Quake: Adaptive Indexing for Vector Search

Jason Mohoney
University of Wisconsin-Madison

Devesh Sarda
University of Wisconsin-Madison

Mengze Tang
University of Wisconsin-Madison

Shihabur Rahman Chowdhury
Apple

Anil Pacaci
Apple

Ihab F. Ilyas
University of Waterloo

Theodoros Rekatsinas
Apple

Shivaram Venkataraman
University of Wisconsin-Madison

Abstract

Vector search, the task of finding the k-nearest neighbors of high-dimensional vectors, underpins many machine learning applications, including recommendation systems and information retrieval. However, existing approximate nearest neighbor (ANN) methods perform poorly under dynamic, skewed workloads where data distributions evolve. We introduce Quake, an adaptive indexing system that maintains low latency and high recall in such environments. Quake employs a hierarchical partitioning scheme that adjusts to updates and changing access patterns, guided by a cost model that predicts query latency based on partition sizes and access frequencies. Quake also dynamically optimizes query execution parameters to meet recall targets using a novel recall estimation model. Furthermore, Quake utilizes optimized query processing, leveraging NUMA-aware parallelism for improved memory bandwidth utilization. To evaluate Quake, we prepare a Wikipedia vector search workload and develop a workload generator to create vector search workloads with configurable access patterns. Our evaluation shows that on dynamic workloads, Quake achieves query latency reductions of 1.5–22× and update latency reductions of 6–83× compared to state-of-the-art indexes SVS, DiskANN, HNSW, and SCANN.

1 Introduction

Vector search, the task of finding the k-nearest neighbors (KNN) of high-dimensional vectors, is fundamental to machine learning applications such as recommendation systems [20,21,27,28,37] and information retrieval [9,12,13,30]. In these applications, each vector represents an item in a metric space, and the distance between vectors reflects semantic similarity. However, performing exact KNN search becomes computationally infeasible on large datasets due to the high dimensionality and volume of data.

To address this challenge, practitioners use approximate nearest neighbor (ANN) indexes, which trade off a controlled amount of search accuracy (recall) for significant reductions in latency. Among these indexes, there are two widely-used types: graph-based indexes and partitioned indexes.

Maintaining low latency, high recall vector search under **dynamic and skewed workloads** remains a significant challenge for existing indexes. Real-world applications often exhibit non-uniform query distributions and evolving data. For example, in an example Wikipedia search application, popular pages like *Lionel Messi* or *LeBron James* receive disproportionately more queries, resulting in *skewed read patterns*. Additionally, pages are frequently added, updated, or deleted, causing *skewed update patterns* that change over time [6]. These factors degrade the performance of existing indexes, leading to increased query latency and reduced recall.

Graph-based indexes, such as HNSW [22] and DiskANN [32,33] construct a proximity graph where each node (vector) is connected to its approximate neighbors. Queries traverse the graph to find approximate nearest neighbors, typically achieving high recall with low latency. However, these indexes face challenges with dynamic workloads because updating the graph structure to accommodate frequent insertions and deletions is computationally intensive [39], due to the random access patterns involved in graph traversal and modification.

Partitioned indexes, such as SCANN [10, 34], SPANN [7, 39], and Faiss-IVF [8], partition the vectors using a clustering algorithm (e.g k-means). Queries are processed by scanning a subset of partitions, balancing recall and latency by adjusting the number of partitions scanned (denoted as *nprobe*). While attractive due to their simplicity, partitioned indexes face a significant search latency gap when compared with graph indexes. For example, on the MSTuring10M benchmark [2], we found Faiss-IVF takes 44ms per search query while Faiss-HNSW takes only 6.8ms. On the other hand, supporting updates in partitioned indexes is less expensive than for graph indexes, as the index structure needs minimal modification when adding or removing vectors. But, existing approaches struggle with dynamic and skewed workloads because they do not adapt to changing access patterns, leading to *imbalanced* partitions that degrade query latency. Recent work has been proposed to resolve imbalances in dynamic workloads by splitting and reclustering imbalanced

partitions [6, 39], however, we find these methods degrade recall as nprobe needs to change as the index structure changes.

In this work, we study the problem of minimizing query latency to meet a fixed recall target for dynamic vector search workloads, where both the queries and the base vectors can change over time. To address this problem, we develop **Quake**, a partitioned index system that minimizes latency by adapting the index structure to the workload. Quake's two primary algorithmic contributions are:

First, Quake employs an **adaptive hierarchical partitioning** scheme that modifies the partitioning by minimizing the cost (derived from a **cost model**) of a query. The cost model tracks partition sizes and access frequencies as the workload is processed and determines which partitions are most negatively contributing to overall query latency. Once identified, we conduct maintenance on them by splitting or merging them based on what is expected to reduce the cost (latency). We also demonstrate our maintenance procedure is stable and converges to a local minimum of the cost model.

