil_btree	ycsb_b	16,777,216	0.93	1,072.8
libart	mixed	4,194,304	2.04	490.0
libart	mixed	16,777,216	1.77	564.1
libart	rand_delete	4,194,304	1.80	556.2
libart	rand_delete	16,777,216	1.32	756.8
libart	rand_insert	4,194,304	2.15	465.0
libart	rand_insert	16,777,216	1.60	626.1
libart	search_after_churn	4,194,304	3.74	267.1
libart	search_after_churn	16,777,216	1.96	509.7
libart	seq_insert	4,194,304	5.17	193.5
libart	seq_insert	16,777,216	5.38	186.0
libart	ycsb_a	4,194,304	4.74	210.8
libart	ycsb_a	16,777,216	4.44	225.2
libart	ycsb_b	4,194,304	3.18	314.4
libart	ycsb_b	16,777,216	2.03	492.0
matryoshka	mixed	4,194,304	1.67	598.7
matryoshka	mixed	16,777,216	1.35	740.9
matryoshka	rand_delete	4,194,304	1.14	879.9
matryoshka	rand_delete	16,777,216	0.97	1,033.3
matryoshka	rand_insert	4,194,304	1.20	831.0
matryoshka	rand_insert	16,777,216	0.95	1,049.9
matryoshka	search_after_churn	4,194,304	1.43	698.1
matryoshka	search_after_churn	16,777,216	0.93	1,069.8
matryoshka	seq_insert	4,194,304	3.45	290.1
matryoshka	seq_insert	16,777,216	3.08	324.2
matryoshka	ycsb_a	4,194,304	3.38	296.2
matryoshka	ycsb_a	16,777,216	2.84	351.9
matryoshka	ycsb_b	4,194,304	1.29	775.5
matryoshka	ycsb_b	16,777,216	1.09	914.3
matryoshka_fence	mixed	4,194,304	1.61	622.0
matryoshka_fence	mixed	16,777,216	1.35	741.5
matryoshka_fence	rand_delete	4,194,304	1.22	818.3
matryoshka_fence	rand_delete	16,777,216	0.97	1,028.0

Continued on next page

Table 5: Full benchmark results (continued).

Library	Workload	N	Mop/s	ns/op
matryoshka_fence	rand_insert	4,194,304	1.19	839.4
matryoshka_fence	rand_insert	16,777,216	0.97	1,030.5
matryoshka_fence	search_after_churn	4,194,304	1.51	662.0
matryoshka_fence	search_after_churn	16,777,216	0.96	1,044.2
matryoshka_fence	seq_insert	4,194,304	3.95	253.2
matryoshka_fence	seq_insert	16,777,216	3.17	316.0
matryoshka_fence	ycsb_a	4,194,304	3.45	290.2
matryoshka_fence	ycsb_a	16,777,216	2.78	359.2
matryoshka_fence	ycsb_b	4,194,304	1.38	724.4
matryoshka_fence	ycsb_b	16,777,216	1.14	880.1
matryoshka_fence_sp	rand_insert	4,194,304	1.06	946.2
matryoshka_fence_sp	rand_insert	16,777,216	0.94	1,067.8
matryoshka_fence_sp	seq_insert	4,194,304	4.69	213.4
matryoshka_fence_sp	seq_insert	16,777,216	3.79	264.0
matryoshka_fence_sp	ycsb_a	4,194,304	4.48	223.1
matryoshka_fence_sp	ycsb_a	16,777,216	3.23	310.0
matryoshka_fence_sp	ycsb_b	4,194,304	1.41	711.6
matryoshka_fence_sp	ycsb_b	16,777,216	0.88	1,139.3
std_set	mixed	4,194,304	0.52	1,912.9
std_set	mixed	16,777,216	0.41	2,425.8
std_set	rand_delete	4,194,304	0.30	3,343.3
std_set	rand_delete	16,777,216	0.23	4,422.1
std_set	rand_insert	4,194,304	0.36	2,802.1
std_set	rand_insert	16,777,216	0.28	3,571.2
std_set	search_after_churn	4,194,304	0.27	3,737.2
std_set	search_after_churn	16,777,216	0.20	4,962.0
std_set	seq_insert	4,194,304	1.26	790.5
std_set	seq_insert	16,777,216	1.06	943.8
std_set	ycsb_a	4,194,304	1.09	916.3
std_set	ycsb_a	16,777,216	0.86	1,166.5
std_set	ycsb_b	4,194,304	0.28	3,624.4
std_set	ycsb_b	16,777,216	0.22	4,643.0
tlx_btree	mixed	4,194,304	1.32	758.3
tlx_btree	mixed	16,777,216	1.05	949.7
tlx_btree	rand_delete	4,194,304	0.87	1,150.8
tlx_btree	rand_delete	16,777,216	0.69	1,443.1
tlx_btree	rand_insert	4,194,304	0.91	1,096.0
tlx_btree	rand_insert	16,777,216	0.72	1,384.2
tlx_btree	search_after_churn	4,194,304	0.83	1,204.3
tlx_btree	search_after_churn	16,777,216	0.67	1,495.3
tlx_btree	seq_insert	4,194,304	4.95	202.0
tlx_btree	seq_insert	16,777,216	4.96	201.6
tlx_btree	ycsb_a	4,194,304	3.75	266.6
tlx_btree	ycsb_a	16,777,216	3.34	299.7
tlx_btree	ycsb_b	4,194,304	0.81	1,232.6
tlx_btree	ycsb_b	16,777,216	0.62	1,607.5

11 Analysis and Diagnosis

11.1 Benchmark Access Patterns

The seven workloads exercise distinct access patterns that interact differently with each data structure's memory layout. Understanding what each workload actually measures is essential for interpreting the results table: raw Mop/s numbers are meaningless without knowing which bottleneck—modification overhead, cache pressure, rebalancing cost, or in-leaf search efficiency—dominates a given workload.

`seq_insert` — **Sequential Append**

Keys arrive in ascending order (1, 3, 5, . . .). This is the *easiest* case for B-trees: leaves fill left-to-right with no splits until full, and the rightmost leaf stays hot in cache across consecutive inserts. Red-black trees rebalance on each insert but maintain temporal locality in the allocator; ART builds monotonically deeper paths in byte-order with no node splitting.

Matryoshka benefits from sequential CL sub-node filling within each page, but the nested sub-tree overhead—navigating two levels of CL internal nodes (2–3 SIMD comparisons) to reach the target CL leaf, then `memmove` within that 64 B node—makes each insert more expensive than a simple sorted-array append in a flat B-tree leaf. This workload primarily tests **per-operation modification overhead**, not cache behaviour, since the hot working set fits in L1/L2 regardless of structure.

`rand_insert` — **Random Cache Pressure**

A Fisher–Yates shuffle of $[0, N)$ scaled to odd values. Every insert touches a random leaf, maximising cache pressure. This is the workload most sensitive to node size and memory layout—the **primary benchmark for cache-conscious designs**.