Second, we design an **adaptive partition scanning** scheme that adjusts the number of partitions scanned on-the-fly to meet recall target for individual queries. We do this by maintaining a recall estimate during query processing based on the a) geometry of the partitioning and b) intermediate results of the query, and once the estimate exceeds the recall target, query processing terminates and the results are returned.

Furthermore, Quake utilizes NUMA-aware parallelism in order to maximize memory bandwidth usage on multi-core machines.

It is a significant challenge to evaluate indexing approaches due to the lack of availability of benchmarks for online vector search. To address this challenge and comprehensively evaluate our approach, we A) prepare a **Wikipedia vector search workload** derived from publicly available query and update patterns of Wikipedia pages and B) develop a **workload generator** for creating workloads with configurable query and update patterns. We will publicly release the Wikipedia workload and workload generator as evaluation tools for the community to use. Using this, we conduct a comprehensive evaluation of Quake in comparison to seven baseline approaches.

1. Quake achieves the lowest search latency across all dynamic workloads in comparison to state-of-the-art graph indexes, with $1.5-8\times$ lower search latency than HNSW and DiskANN while having $6-83\times$ lower update latency.
2. We also find that APS Scanning matches the nprobe of an oracle across recall targets on Sift1M, with only a 20% increase in latency relative to the oracle. In addition, APS exhibits stable recall on the Wikipedia-12M workload, with only a .005% average deviation from a recall target of 90% over the course of the workload.
3. Quake's NUMA-aware intra-query parallelism exhibits linear scalability and saturates memory bandwidth on the MSTuring100M dataset with 100-million records. Quake achieves $16\times$ lower query latency when compared to a single-threaded version and $4\times$ lower latency compared to a non-NUMA aware configuration.

2 Motivation and Challenges

Efficient vector search is critical for large-scale systems used in recommendation, semantic search, and information retrieval. These applications demand the ability to process a high volume of nearest neighbor queries with low latency, even as the underlying data evolves. To meet these requirements, vector databases—such as Milvus [36], Pinecone [3], Analytic-DBV [38], VBASE [41], and Qdrant [1]—utilize specialized vector indexes (e.g., Faiss-IVF, HNSW, Vamana) that support fast approximate nearest neighbor (ANN) queries. However, serving these dynamic workloads introduces significant challenges in maintaining query performance and accuracy as data and query patterns shift over time.

2.1 Vector Search Workload

A vector search workload is a continuous, evolving stream of **queries** and **updates**:

- **Queries:** Given a query vector q, the goal is to find the top-k nearest neighbors in a set $\mathbf{X}$. Exact linear search is too slow for large, high-dimensional datasets, so ANN indexes are used. These indexes approximate nearest neighbors with controlled recall to lower latency by orders of magnitude.
- **Updates:** The dataset evolves over time. Insertions add new vectors representing fresh content (e.g., new products, trending news articles), and deletions remove outdated entries. Typically, updates are applied in a batched fashion.

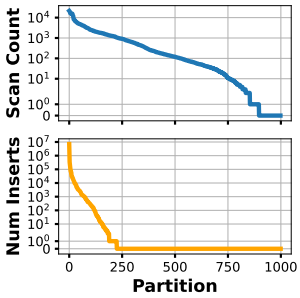
Recall@k is the standard metric for accuracy, defined as: $\frac{|\mathbf{G}\cap\mathbf{R}|}{k}$ where $\mathbf{R}$ is the vectors returned by the approximate search, and $\mathbf{G}$ is the ground truth set. Maintaining a consistent recall target (e.g., > 0.9) and low latency (e.g., milliseconds per query) as both data and query patterns shift is a key challenge. The complexity of these workloads stems from their inherently dynamic and skewed nature, which few existing indexing methods handle gracefully.

2.2 Why Real-World Workloads are Hard

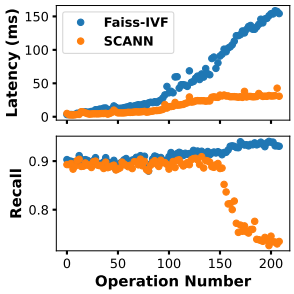
Skewed Read Patterns In practice, user queries concentrate on popular items. For example, queries against a Wikipedia-derived dataset tend to focus on a small subset of entities at any given time. As a result, certain partitions or graph regions of the index are accessed disproportionately often.

Skewed Write Patterns Insertions and deletions are also rarely uniform. New data often arrives in bursts—e.g., new Wikipedia pages added monthly, new products introduced ahead of a shopping season, or newly relevant embeddings generated by continuously updated language models.

Real-World Example: Wikipedia-12M In our evaluation, we prepared Wikipedia-12M, a workload based on a subset of Wikipedia articles derived from publicly available monthly pageview statistics [4]. Over 103 months, the dataset grows



(a) Read (top) and write skew.



(b) Query performance.

Figure 1: Skewed access patterns of Faiss-IVF index partitions in the Wikipedia-12M workload and their effect on query performance for Faiss-IVF and SCANN

Table 1: Comparison of updatable vector indexes. **Tuning**: Requires manual parameter tuning in indexing and query processing. **Maintenance**: Restructures index with incremental updates. **Adaptive**: Utilizes query information to inform indexing.

Method	Tuning	Maint.	Adaptive
Quake (**Ours**)	✗	✓	✓
Faiss-IVF [8]	✓	✗	✗
DeDrift [6]	✓	✓	✗
SpFresh [39]	✓	✓	✗
SCANN [10, 34]	✓	✓	✗
DiskANN [26, 32]	✓	✓	✗
Faiss-HNSW [22]	✓	✗	✗
SVS [5]	✓	✓	✗

from millions to tens of millions of vectors. Popular articles dominate query traffic, while embeddings of newly created pages accumulate in certain regions of the embedding space. This workload shows read skew and write skew, as evidenced by Figure 1a, reads and writes predominately affect a small portion of the index.

2.3 Shortcomings of Existing Approaches

Existing indexes were often developed and evaluated under assumptions of static data distributions; conditions not met in real-world use cases. Table 1 compares a range of well-known and state-of-the-art vector indexes. Although they are widely adopted in vector databases, none fully solve the problem of maintaining low-latency, high-recall search under dynamic, skewed workloads without constant manual intervention or offline tuning.

Graph Indexes Graph-based index systems, such as HNSW [22], DiskANN [26, 32], and SVS [5] construct a proximity graph where each node represents a vector connected to its approximate neighbors. These indexes achieve high recall with low latency in static settings by efficiently traversing the graph to locate nearest neighbors using a process known as *greedy traversal*. However, maintaining the graph structure under frequent insertions and deletions is computationally intensive, as each update may require rewiring multiple edges to preserve graph connectivity and proximity properties. Our evaluation (Table 1) shows that update latency can be multiple orders of magnitude higher than partitioned indexes. Additionally, graph traversal involves random memory accesses, leading to poor memory bandwidth utilization.

Partitioned Indexes Partitioned indexes such as Faiss-IVF [8], SCANN [10, 34], and SpFresh [39] divide the vector space into disjoint partitions using a clustering algorithm such as k-means. Queries are processed by scanning a subset of partitions to retrieve approximate nearest neighbors. Partitioned indexes are more update-friendly than graph-based method since insertions and deletions leads to sequential access. For write skewed workloads some partitions become significantly larger, degrading query latency, this can be exacerbated by read skew if large partitions are also more frequently accessed ("hot partitions"). Query processing is *memory-bound*, as achieving high recall requires scanning many megabytes of data across multiple partitions, for example reaching a recall target of 90% on the MSTuring100M dataset requires each query to scan 1GB of vectors. Moreover, most partitioned indexes use a fixed number of partitions to probe (*nprobe*), which does not adapt to changing data distributions or query patterns, leading to either insufficient recall or excessive data scanning. The challenges yield subpar performance for partitioned indexes on real-world workloads. For example, Figure 1b shows the degradation of latency and recall over time when using Faiss-IVF and SCANN with a fixed nprobe on Wikipedia-12M (workload details in Section 7).

Early Termination Early-termination methods have been proposed to reduce query latency or meet recall targets in partitioned indexes by dynamically adjusting the number of partitions scanned per query. **SPANN** [7] applies a simple rule: it prunes partitions once the centroid distance exceeds a user-tuned threshold relative to the closest centroid. **LAET** [17] is a learning-based approach that predicts the required *nprobe* per query using a trained model, but still requires dataset-specific training and calibration for each recall target. **Auncel** [42] uses a geometric model to estimate when recall for a given query, setting *nprobe* per query, but its conservative estimation leads to substantial overshooting of the recall target (See Figure 13 in [42]). All three methods require tuning or calibration and do not adapt to changes in the index structure or data distribution.

2.4 Technical Challenges for Partitioned Indexes

The following technical challenges have yet to be solved by existing partitioned indexes

1. **Adaptation to Queries** Query adaptivity is overlooked by existing partitioned index approaches and exhibits an opportunity for optimization, particularly for maintaining

hot partitions induced by read skew.

2. **Online Adjustment of Nprobe** As the index structure and data changes, partitioned indexes need to adjust the number of partitions scanned otherwise recall will suffer. Existing early termination works are insufficient as they assume a static index and require retuning as the index and data changes.
3. **Performance Gap with Graph Indexes** Standard partitioned indexes such as Faiss-IVF are memory bound, and exhibit an order of magnitude higher query latency in comparison than graph indexes.

Quake is our solution to these technical challenges. Quake A) adapts the index structure to queries by utilizing maintenance that minimizes a cost model for query latency, B) using a recall estimation model, Quake individually sets nprobe for queries to meet recall targets as the index structure changes, and C) uses NUMA-aware parallelism in order to saturate memory bandwidth during query processing, closing the performance gap with graph indexes. We next describe Quake in detail.