- **Wide B-tree leaves** (tlx, abseil at 256–4096 B) amortise random access: one cache miss loads many keys, so a linear or SIMD scan within the leaf is cheap relative to the miss.
- **Pointer-chasing structures** (`std::set`) incur a cache miss per tree level ($\sim \log_2 N$ levels); at $N = 16\text{M}$ that is ∼24 dependent misses.
- **ART**'s fixed-depth byte-trie (4-byte key $\Rightarrow \leq 4$ levels) limits pointer-chase misses to at most 4, but each node may be 48–256 B depending on type.
- **Matryoshka**'s 4 KiB pages are cache-friendly, but the nested CL sub-tree adds 2–3 cache-line touches within each page. The pointer-tagging optimisation (§**??**) specifically targets this workload: prefetching the CL root while still in the outer tree removes one serial miss from the critical path.

`rand_delete` — **Rebalance Stress Test**

Bulk-loads N sorted keys (giving every structure its optimal starting layout), then deletes all keys in shuffled order. This isolates **rebalancing cost** from insertion:

- Red-black trees perform $O(1)$ rotations per deletion with small constant factors (3 pointer writes + colour flip).
- B-trees `memmove` within leaves and occasionally merge or redistribute siblings, touching 1–2 nodes.
- Matryoshka merges CL sub-nodes within a page (cheap—same cache line or adjacent lines, no system calls) and only performs expensive page-level merges when total occupancy drops below $\lfloor 855/4 \rfloor = 213$ keys. The two-level underflow propagation (CL merge $\rightarrow$ page merge) is more complex than a flat B-tree merge but amortises well since most deletions only affect CL sub-nodes.

`mixed` — **Steady-State Churn**

Alternating insert (new key beyond current max) and delete (random existing key) on a pre-loaded tree of size N. The tree size fluctuates around N, creating a **realistic steady-state** workload. This tests whether structures waste work on structural oscillation: a tree that aggressively splits a node on insert and immediately merges it on the next delete pays the cost of both operations with no net benefit. Structures with hysteresis between split and merge thresholds (matryoshka uses max/4 for underflow vs. max for split) handle this efficiently.

`ycsb_a` **(95% insert / 5% search) — Write-Heavy OLTP**
Modelled on the Yahoo! Cloud Serving Benchmark "Workload A." Inserts are sequential (monotonically increasing keys), so **append-path efficiency** dominates the 95% write portion. The 5% predecessor searches target recently-inserted regions that are likely still hot in L1/L2 cache, favouring structures with good temporal locality in their leaf layer. This workload reveals whether a structure's insert path is cheap enough to sustain high write throughput without being dragged down by occasional search overhead.

`ycsb_b` **(50% delete / 50% search) — Shrinking Tree**
Deletes keys from a pre-loaded tree interleaved with random predecessor searches. As the tree shrinks, occupancy drops and the ratio of useful data to allocated memory worsens. This tests whether **structural changes degrade search performance**: partially-filled pages waste cache capacity (fewer keys per cache miss), and ongoing merges may leave the tree in a suboptimal layout for search. Structures that reclaim space eagerly (matryoshka's CL sub-node merging) maintain higher effective density than those that leave tombstones or half-empty nodes.

`search_after_churn` **— Pure Search Isolation**
The tree undergoes insert/delete churn (untimed setup phase), then runs 5,000,000 random predecessor searches as the timed workload. This **isolates search throughput** from modification cost, making it the purest measure of in-leaf search efficiency and memory layout quality. The workload is most sensitive to:

1. *In-leaf search cost*: SIMD width (SSE2 at 4 keys vs. AVX2 at 8 keys per comparison), number of CL levels traversed, and branch misprediction rate.
2. *Memory layout*: cache-line utilisation (how many useful keys per 64 B line fetched) and whether the post-churn layout retains spatial locality.
3. *Tree height*: fewer outer-tree levels means fewer pointer-chase misses before reaching the leaf.

ART caveat: ART lacks native predecessor search; its wrapper falls back to point lookup (`art_search`), which is a fundamentally different and easier operation. ART's numbers in this workload are not directly comparable to the other structures.

Key encoding. All workloads use 4-byte `int32_t` keys encoded as odd values ($2i+1$), ensuring no key equals zero (used as a sentinel in CL sub-node headers). Keys are generated by xorshift64 with fixed seeds for reproducibility across runs and platforms.

11.2 Matryoshka: Nested Sub-Tree Tradeoffs

Each insert or delete navigates the page-level CL sub-tree (2–3 SIMD comparisons on 12–15 keys per level) to a target CL leaf, then performs a scalar `memmove` of at most 14 keys within that 64 B cache-line sub-node. The cost per modification is $O(h_s \times b)$ where $h_s \leq 2$ is the sub-tree height and $b = 15$ is the CL leaf capacity—roughly 30–45 key touches, all within a single 4 KiB page.

CL sub-node splits and merges occur only when a CL leaf overflows (15 keys) or underflows (< 7 keys). Page-level splits occur only when all 63 CL slots are exhausted ($\sim$855 keys/page).

The nested design adds a constant overhead per operation compared to flat sorted-array B-tree leaves, where a single `memmove` suffices. At N=1,048,576: matryoshka achieves 0.95 Mop/s on random insert, vs. 0.85 (Abseil) and 0.72 (TLX).

However, SIMD search through the CL sub-tree is used during both search *and* the navigation phase of insert/delete. This yields search-after-churn throughput of 0.93 Mop/s at N=1,048,576. The key advantage is that modifications touch a single cache-line sub-node rather than shifting an entire sorted leaf array.

11.3 Detailed Comparison: std::set (Red-Black Tree)

`std::set` uses a balanced binary search tree with one heap-allocated node per key (40–48 B on 64-bit systems: two child pointers, parent pointer, colour bit, key, allocator overhead).

Insert and delete. At N=16,777,216: 0.28 Mop/s (random insert) and 0.23 Mop/s (random delete)—the slowest of all libraries tested. Each operation traverses $O(\log_2 N)$ levels with a pointer dereference (and likely cache miss) at every level. The 40–48 B node size means $\sim$1 useful key per cache line, so every level is a full cache miss for large N.

Sequential insert. At 1.06 Mop/s (N=16,777,216), sequential insert is only modestly better than random because the allocator provides some temporal locality, but the red-black tree still requires $O(\log N)$ pointer chases and rotations.

Search. Search-after-churn: 0.20 Mop/s. Binary search through $\log_2 N \approx 24$ levels of pointer chasing is inherently cache-unfriendly. `std::set` has no mechanism for SIMD-accelerated search or cache-line-aware layout.