3 Solution Overview

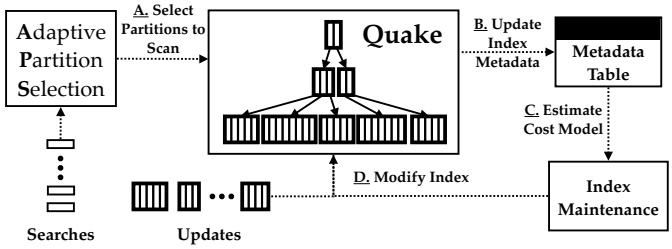


Figure 2: Quake Architecture Diagram. Search queries use Adaptive Partition Selection (APS) to determine which partitions to scan (**A**). Scanning partitions modifies access patterns of the index, tracked in the metadata table (**B**). A cost model is used to determine which maintenance actions to take (**C**) where the chosen maintenance actions modify the index (**D**). This process operates in a continuous online fashion as search and update queries (inserts/deletes) are issued to the index.

Index Structure In Quake, vectors are partitioned into disjoint partitions (using k-means) where each partition has a representative centroid. Centroids can be further partitioned in a similar manner to add additional levels to the index. Search queries scan the index structure in a top-down fashion, finding the nearest centroids to determine the partitions to scan in the next level. Partitions in the bottom level contain the base vectors and subsets of these partitions are scanned to return the k-nearest neighbors. Utilizing a multi-level design mitigates the cost of scanning centroids, allowing us to use fine-grained partitioning of vectors at large scale, which prior work has shown to be beneficial [7].

Adaptive Incremental Maintenance Inserts and deletes modify the Quake data structure by appending vectors to and removing vectors from index partitions. Insertions traverse the index structure top-down to find the nearest partition in the base level to the inserted vector and append to that partition. Deletes use a map to find the partition containing the vector to be deleted and the vector is removed from the partition with immediate compaction. As demonstrated in Section 2, modifications can negatively affect index performance over time, requiring maintenance (Figure 1b). Quake uses the following maintenance actions in order to minimize query latency:

1. **Split Partition:** Uses k-means to split a partition into two, removing the old partition and its centroid and adding two new partitions and centroids. To mitigate potential overlap due to the new partitions, we perform additional iterations of k-means over the partitions neighboring the split partitions (by centroid distance).
2. **Delete Partition:** Removes a partition and its centroid, reassigning the vectors of the deleted partition to the remaining partitions the index.
3. **Add Level:** Adds a level of partitioning to the index by partitioning the current top-level using k-means.
4. **Delete Level:** Removes current top-level and merges the partitions in the next level.

Quake uses a cost model that estimates query latency to determine if maintenance actions should be taken and which partitions to apply them to. The cost model is a function of partition access patterns and sizes to determine which partitions are most contributing to the overall query latency. We check for maintenance after each operation by evaluating the cost model, but the maintenance frequency is configurable. Partitions with the largest cost contribution are considered for split or deletion. Intuitively, frequently accessed and/or large partitions are split and infrequently accessed and/or small partitions are deleted as they do not justify the overhead of maintaining a centroid. See Section 4 for details on the cost model and maintenance methodology.

Adaptive Partition Scanning In order to determine the number of partitions a search query should scan to reach a given recall target, we apply *Adaptive Partition Scanning* (APS) at each level of the index. APS solves a critical problem for partitioned indexes when applied to dynamic workloads: as the number and contents of partitions change, the number of partitions scanned needs to change, otherwise recall will degrade (Figure 1b). APS maintains a recall estimator based on the intermediate top-k results of the query and the geometry of neighboring partitions. As more partitions are scanned, the intermediate results and recall model are updated and when the recall estimate exceeds the target recall, the results are returned. To mitigate overheads introduced by the recall estimator, we use pre-computation of expensive geometric functions and only update the estimate when the intermediate results have changed significantly. APS supports both euclidean and inner-product distance metrics. We cover APS

in Section 5.

NUMA-Aware Query Processing Modern multi-core servers often use Non-Uniform Memory Access (NUMA) architectures, where memory close to a processor's local node is faster to access than remote memory. Quake is designed to capitalize on this heterogeneous memory. It distributes index partitions across NUMA nodes. To minimize remote memory access, Quake employs affinity-based scheduling, and supports work stealing within a NUMA node to mitigate workload imbalances. By co-locating computation with the relevant data, Quake reduces remote memory accesses, saturates memory bandwidth, and thus lowers query latency. See Section 6 for further details on Quake's NUMA-aware optimizations.

4 Adaptive Incremental Maintenance

We present our adaptive maintenance methodology, beginning with a cost model that estimates each partition's contribution to query latency and guides maintenance decisions. Next, we describe the available maintenance actions, analyzing their impact on the cost model. We then detail the multi-stage decision workflow that prioritizes beneficial actions and conclude with a concrete example.