Scaling. `std::set` shows the steepest throughput degradation from small to large N (random insert scales 1.3× from N=4,194,304 to N=16,777,216) because the working set of pointer-chased nodes quickly exceeds cache capacity.

11.4 Detailed Comparison: TLX btree_set

TLX implements a B+ tree with sorted-array leaves. Leaf capacity is typically 64–128 keys (depends on template parameters and key size). Internal nodes use sorted arrays of separator keys with binary search.

Insert and delete. At N=16,777,216: 0.72 Mop/s (random insert) and 0.69 Mop/s (random delete). Each leaf insert is a binary search followed by a `memmove` of the leaf's sorted array. The average shift is $B/2 \approx 32$–64 keys per insert, but the entire operation stays within one or two cache lines for small leaves.

Search. Search-after-churn: 0.67 Mop/s. TLX uses scalar binary search within leaves, which incurs $\lceil \log_2 B \rceil$ comparisons with data-dependent branches. This is slower than SIMD linear scan for the same leaf size.

Sequential insert. 4.96 Mop/s. The B+ tree append path is efficient: new keys land at the rightmost leaf with minimal shifting, and splits propagate only when the leaf is full.

Comparison to matryoshka. TLX's flat sorted-array leaves have a lower constant factor per insert (one `memmove` vs. CL sub-tree navigation), but matryoshka's wider pages (855 keys vs. $\sim$128) reduce the outer tree height and number of leaf splits. At large N, the outer-tree traversal cost dominates, and matryoshka's SIMD-accelerated outer internal node search closes the gap.

11.5 Detailed Comparison: Abseil btree_set

Abseil's B-tree uses a similar sorted-array design to TLX but with different node sizes and allocation strategies. Leaf nodes hold up to ∼256 keys in a single sorted array.

Insert and delete. At N=16,777,216: 0.85 Mop/s (random insert) and 0.92 Mop/s (random delete). The wider leaves mean fewer tree levels and splits, but each `memmove` within a leaf shifts more keys on average ($B/2 \approx 128$).

Search. Search-after-churn: 1.02 Mop/s. Abseil uses scalar binary search within leaves. The wider leaves reduce tree height (fewer pointer chases) but increase the number of comparisons per leaf ($\lceil \log_2 256 \rceil = 8$ vs. $\lceil \log_2 128 \rceil = 7$ for TLX).

TLX vs. Abseil. On random insert at N=16,777,216, TLX and Abseil are within 15% of each other. Abseil's wider leaves trade cheaper outer traversal (fewer levels) for more expensive in-leaf operations (larger `memmove`). The two designs converge in throughput because the cache miss cost of locating the target leaf dominates at large N.

11.6 Detailed Comparison: libart (Adaptive Radix Tree)

ART uses a radix/trie structure with adaptive node types (Node4, Node16, Node48, Node256) that compact sparse levels. For 4-byte keys, the tree has at most 4 levels regardless of N.

Insert and delete. At N=16,777,216: 1.60 Mop/s (random insert) and 1.32 Mop/s (random delete). ART's $O(k)$ complexity (key length, not tree size) means insert cost is nearly constant across dataset sizes. The scaling ratio from N=4,194,304 to N=16,777,216 is 1.3× —the flattest of all structures tested.

Search. Search-after-churn: 1.96 Mop/s. ART achieves the highest absolute search throughput because its point lookups traverse ≤ 4 levels, each requiring a single indexed array access (no comparison-based search within nodes for Node256). *However*, the benchmark uses point lookups for ART rather than predecessor search (which ART does not natively support), so this comparison is not apples-to-apples with the other structures.

Access pattern interaction. ART's per-byte radix decomposition means key distribution matters less than key length. The uniform random keys in these benchmarks create well-distributed tries with few path-compressed nodes. A workload with clustered keys sharing long common prefixes would trigger more path compression and potentially different performance characteristics.

Memory overhead. ART's adaptive node types (4, 16, 48, or 256 children) trade memory for access speed. At high occupancy, most internal nodes are Node48 or Node256, using 256–2048 B per node regardless of actual fan-out—significantly more memory per key than B-tree or matryoshka designs.

11.7 Hardware Counter Comparison

Table 6: Hardware counters across all libraries (rand_insert, N=16,777,216).

Library	Cache Miss (%)	L1d Miss (%)	LLC Miss (%)	dTLB Miss (/1K ops)	Branch Miss (%)	IPC
`matryoshka`	79.1	3.9	28.9	2.6	3.9	0.55
`matryoshka_fence`	77.2	3.2	28.8	2.6	5.0	0.59
`matryoshka_fence_sp`	56.7	6.7	13.9	5.8	8.8	0.44
`std_set`	68.9	13.7	45.8	65.8	2.9	0.14
`tlx_btree`	68.6	4.6	34.8	14.9	10.4	0.46
`libart`	63.7	6.6	40.1	35.0	1.2	0.60
`abseil_btree`	67.9	6.6	37.7	15.9	7.0	0.61

The hardware counters reveal distinct micro-architectural profiles:

dTLB pressure. Matryoshka's arena allocator (2 MiB hugepage-backed regions) yields the lowest dTLB miss rate (2.6/1,000 ops). `std::set` has the highest (65.8/1,000) because each pointer chase to a heap-allocated node potentially touches a new 4 KiB page. The B-tree libraries (TLX, Abseil) fall between: their wider nodes reduce the number of distinct pages accessed per operation, but they use standard `malloc` without hugepage awareness. ART's adaptive nodes are heap-allocated but fewer in number than red-black tree nodes, yielding moderate dTLB pressure.

LLC misses. At N=16,777,216 (random insert), all structures exceed LLC capacity. Matryoshka's LLC miss rate is 287.9/1,000 ops vs. 457.7 for `std::set`. Matryoshka confines each leaf operation to a single 4 KiB page (or a few adjacent cache lines for CL sub-node navigation), while red-black tree operations touch $\log_2 N$ widely-scattered nodes. The B-tree libraries achieve comparable LLC rates because their wide leaves also localise work.

IPC. Matryoshka achieves IPC of 0.59 via SIMD search (`_mm_cmpgt_epi32`/`_mm_movemask_ps`) at every level, avoiding data-dependent branches. `std::set`'s pointer-chasing and branch-heavy traversal yields the lowest IPC. ART's indexed array lookups (no comparisons for Node256) can achieve high IPC on cache-hot paths.

Branch misprediction. SIMD mask arithmetic in matryoshka replaces conditional branches, yielding low branch misprediction rates. `std::set` and the scalar-binary-search B-trees (TLX, Abseil) have higher misprediction rates because each comparison is a data-dependent branch. ART's indexed-lookup approach avoids comparison branches within nodes but has conditional branches for node type dispatch.

11.8 Access Pattern Interactions with Data Structure Layout

The interaction between access pattern and memory layout explains much of the performance variation:

Sequential vs. random insert. Sequential insert favours structures with efficient append paths. All B-tree variants (matryoshka, TLX, Abseil) benefit because new keys land at the rightmost leaf. `std::set` benefits less because red-black rebalancing is oblivious to key order.

The throughput ratio (seq/rand) at N=16,777,216 reveals how much each structure benefits from locality: matryoshka 2.91× vs. std::set on sequential, 3.40× on random.

Delete after bulk-load vs. interleaved. The `rand_delete` workload starts from a bulk-loaded (optimally packed) tree, giving every structure its best starting point. The `mixed` workload, by contrast, operates on a tree that is continuously modified, creating internal fragmentation. Structures that maintain good occupancy under churn (B-trees with merge/redistribute) degrade less between these workloads than structures with per-node allocation (`std::set`).

The 4 KiB page boundary. Matryoshka's 4 KiB page leaves are sized to match the OS page size, ensuring that navigating the CL sub-tree within a leaf never crosses a page boundary. TLX and Abseil leaves are smaller (<1 KiB), so multiple leaves may share a page—good for spatial locality of adjacent leaves, but each leaf may straddle two cache lines for the `memmove` operation. `std::set` nodes are scattered across the heap with no page-alignment guarantees.

SIMD and branch prediction. Matryoshka's SIMD search produces a bit mask rather than a conditional branch, making it prediction-friendly. The sorted-array B-trees (TLX, Abseil) use scalar binary search with $O(\log B)$ data-dependent branches per leaf, which the branch predictor struggles with for uniform random keys (50/50 taken probability at each comparison).

11.9 Proposed Additional Access Patterns

Several workloads not currently benchmarked would reveal different performance relationships:

Zipfian (skewed) insert
A Zipfian distribution concentrates inserts on a small number of "hot" leaves. B-tree variants would benefit from cache-hot leaves; `std::set` would benefit from a cache-hot path of recently accessed nodes. Matryoshka's per-page CL sub-tree might show more frequent CL splits under concentrated load.

Range scan after insert
Iterate over a range of k keys (e.g. $k = 1000$) after building the tree. Matryoshka's linked leaf pages and dense packing should excel; `std::set`'s in-order traversal via parent pointers would lag. This would highlight the spatial locality advantage of contiguous leaf storage.

Interleaved point lookup and insert
A read-modify-write pattern ("contains then insert if absent") would test whether search and insert share cache-hot state. Matryoshka's search and insert paths share the same CL sub-tree navigation code, so a just-searched path remains cache-hot for the subsequent insert.

Large-key workload
Keys longer than 4 bytes (e.g. 16- or 32-byte strings) would stress ART's strength (key-length-dependent, not N-dependent traversal) while increasing matryoshka's CL sub-node overhead (fewer keys per 64 B cache line). B-tree `memmove` cost would grow linearly with key size.

Delete-heavy with searches (YCSB-D)
A workload where the tree shrinks from N to near-empty while servicing read queries. This would stress merge and redistribute paths, and test whether search throughput degrades as the tree becomes sparsely populated. Matryoshka's CL sub-node merge and page-level redistribute are designed for graceful degradation, but extreme sparsity (few keys spread across many pages) could hurt cache utilisation.

Bulk-load comparison
Timing the bulk-load operation itself (currently untimed) would highlight structural differences: matryoshka distributes keys across CL sub-nodes within pages in a single bottom-up pass; TLX and Abseil use repeated insertion; `std::set` has no bulk-load optimisation.

11.10 Overall Assessment

The matryoshka nesting design achieves competitive insert and delete throughput while preserving SIMD-accelerated search at every level of the hierarchy. Each modification touches a single cache-line sub-node rather than rebuilding an entire sorted leaf array. At the largest dataset size (N=16,777,216), matryoshka's throughput is within 1.7× of the best B-tree competitor on insert-heavy workloads and 1.4× on delete-heavy workloads.

The primary cost of the nesting design is a constant-factor overhead per modification (CL sub-tree navigation), which the flat sorted-array B-trees avoid. This overhead is most visible at small N where the outer tree is shallow and the per-operation cost is dominated by in-leaf work. At large N, where the outer tree traversal and cache misses dominate, the nesting overhead is amortised and matryoshka's SIMD search and dense page layout become decisive advantages.

12 Improvements Since Initial Report

12.1 Superpage-Level Nesting (Implemented)

The nesting now extends to three levels: CL sub-nodes (64 B) within 4 KiB pages within 2 MiB superpages. Each superpage contains a B+ tree of page-level sub-nodes, with page-level internal nodes (681 separator keys, 682 children per 4 KiB page) routing searches to up to 510 page leaves. Maximum capacity per superpage: 510 × 855 ≈ 436K keys. This confines most operations to a single TLB entry and reduces outer-tree height. Enable via `mt_hierarchy_init_superpage`.

12.2 Wider SIMD: AVX2 and AVX-512 (Implemented)

Compile-time SIMD width selection via `-DMT_SIMD=avx2` or `-DMT_SIMD=avx512`. AVX2 (256-bit) processes 8 keys per comparison in CL leaf predecessor search, CL internal search, and outer internal node search. AVX-512 (512-bit) processes 16 keys per comparison using masked operations. Unaligned loads handle the 4-byte header offset within CL sub-nodes. SSE2 (128-bit, 4 keys) remains the baseline fallback.

12.3 Batch Insert and Delete API (Implemented)

`matryoshka_insert_batch(tree, keys, n)` and `matryoshka_delete_batch(tree, keys, n)` sort incoming keys, group them by target leaf, and amortise outer-tree traversal across each group. On page-full or underflow, the path is re-navigated for remaining keys. Both functions work with page leaves and superpages.

12.4 CL Sub-Tree Cache-Miss Optimisation: Fence Keys vs. Eytzinger (Implemented)

Instruction-level profiling (`perf record -e cache-misses` with `addr2line`) of the baseline matryoshka on `rand_insert` at N=16M identified three serial cache-miss hotspots within the CL sub-tree traversal:

% Miss	Source	Description
59.8%	`leaf.c:323`	`page->header.sub_height` — first access to the page header cache line
38.0%	`leaf.c:199`	`cl->nkeys` in `cl_inode_search` — loading the child CL internal node
76.3%*	`leaf.c:451`	`cl->nkeys < MT_CL_KEY_CAP` in `mt_page_insert` — loading the target CL leaf

*Percentage of `mt_page_insert`'s cache misses, not total.

These form a serial dependency chain: header → CL root internal → child CL node. Each load depends on data from the previous load, so no out-of-order execution or hardware prefetching can overlap them. The pointer-tagging optimisation (§**??**) already addresses hotspot 1 by prefetching the CL root from tagged pointers while still in the outer tree, but hotspots 2 and 3 remain. Two strategies were implemented and benchmarked head-to-head:

12.4.1 Strategy A: Fence Keys

Store the CL root internal's separator keys and child slot indices in the 32 spare bytes of the page header (previously `_reserved`). Since the page header is always loaded first (hotspot 1 is unavoidable), the fence data comes "for free"—the CL root internal can be skipped entirely for height-1 sub-trees with ≤ 6 separators.