4.1 Cost Model

The cost model estimates the query latency contributed by each partition, in the index. Estimating the per-partition latency contribution enables targeted maintenance to the partitions most affecting query performance.

Partition Properties: Consider an index with L levels, numbered $l = 0, 1, \ldots, L-1$. Level l contains N_l partitions. The base level corresponds to $l = 0$ and contains partitions of the original dataset vectors. Higher levels contain partitions of centroid vectors that summarize the partitions in the level below. At the top level, $l = L-1$, there is a single partition containing top-level centroids.

Each partition j at level l has a size s_{lj} (the number of vectors it contains) and an access frequency $A_{l,j} \in [0.0, 1.0]$. $A_{l,j}$ denotes the fraction of queries, measured in a sliding window W, that scan the partition j at level l. The cost model is primarily driven by these sizes and access frequencies.

Partition Cost A partition (l, j) contributes latency proportional to its size and how frequently it is accessed. Let $\lambda(s)$ be the latency function for scanning s vectors. We measure $\lambda(s)$ through offline profiling. The cost of partition (l, j) is:

$$C_{lj} = A_{lj} \cdot \lambda(s_{lj}) \tag{1}$$

Total Cost: The overall query latency (cost) estimate is the sum across all levels and partitions

$$C = \sum_{l=0}^{L-1} \sum_{j=0}^{N_l-1} A_{lj} \cdot \lambda(s_{lj}) \tag{2}$$

Interpretation The cost model reflects the relationship between partition size, access frequency, and query latency. The fundamental trade-off that needs to be balanced is the number and size of partitions. Larger partitions require more time to scan, increasing latency, but reducing the total number of partitions and the overhead of scanning centroids. Conversely, smaller, fine-grained partitions reduce the number of vectors needed to scan to reach a high recall but increase the overhead of scanning centroids. The model further captures that frequently accessed partitions dominate the total cost, motivating targeted maintenance actions to balance these trade-offs.

Guiding Maintenance Decisions Maintenance actions such as splitting or deleting aim to reduce the total cost C. Each action is evaluated based on its predicted change in cost:

$$\Delta C = C_{\text{after}} - C_{\text{before}} \tag{3}$$

where C_{before} and C_{after} are the total costs before and after the action, respectively. Actions are applied only if $\Delta C < -\tau$, where τ is a non-negative tunable threshold, ensuring monotonic improvement in query performance. By focusing on reducing C, the index is dynamically restructured to maintain efficient query performance under varying workloads.

4.2 Conducting Maintenance

Maintenance at each level of the index proceeds in three phases—*estimate*, *verify*, and *commit / reject*. We first list the available actions, then derive their cost deltas, describe the workflow, and finally walk through a concrete split-verification example.

4.2.1 Maintenance Actions

To minimize query latency, Quake employs a series of maintenance actions that dynamically adjust the index structure in response to evolving workloads. Here we define the maintenance actions and then analyze the impact of each maintenance action on the overall cost model.

Split Partition: If a partition (l, j) is too large or frequently accessed, we consider splitting it into two partitions (l, j_L) and (l, j_R). We apply k-means clustering within that partition, forming two smaller partitions with their own centroids. The original partition is removed and its vectors are reassigned. A subsequent *partition refinement* step adjusts vector assignments to ensure minimal overlap and balanced partition sizes.

Partition Refinement After a split, refinement uses k-means (seeded by current centroids) on nearby partitions to mitigate overlap and ensure that each vector is assigned to its most representative partition. Nearby partitions are determined by finding the r_f nearest centroids to the split centroids, where r_f is a tunable parameter (typically between 10 and 100). This is a generalization of the reassignment procedure used in SpFresh [39], using additional rounds of k-means prior to reassignment. Refinement ensures vectors are correctly assigned to their most representative partition and avoids performance degradation due to excessive overlap.

Delete Partition If a partition is rarely accessed and below a minimum size threshold, we consider deleting it to remove the cost of maintaining its centroid. After deletion, the vectors are reassigned to their respective nearest existing partitions. This can reduce total cost by removing a low-benefit partition, although the reassignment may increase the size (and thus cost) of other partitions, and therefore careful consideration is needed before conducting a delete.

Adding and Removing Levels If the number of centroids in the top level grows beyond a threshold, we add a new top level by clustering those centroids into fewer groups. Conversely, if the top level becomes too sparse (below a configured lower threshold), we remove the top level and merge its centroids in the level below. Both actions help maintain hierarchy balance and control centroid-scanning overhead (details in Appendix).

4.2.2 Cost Deltas

The maintenance loop treats every candidate action as a *proposed edit* to the index and scores it by the change it would induce in the total cost (Eq. (3)).