Header layout.

```
/* Replaces _reserved[32] in mt_page_header_t */
int32_t fence_keys[6]; /* 24 bytes */
uint8_t fence_slots[7]; /* 7 bytes */
uint8_t nfence; /* 1 byte */
                        /* 32 bytes total */
```

When fence keys apply.

- Height 1, ≤ 7 children: fence keys fully resolve the CL leaf → **skip CL internal entirely**
- Height 2, root has ≤ 6 children: fence keys skip the root internal → saves one level
- Height 1, 8–13 children: `nfence=0`, transparent fallback to normal path

Maintenance. A `refresh_fence_keys()` helper copies $\min(\text{nkeys}, 6)$ keys and $\min(\text{nkeys} + 1, 7)$ child slots from the CL root internal into the header. It is called after bulk load, CL root split (in `mt_page_insert`), and CL root collapse (in `mt_page_delete`).

Page capacity. Unchanged (855 keys/page). Outer tree structure unchanged. Enable via `mt_hierarchy_init_fence`.

12.4.2 Strategy B: Eytzinger Dense BFS Layout

Fix the CL sub-tree to height ≤ 1 with a dense BFS layout. The root always occupies slot 1; children occupy slots $2, 3, \ldots, N+1$. Since child positions are arithmetic (not stored in the CL internal), all children can be prefetched simultaneously while the root cache line is still loading—breaking the dependency chain between hotspots 1 and 2.

CL internal type. A new 64-byte Eytzinger internal stores 15 separator keys with no `children[]` array (positions are implicit):

```
typedef struct mt_cl_inode_eytz {
    uint8_t type; /* MT_CL_INTERNAL */
    uint8_t nkeys; /* 0-15 */
    uint8_t nchildren; /* 1-16 */
    uint8_t _pad;
    int32_t keys[15]; /* 60 bytes */
} mt_cl_inode_eytz_t; /* 64 bytes total */
```

Key trade-off. With 16 children × 15 keys = 240 keys max per page, pages split at 240 keys instead of 855. This means ∼3.5× more pages in the outer tree, but each page operation is faster due to the eliminated serial miss. When a CL leaf splits within an Eytzinger page, the dense BFS invariant is restored by extracting all keys and rebuilding the layout—amortised to ∼16 key copies per insert.

Enable via `mt_hierarchy_init_eytzinger`.

12.4.3 Benchmark Results

Table 7: Fence keys vs. Eytzinger vs. baseline: `rand_insert` (ns/op, lower is better).

N	Baseline	Fence	Δ	Eytzinger	Δ
65,536	307	245	−20%	791	+158%
262,144	352	276	−21%	819	+133%
1,048,576	406	388	−4%	803	+98%
4,194,304	652	510	−22%	1,051	+61%
16,777,216	719	706	−2%	1,123	+56%

Table 8: Fence keys vs. Eytzinger vs. baseline: `search_after_churn` (ns/op).

N	Baseline	Fence	Δ	Eytzinger	Δ
65,536	232	193	−17%	228	−2%
16,777,216	785	668	−15%	771	−2%

12.4.4 Hardware Counter Analysis

`perf stat` on `rand_insert` at N=16M (P-core):

Table 9: Hardware counters: fence keys vs. Eytzinger vs. baseline.

Variant	Cache Miss (M)	L1d Miss (M)	Instr (B)	IPC
Baseline	393	145	12.6	0.65
Fence	384	130	12.7	0.69
Eytzinger	504	172	25.3	0.85

Fence keys reduced L1d misses by 10.3% and total cycles by 4.5%, with IPC improving from 0.65 to 0.69. The instruction count is essentially unchanged (+1%), confirming that the fence key fast path adds negligible overhead—it merely resolves the CL leaf from data already in the header cache line.

Eytzinger achieved dramatically better IPC (0.85 vs. 0.65) and a lower cache miss *rate* (43.8% vs. 66.6%), confirming that the mass-prefetch of all 16 children is working as intended: the CPU executes useful computation while prefetches resolve in parallel. However, the absolute miss count is 28% higher (504M vs. 393M) and the instruction count doubled (25.3B vs. 12.6B) due to the taller outer tree (3.5× more pages) and the $O(240)$ extract-rebuild on every CL leaf split.

12.4.5 Cache-Miss Profile Shift with Fence Keys

`perf record -e cache-misses` on `rand_insert` at N=16M:

Table 10: Cache-miss attribution: baseline vs. fence keys.

Function	Baseline	Fence	Change
`page_find_leaf`	51.6%	48.2%	−3.4 pp
of which `cl_inode_search`	31.7%	22.6%	−9.1 pp
`mt_page_insert`	25.3%	28.8%	+3.5 pp
`mt_inode_search`	14.2%	14.2%	∼0
`matryoshka_insert`	3.0%	3.1%	∼0

The fence keys reduced `cl_inode_search` cache misses by 9.1 percentage points—exactly the CL root internal loads being skipped for height-1 sub-trees with ≤ 6 separators. The residual 22.6% comes from: (1) height-2 pages where fence keys skip the root but a level-1 internal still requires loading; and (2) height-1 pages with 7–12 separators that exceed the 6-key fence capacity.

Instruction-level attribution of the remaining hotspots with fence keys:

`page_find_leaf` at `leaf.c:199` (35.6%)
`int n = cl->nkeys` in `cl_inode_search`—loading the child CL internal after fence keys resolved the root. This fires on height-2 pages and height-1 pages with > 6 separators.

`mt_page_insert` at `leaf.c:62` (17.7% overall)
`int lo = 0, hi = cl->nkeys` in `cl_leaf_lower_bound`—loading the target CL leaf to find the insertion point. This is the third miss in the serial chain: header → (fence skip root) → child CL leaf. It is now the **dominant remaining bottleneck**.

`mt_inode_search` at `inode.c:91` (13.6%)
Binary search loop within outer tree internal nodes—standard outer-tree traversal cost, unaffected by CL-level optimisations.

12.4.6 Conclusions

1. **Fence keys are the clear winner.** 17–22% faster on insert and search at sizes fitting L3 cache (up to N=4M), with 2–15% improvement persisting at N=16M. Zero impact on page capacity or outer tree structure.
2. **Eytzinger is a negative trade-off.** Despite achieving better IPC (0.85 vs. 0.65) and a lower cache miss rate, the 3.5× page count increase (240 vs. 855 keys/page) and $O(240)$ extract-rebuild on every CL leaf split make it consistently slower—56–158% slower on insert across all sizes.
3. **The dominant remaining bottleneck** is loading the target CL leaf (`cl_leaf_lower_bound` at 17.7% of total cache misses). This is the irreducible third step in the serial chain that neither strategy can eliminate: the CL leaf's address is only known after searching the CL internal one level above.
4. **Fence keys hurt delete** by ∼15% due to the `refresh_fence_keys()` overhead after structural changes (merge, redistribute, root collapse). Future work could defer fence refresh to the next search, amortising the cost across multiple modifications.