We tentatively accept an action whenever $\Delta C < -\tau$. Below we give the exact ΔC formulas for the primary maintenance actions: *split* and *delete*. Full derivations are in the Appendix; here we show only the final expressions and explain the terms.

Exact split delta. Splitting a hot or oversized partition (l, j) into children (l, j_L) and (l, j_R) inserts one new centroid at the parent level, changing the overhead by $\Delta O^+ = \lambda(N_l + 1) - \lambda(N_l)$. The resulting cost difference is

$$\begin{aligned}\Delta\text{Split}_{l,j} \;=\; &\underbrace{\Delta O^+}_{\text{new centroid}} - A_{l,j}\,\lambda(s_{l,j}) \\ &+ A_{l,j_L}\,\lambda(s_{l,j_L}) + A_{l,j_R}\,\lambda(s_{l,j_R}).\end{aligned} \tag{4}$$

where the first term pays for the extra centroid, the second removes the old scan cost, and the last two add the costs of scanning the new, smaller partitions. Note that we do not explicitly model the effect of refinement, as refinement does not change the number of partitions. Its impact is captured automatically as statistics are collected from future queries, so we omit it from the Δ-formula and let later maintenance iterations adjust if necessary.

Exact delete delta. Deleting a cold, tiny partition (l, j) removes a centroid ($\Delta O^- = \lambda(N_l - 1) - \lambda(N_l)$) and redistributes its vectors to a receiver set $R_{l,j}$. Let Δs_m and ΔA_m be the resulting size and frequency bumps for each receiver m. Then

$$\begin{aligned}\Delta\text{Del}_{l,j} \;&= \Delta O^- - A_{l,j}\,\lambda(s_{l,j}) \\ &+ \sum_{m \in R_{l,j}} \big[(A_m + \Delta A_m)\,\lambda(s_m + \Delta s_m) - A_m\,\lambda(s_m)\big].\end{aligned} \tag{5}$$

captures both the benefit of deleting the partition and the penalty of swelling its neighbors.

The need for an estimate At decision time we do not yet know the post-action quantities $\{s_{l,j_L}, A_{l,j_L}, \dots\}$ or the true $\Delta s_m, \Delta A_m$. We therefore use a lightweight *estimate* based on two assumptions: 1) **Balanced Split:** $s_{l,j_L} \approx s_{l,j_R} \approx \frac{s_{l,j}}{2}$, and 2) **Proportional-Access Scaling**: each child inherits a fixed fraction α of the parent's frequency.

Under these assumptions the split estimate becomes

$$\begin{aligned}\Delta\text{Split}'_{l,j} \;&= \Delta O^+ - A_{l,j}\,\lambda(s_{l,j}) \\ &+ 2\alpha A_{l,j}\,\lambda\big(\tfrac{s_{l,j}}{2}\big).\end{aligned} \tag{6}$$

and the analogous delete estimate, derived with a uniform redistribution assumption, is located in the Appendix. Immediately after a tentative action we measure the *actual* sizes (and, for deletes, the exact receiving partitions) and re-evaluate Eqs. (4) or (5). If the recomputed gain is still below $-\tau$ the action is committed; otherwise it is rolled back (§4.2.3). This "estimate-then-verify" strategy is crucial for ensuring monotonic cost improvement.

4.2.3 Decision Workflow

Maintenance is a *bottom-up* pass over the hierarchy. Each level executes the five stages below starting from the base level. This workflow is assumed to be triggered by the user, an interesting avenue for future work is to develop scheduling and budgeting policies to call this workflow and limit its scope.

Stage 0 — Track statistics. At the end of each query batch we update, for every partition (l, j): (i) size $s_{l,j}$, (ii) access count over the sliding window of queries W, giving $A_{l,j} = \text{hits}(l, j)/|W|$. These values feed directly into the cost formulas.

Stage 1 — Estimate. For the current level l compute the estimate Δ' (§4.2.2) of splitting and deleting for every partition. Tentatively apply any action with $\Delta' < -\tau$.

Stage 2 — Verify. Immediately after performing a tentative action, we measure the actual resulting partition sizes (and the exact receiver partitions for deletes). We recompute the cost delta using these known values but retain the original frequency assumptions from Stage 1.

Stage 3 — Commit / Reject.

$$\Delta < -\tau \;\rightarrow\; \textit{commit}, \quad \Delta \geq -\tau \;\rightarrow\; \textit{reject}.$$

Rejection discards the action and keeps the previous state of the partition(s), in order to prevent cost increases.

Stage 4 — Propagate upward. Repeat Stages 1-3 on the next level l+1,

Safety. Because every level enforces the same $\Delta < -\tau$ guard, total cost across *all* levels monotonically decreases and the hierarchy converges to a stable state under a fixed workload distribution (proof in Appendix).

4.2.4 Example Maintenance Workflow

Below we walk through the *estimate → verify → commit / reject* loop for two example partitions and show how an imbalanced split is automatically rejected to prevent accidental cost increases.

Set-up. Consider partitions P_1 and P_2, where each contain $s = 500$ vectors and appear in $A = 0.10$ of queries. From profiling we observe scan latencies for the following partition sizes: $\lambda(50) = 250\mu s$, $\lambda(250) = 550\mu s$, $\lambda(450) = 1050\mu s$, $\lambda(500) = 1200\mu s$ Adding a centroid costs $\Delta O^{+} = 60\mu s$. We use a decision threshold of $\tau = 4\mu s$ and $\alpha = .5$

1. **Estimate.** The estimate assumes a balanced $250/250$ split and $\alpha = 0.5$ traffic per child:

$$C_{\text{before}} = 0.10 \times 1200 = 120\mu s,$$
$$C_{\text{est}} = 0.05 \times (550 + 550) = 55\mu s,$$
$$\Delta' = 60 - 120 + 55 = -5\mu s.$$

   Because $\Delta' < -\tau$, *both partitions are tentatively split*.
2. **Verify.** After splitting we see that P_1 does split $250/250$, but P_2 comes out $450/50$:

$$C_{\text{verify}}(P_2) = 0.05 \times (1050 + 250) = 65\mu s,$$
$$\Delta(P_2) = 60 - 120 + 65 = +5\mu s.$$

3. **Decision.**
   - P_1: *commit* because $\Delta = -5\mu s < -\tau = -4\mu s$
   - P_2: *reject* because $\Delta = +5\mu s > -\tau = -4\mu s$

   The verify step therefore blocks an imbalanced split that would otherwise *increase* query latency.

5 Adaptive Partition Scanning (APS)

Adaptive Partition Scanning (APS) dynamically determines the number of partitions to scan per query to achieve a specified recall target τ_R with minimal latency. APS adapts to evolving workloads and changing index structures, making it particularly effective in dynamic data settings. We first introduce the geometric model underlying APS, followed by a detailed description of the scanning algorithm, and conclude with key performance optimizations. We apply APS at each level of the index independently.

Let ρ denote the distance from query $\mathbf{q}$ to its current k-th nearest neighbor, defining a d-dimensional ball $\mathcal{B}(\mathbf{q}, \rho)$.

Geometric Model To estimate the probability that each partition contains one of the query's k nearest neighbors, APS uses a geometric interpretation. Given query $\mathbf{q}$ and the current search radius ρ (distance to the k-th nearest neighbor found so far), consider the hypersphere $\mathcal{B}(\mathbf{q}, \rho)$. Under a uniform-density assumption, the fraction of this sphere's volume intersecting partition $\mathcal{P}_i$ estimates the probability that $\mathcal{P}_i$ holds at least one true nearest neighbor:

$$p_i = \frac{\text{Vol}\big(\mathcal{B}(\mathbf{q}, \rho) \cap \mathcal{P}_i\big)}{\text{Vol}\big(\mathcal{B}(\mathbf{q}, \rho)\big)}, \tag{7}$$

Intersection Volume Approximation Exact computation of intersection volumes between a sphere and high-dimensional Voronoi partition boundaries is infeasible, as partitions are intersections of multiple half-spaces. Instead, we approximate each partition as a single half-space, defined by the perpendicular bisector between the query's nearest centroid $\mathbf{c}_0$ and each neighboring centroid $\mathbf{c}_i$. This simplification results in a hyperspherical cap whose volume v_i has a closed-form expression via the regularized incomplete beta function [18] (see Appendix).

Nearest Partition Volume Approximation The half-space approximation is invalid for the nearest partition $\mathcal{P}_0$, since the query lies within it. Instead, we first compute hyperspherical cap volumes v_j for the remaining $M-1$ candidate partitions and normalize these so that $\sum_{j=1}^{M-1} v_j = 1$. The probability p_0 that no neighbor is located outside $\mathcal{P}_0$ is:

$$p_0 = \prod_{j=1}^{M-1} (1 - v_j),$$

with the remaining probability distributed proportionally among other partitions:

$$p_i = (1 - p_0)\, v_i.$$

Thus, these probabilities sum to 1 and provide valid recall estimates.

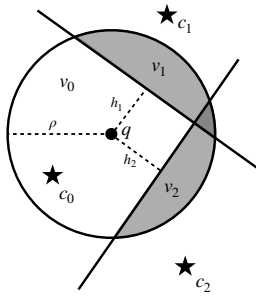


Figure 3: The query ball (centered at $\mathbf{q}$ with radius ρ) intersecting partition boundaries. The intersection volumes v_1 and v_2 correspond to the probability of finding a nearest neighbor in partitions $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively.