12.4.7 Follow-Up: Mass Prefetch of Fence Children

The remaining bottleneck after fence keys is loading the target CL leaf (`cl_leaf_lower_bound` at 17.7% of total cache misses). The fence key fast path already knew all child slot positions (from `fence_slots[]` in the header), but issued a prefetch only *after* the fence search resolved the target child—leaving insufficient latency hiding between the prefetch and the first access.

The fix: prefetch *all* fence children *before* the fence search. At most 7 prefetches (7×64 B = 448 B), all within the same 4 KiB page (single TLB entry). Spurious prefetches cost only L1/L2 fill bandwidth. The fence search (≤ 6 comparisons) provides the latency window for the target child's cache line to arrive.

```
/* Prefetch ALL fence children before searching. */
int nf = page->header.nfence;
for (int c = 0; c <= nf; c++)
    __builtin_prefetch(get_slot_c(page,
                       page->header.fence_slots[c]), 0, 1);
int ci = fence_search(page->header.fence_keys, nf, key);
```

Table 11: Mass prefetch vs. fence (no mass prefetch) vs. baseline.

Workload	N	Baseline	Fence	Fence+MP	Δ vs Base
rand_insert	4M	768.4	510	577.2	−24.9%
rand_insert	16M	801.5	706	784.9	−2.1%
search_after_churn	16M	769.5	668	707.5	−8.1%

Throughput results (ns/op, lower is better).

Table 12: Hardware counters: mass prefetch vs. baseline.

Variant	Cache Miss (M)	L1d Miss (M)	Instr (B)	IPC
Baseline	364	145	12.5	0.66
Fence+MP	467	131	12.7	0.71

Hardware counters (P-core, `rand_insert`, N=16M). The mass prefetch reduced L1d misses by 9.7% (145M → 131M) and improved IPC from 0.66 to 0.71. Total reported "cache misses" rose (364M → 467M) because hardware counters include the speculative prefetches that hit L2/L3 but were not consumed—the *useful* miss rate (misses that stall the pipeline) decreased, as evidenced by the 4% wall-clock improvement at N=16M and 25% at N=4M.

Table 13: Cache-miss profile: fence+mass prefetch vs. baseline.

Function	Baseline	Fence+MP	Change
`page_find_leaf`	48.2%	32.0%	−16.2 pp
`mt_page_insert`	28.8%	22.2%	−6.6 pp
`mt_inode_search`	13.6%	2.8%	−10.8 pp
`make_shuffled_keys`	—	34.5%	(setup noise)

Cache-miss attribution with mass prefetch. Instruction-level annotation confirms the prefetch loop itself incurs 0% cache misses (the prefetch instructions never stall), and `page_find_leaf`'s remaining 32% is entirely the unavoidable first touch of the page header (`sub_height` load at function entry). The 88% concentration of `mt_page_insert`'s misses at `cl->nkeys` (offset 1 of the CL leaf) confirms that the serial chain has been compressed to its theoretical minimum: header → CL leaf, with the CL root internal eliminated by fence keys and the CL leaf latency partially hidden by mass prefetch.

12.4.8 Cross-Library Hardware Counter Comparison

Table 14: Hardware counters across all libraries (`rand_insert`, N=16M, P-core).

Library	Instr (B)	L1d Miss (M)	Cache Miss (M)	LLC Miss (M)	IPC	ns
matryoshka (fence+MP)	12.7	131	467	17.9	0.71	
matryoshka (baseline)	12.5	145	364	15.8	0.66	
Abseil btree	13.4	285	436	93.1	0.76	
TLX btree	11.7	212	267	46.0	0.60	
libart (ART)	13.9	308	495	137.7	0.91	
std::set (RB tree)	11.1	652	710	245.5	0.20	2

Key observations:

Matryoshka has the fewest L1d misses (131M) and LLC misses (17.9M).
The 4 KiB page design confines each operation to a single page, and the arena allocator's hugepage-backed regions keep the outer tree compact. Abseil has 2.2× the L1d misses and 5.2× the LLC misses despite similar throughput—it compensates with wider SIMD-free leaves that amortise each miss over more keys.

libart achieves the highest IPC (0.91) and throughput (384 ns/op).
ART's fixed-depth radix trie (≤ 4 levels for 4-byte keys) limits the serial dependency chain to ≤ 4 pointer chases. Its high cache miss count (495M) and LLC miss count (137.7M) are tolerated because each miss resolves an entire trie level, and the short chain means the CPU spends less time stalled per operation. ART also has the lowest branch misprediction rate (1.2%), since node dispatch uses indexed array lookups rather than comparison branches.

std::set is memory-bound (IPC 0.20).
Each of the ∼24 pointer chases per operation is a full cache miss, and the CPU stalls waiting for each one before computing the next address. The 652M L1d misses and 245.5M LLC misses reflect one miss per tree level per operation. Despite having the lowest instruction count (11.1B), the serial dependency chain dominates.

TLX has the fewest total cache misses (267M) but lowest IPC (0.60).
TLX's scalar binary search within sorted-array leaves uses data-dependent branches that the CPU struggles to predict (10% branch miss rate), causing pipeline flushes on every misprediction. The low cache miss count reflects good spatial locality of the moderate-sized leaves, but the branch mispredictions prevent the CPU from filling the pipeline efficiently.

12.5 Bottleneck Analysis at N=16M (Fence + Mass Prefetch)

At N=16M with `rand_insert`, the fence + mass prefetch variant achieves 776 ns/op (1.29 Mop/s) with the following per-operation budget:

Table 15: Per-operation budget (`rand_insert`, N=16M, P-core at ∼1.4 GHz effective).

Function	**Cycles/op**	**% cycles**	**Role**
`page_find_leaf`	412	37.3	Fence search + mass prefetch + CL sub-tree descent
`mt_page_insert`	280	25.3	CL leaf binary search + insert + split propagation
`mt_inode_search`	240	21.7	Outer tree binary search (339-key inodes)
`matryoshka_insert`	75	6.8	`find_leaf` dispatch + pointer retag
Other	99	8.9	`memmove`, setup, shuffled-key generation

Table 16: Cache-miss attribution (`rand_insert`, N=16M, P-core).

Function	**% misses**	**Dominant instruction**
`page_find_leaf`	44.1%	`sub_height` load — page header first touch
`mt_page_insert`	34.9%	`cl->nkeys` load — CL leaf first touch
`mt_inode_search`	13.5%	Binary search `cmp` + prefetch loads in loop

12.5.1 Bottleneck 1: Page Header Cold Miss (44% of cache misses)

The first touch of each 4 KiB leaf page (`sub_height` at byte offset 5) accounts for 44% of all cache misses. At N=16M, the working set is ∼78K pages × 4 KiB =∼312 MiB of leaf data — far exceeding the 24 MB L3 cache. Only ∼6K of the 78K pages fit in L3 at any time, yielding a ∼92% miss rate on random page access.

The pointer-tag prefetch in `find_leaf` fires the page header prefetch at the last outer-tree level, but only ∼30 cycles of inode search computation separate the prefetch from the first access in `page_find_leaf` — insufficient to hide an LLC miss (∼40–70 cycles) or DRAM miss (∼200 cycles).

12.5.2 Bottleneck 2: CL Leaf Cold Miss (35% of cache misses)

After fence keys resolve the target CL leaf slot, the mass prefetch fires all fence children before the fence search. The fence search provides ∼12–18 cycles of latency window (6 comparisons). At N=16M, the CL leaf's cache line has been evicted since the last visit to this page, and ∼18 cycles is insufficient to hide the LLC/DRAM latency. The prefetch converts a *stall* into a *partial overlap*, but the residual miss remains the dominant bottleneck within `mt_page_insert` (74.4% of its cache misses at the `cl->nkeys` load).

12.5.3 Bottleneck 3: Outer Inode Binary Search (13.5% of cache misses, 21.7% of cycles)

`mt_inode_search` binary-searches 339-key outer internal nodes (4 KiB each, spanning 64 cache lines). The binary search touches $\lceil \log_2 339 \rceil \approx 9$ cache lines per search, with the prefetch-both-halves technique overlapping adjacent iterations. The prefetch instructions themselves show 0% cache misses (they never stall), but the *data loads* after the prefetch still miss when the prefetch hasn't completed in time. The cache misses are spread across the loop body: `cmp` instruction at 5.3%, `movslq` at 12.9%, and prefetch targets at 7.6%.

At N=16M the outer tree has height 1 (root inode with ∼230 children). The root inode's 4 KiB page spans 64 cache lines; repeated accesses keep the hot lines in L2, but the binary search's access pattern touches ∼9 lines per query across different regions of the node. Branch misprediction contributes 9.3% of all branches missed (∼11 mispredicts per insert), with ∼4 coming from the outer inode binary search (data-dependent `cmp/jle` at ∼50% taken probability).

12.5.4 Hardware Counters at N=16M

Table 17: Hardware counters: fence + mass prefetch (`rand_insert`, N=16M, P-core).

Counter	Value
Cycles	18,551,518,503
Instructions	12,968,103,304
IPC	0.70
Cache references	691,719,328
Cache misses	482,020,663 (69.7%)
L1d load misses	134,692,327
LLC load misses	18,685,337
dTLB load misses	8,565,404
Branch misses	185,437,635 (9.3% of 1,998,387,729)

Per-operation: ∼1106 cycles, ∼773 instructions, ∼8.0 L1d misses, ∼1.1 LLC misses, ∼0.5 dTLB misses, ∼11 branch mispredicts. The 0.70 IPC indicates that ∼30% of potential throughput is lost to cache and branch stalls.

12.6 Potential Throughput Improvements

12.6.1 A. Enable Superpages (Already Implemented)

With 2 MiB superpages (up to 510 pages per superpage), the 78K leaf pages at N=16M fit in ∼153 superpages. The outer tree shrinks from 78K children (height 1 with a 4 KiB root inode) to 153 children (root inode fits in a few hundred bytes, entirely in L1). Page-level internal nodes within each superpage are co-located in the same 2 MiB region, improving TLB and spatial locality.

The superpage infrastructure is already implemented (`mt_hierarchy_init_superpage`); enabling it for the fence variant is a configuration change.

Expected improvement: 15–25% at $N > 4$M (eliminates most outer-tree overhead, reduces TLB misses).

12.6.2 B. Batched Prefetch Pipelining

For bulk workloads, sort incoming keys and group by target leaf. While inserting key[i] into its CL leaf, prefetch the page header for key[i+1]'s target leaf. This completely hides the 44% page-header miss by overlapping it with the previous key's CL leaf insert (∼280 cycles of useful work — more than enough to hide even a DRAM miss).

The batch insert API already exists (`matryoshka_insert_batch`) but does not currently pipeline prefetches across keys.

Expected improvement: 20–30% for batch workloads (eliminates the dominant page-header cold miss entirely).

12.6.3 C. Branchless Binary Search in Outer Inodes

Replace the data-dependent `cmp/jle` branches in `mt_inode_search`'s binary search loop with CMOV-based branchless updates:

```
while (lo < hi) {
    int mid = lo + (hi - lo) / 2;
    int go_right = (keys[mid] <= key);
    lo = go_right ? mid + 1 : lo; /* CMOV */
    hi = go_right ? hi : mid; /* CMOV */
}
```

This eliminates ∼4 branch mispredicts per insert × ∼15 cycle penalty =∼60 cycles saved per operation. The CMOV version is also more amenable to the CPU's load speculation since there are no mispredicted branches to flush.

Expected improvement: 5–8% (reduces branch misprediction from 9.3% to ∼6%).

12.6.4 D. FAST-Layout Outer Inodes

The FAST paper [[1]] describes a cache-line-blocked layout for internal nodes that reorganises keys so each binary search step lands on a different cache line, with the tree structure matching the cache hierarchy. Applied to the 339-key outer inodes:

- Level 0 (root line): 1 key, splits the node in half
- Level 1: 2 keys, one per half
- . . . until cache-line-sized blocks of ∼15 keys

Each level's cache line can be prefetched one level ahead. The current sorted-order layout requires ∼9 random cache line accesses within the 4 KiB inode; the FAST layout converts this to a BFS descent where each level is a single cache line, enabling systematic prefetching.

Expected improvement: 10–15% (reduces outer inode search from ∼240 to ∼150 cycles by eliminating ∼2 serial cache line misses).

12.6.5 E. Compile-Time Strategy Specialisation

The fence key fast path currently checks `hier->cl_strategy == MT_CL_STRAT_FENCE` at runtime on every insert. Compile-time specialisation (via `__attribute__((flatten))` or template instantiation) would:

- Eliminate runtime strategy branches (∼2–3 per insert)
- Allow the compiler to inline `page_find_leaf` into `mt_page_insert`, fusing the CL leaf resolution with the insert
- Reduce function call overhead (∼20 cycles for `callq`/`retq` + register save/restore)

Expected improvement: 3–5% (reduces instruction count and eliminates function call overhead).

12.6.6 Summary

Table 18: Potential throughput improvements for `rand_insert` at N=16M.

Approach	**Bottleneck addressed**	**Est. gain**	**Effort**
Superpages (config change)	Outer tree height + TLB	15–25%	Low
Batched prefetch pipelining	Page header cold miss (44%)	20–30%	Medium
Branchless inode binary search	Branch misprediction (9.3%)	5–8%	Small
FAST-layout outer inodes	Inode search cache misses	10–15%	Large
Compile-time specialisation	Runtime dispatch overhead	3–5%	Medium

Approaches A–C are independent and composable; their improvements should be roughly additive. Approach D (FAST layout) is orthogonal to A–C but requires significant restructuring of the outer inode format. The combined ceiling from A+B+C is ∼40–60% improvement for batch workloads, potentially bringing matryoshka to ∼500 ns/op at N=16M — competitive with libart's 384 ns/op.

12.7 Implementation Results: Approaches A–C

Approaches A (superpages + fence keys), B (batched prefetch pipelining), and C (branchless binary search in outer inodes) have been implemented and benchmarked.

12.7.1 C. Branchless Binary Search — Implementation

The `mt_inode_search` binary search loop (lines 89–102 of `inode.c`) was converted from a data-dependent `if/else` branch to arithmetic bit-masking:

```
int cmp  = (keys[mid] <= key);   // 0 or 1
int mask = -cmp;                 // 0x00000000 or 0xFFFFFFFF
lo += ((mid + 1) - lo) & mask;   // lo = mid+1 if cmp, else unchanged
hi += (mid - hi) & ~mask;        // hi = mid if !cmp, else unchanged
```

Disassembly confirms the compiler emits `setle` + `cmovle` + `and` instead of conditional jumps in the inner loop body. The only branch is the well-predicted `while (lo < hi)` loop exit.

12.7.2 B. Batched Prefetch Pipelining — Implementation

The `matryoshka_insert_batch` function was enhanced with two optimisations:

1. **Sibling-advance fast path.** When sorted keys cross a leaf boundary, the code checks whether the next key falls into the next sibling child within the same parent inode. Since outer inodes hold up to 340 children, this skips the full outer-tree re-walk for ∼339/340 of leaf transitions.
2. **Eager next-sibling prefetch.** At the start of each leaf group, the code prefetches the next sibling leaf's page header and CL root slot via the parent's tagged child pointer. The ∼400–855 inserts for the current leaf provide ample latency to hide the DRAM fetch.

These optimisations apply only to the batch insert API; single-key `matryoshka_insert` is unchanged.

12.7.3 A. Superpages + Fence Keys — Implementation

A new hierarchy initialiser `mt_hierarchy_init_fence_sp` combines fence keys (`cl_strategy = MT_CL_STRAT_FENCE`) with 2 MiB superpage allocation (`use_superpages = true`). The two features compose naturally: the superpage code delegates to page-level functions that already check `cl_strategy`.

12.7.4 Benchmark Results: Approaches A–C

Table 19: Throughput (Mop/s) for `rand_insert` after implementing approaches A–C. All matryoshka variants include branchless binary search (approach C). Batch prefetch (B) only affects `insert_batch`.

Variant	N=4M	N=16M	ns/op (16M)	**vs. ART**
Matryoshka (baseline)	1.87	1.11	901	0.54×
Matryoshka + fence	1.32	0.98	1,020	0.48×
Matryoshka + fence + SP	1.17	0.94	1,064	0.46×
libart (ART)	2.52	2.06	486	1.00×
TLX btree_set	1.20	0.83	1,205	0.40×
Abseil btree_set	1.41	0.97	1,031	0.47×

12.7.5 Hardware Counter Impact

Table 20: Selected hardware counters for `rand_insert` at N=16M. Cache-miss ratio = cache-misses / cache-references.

Metric	**Baseline**	**Fence**	**Fence+SP**	**libart**
Cache-miss ratio	76.9%	76.1%	53.3%	64.8%

Key findings from the hardware counter analysis:

- **Cache miss rate:** Superpages reduce the cache-miss ratio from ∼76% (fence) to ∼53% (fence+SP) at N=16M. The 2 MiB TLB entries and spatial co-location of pages within superpages substantially improve LLC hit rates.
- **Branchless binary search:** The `setle/cmovle` codegen eliminates data-dependent branches in the outer inode search. For inodes with >64 keys (where the AVX2 linear scan falls back to binary search), this saves ∼4 mispredictions per search at ∼15 cycles each.
- **Superpage throughput:** Despite the dramatic cache-miss reduction, the fence+SP variant shows slightly lower throughput than plain fence keys for random inserts. The additional indirection through the superpage page-level B+ sub-tree (finding the correct 4 KiB page within the 2 MiB superpage) adds instructions that offset the cache improvement.
- **Known limitation:** Large-scale random deletion with superpages triggers a pre-existing infinite loop in the superpage rebalance code (`rebalance_sp`). The issue is in the superpage merge path when multiple superpages need to be consolidated. Delete-heavy workloads (rand_delete, mixed, search_after_churn) are not available for the fence+SP variant at $N \geq$2M.

12.7.6 Progress Toward libart Parity

At N=16M random insert, the current standing:

- **libart (ART):** ∼2.06 Mop/s (486 ns/op)
- **Matryoshka + fence (with branchless search):** ∼0.98 Mop/s (1020 ns/op) — 2.1× slower than ART

- **Matryoshka + fence + superpages:** ∼0.94 Mop/s (1064 ns/op) — cache misses cut from 76% to 53%, but superpage overhead offsets the gain

The gap remains substantial (2×). The primary remaining bottleneck is the CL sub-tree traversal within each leaf page, which requires 2–3 serial cache-line loads. The approaches implemented here (branchless search, batch prefetch, superpages) address the outer-tree descent but cannot reduce the intra-page serial dependency chain. Closing the gap would require restructuring the page-level layout (e.g., FAST-layout inodes, or flattening the CL sub-tree to a sorted array with SIMD search for small pages).

12.8 Future: Variable-Length Keys

The current 4-byte `int32_t` key format could be extended to variable-length keys by storing key offsets or using indirection within CL sub-nodes. This would broaden applicability at the cost of some cache efficiency.

References

[1] C. Kim, J. Chhugani, N. Satish, E. Sedlar, A. D. Nguyen, T. Kaldewey, V. W. Lee, S. A. Brandt, and P. Dubey. *FAST: Fast Architecture Sensitive Tree Search on Modern CPUs and GPUs.* SIGMOD '10, pp. 339–350, 2010.

[2] R. Bayer and E. McCreight. *Organization and Maintenance of Large Ordered Indexes.* Acta Informatica, 1(3):173–189, 1972.

[3] V. Leis, A. Kemper, and T. Neumann. *The Adaptive Radix Tree: ARTful Indexing for Main-Memory Databases.* ICDE '13, pp. 38–49, 2013.

[4] J. Rao and K. A. Ross. *Making B+-Trees Cache Conscious in Main Memory.* SIGMOD '00, pp. 475–486, 2000.