5.1 APS algorithm

Algorithm 1 details the APS procedure. Given query $\mathbf{q}$, recall target τ_R, and an initial set of M candidate centroids:

1. Scan partition $\mathcal{P}_0$, initializing the query radius ρ.
2. Compute probabilities p_i for each remaining candidate partitions based on radius ρ.
3. Iteratively scan partitions in descending probability order until cumulative recall exceeds target τ_R, updating radius ρ and recomputing probabilities whenever ρ shrinks significantly (beyond threshold τ_ρ).

Table 2: Mean single-threaded query latency and recall for APS variants on SIFT1M dataset at recall target 0.9. **APS-RP:** recomputes probabilities after each partition scan without precomputation. **APS-R:** recomputes after each partition scan with precomputation. **APS:** recomputes probabilities only if query radius changes by more than $\tau_\rho = 1\%$, using precomputed beta function values.

Configuration	Recall	Search Latency
APS	.901	.52 ms
APS-R	.901	.61 ms
APS-RP	.901	.70 ms

This process is conducted at each level of the index. To avoid propagating errors from searching higher levels of the index, we fix the recall target to .999 for the higher levels.

Algorithm 1 Adaptive Partition Scanning (APS)

Input: query $\mathbf{q}$, centroids $\mathcal{C}$, recall target τ_R, initial candidate count M, recompute threshold τ_ρ, k
Output: k nearest neighbors of $\mathbf{q}$

$R \leftarrow$ empty max-heap of size k
$\mathcal{S} \leftarrow M$ nearest centroids to $\mathbf{q}$
scan $\mathcal{P}_0$; update R; set ρ
for each $(\mathbf{c}_i, \mathcal{P}_i) \in \mathcal{S} \setminus \{\mathbf{c}_0\}$ **do**
    compute p_i
$r \leftarrow p_0$; $m \leftarrow 1$
while $r < \tau_R$ **and** unscanned candidates remain **do**
    choose i with maximal p_i; scan $\mathcal{P}_i$; update $\mathcal{H}$
    $m \leftarrow m + 1$
    $\rho' \leftarrow$ distance to k-th in $\mathcal{H}$
    if $|\rho' - \rho| > \tau_\rho \rho$ **then**
        $\rho \leftarrow \rho'$; recompute p_j
    $r \leftarrow r = \sum_{i=0}^{m-1} p_i$
return R

Performance Optimizations APS incorporates two optimizations to minimize computational overhead. First, it precomputes values of the regularized incomplete beta function at 1024 evenly spaced points in $[0,1]$ and linearly interpolates during queries. Second, partition probabilities are recomputed only when the query radius ρ shrinks by more than a relative threshold τ_ρ. Table 2 shows these optimizations reduce query latency by 25% on SIFT1M without sacrificing recall.

6 Quake Implementation

Here we discuss NUMA-aware query processing and implementation details of Quake.

NUMA Data Placement and Query Processing Query processing in partitioned vector indexes is memory-bound, and therefore increasing the effective memory bandwidth available to the system will reduce query latency. NUMA-aware intra-query parallelism has been applied in the context of relational database systems to great success [16, 29], but has yet to be applied to vector databases.

In order to maximize memory bandwidth utilization, Quake distributes index partitions across NUMA nodes and ensures that cores only scan partitions resident in their respective node. Quake assigns index partitions to specific NUMA nodes using round-robin assignment. This assignment procedure allows for simple load balancing as partitions are added to the index by the maintenance procedure.

To maximize the benefits of data placement, Quake employs partition affinity and NUMA-aware work scheduling. Partitions are bound to specific CPU cores. This binding ensures that partitions are always scanned by the same core to maximize cache utilization. Queries are scheduled to worker threads based on the location of the data partitions they need to access. When a query requires scanning multiple partitions, the work is divided among threads on the relevant NUMA nodes where the partitions reside. By aligning thread execution with data placement, Quake minimizes remote memory accesses and ensures maximum utilization of memory bandwidth.

NUMA-Aware Query Execution with APS Quake integrates NUMA-aware processing with Adaptive Partition Selection (APS) to dynamically select which partitions to scan based on query requirements and desired recall. The query processing involves both worker threads scanning local partitions and a main thread coordinating the process.

Algorithm 2 NUMA-Aware Query Processing with Adaptive Partition Selection

Input: Query vector q, Index partitions P_i with locations $Node_j$, Recall threshold τ_R, Recall model M, Period to check recall T_{wait}
Output: Top-k nearest neighbors to q satisfying recall threshold τ

Initialize: $R \leftarrow \emptyset$ (global result set), $S \leftarrow$ sorted list of partitions based on distance to q (obtained from searching parent)
Distribute q to local memory of NUMA nodes
for all NUMA nodes $Node_j$ **in parallel do**
    $W_j \leftarrow$ worker threads on $Node_j$
    $P_j \leftarrow$ partitions on $Node_j$ from S
    Enqueue partitions P_j to local job queue
while not **all** partitions in S have been processed **do**
    Main Thread:
    Wait for a predefined interval T_{wait}
    Merge partial results from worker threads into R
    Estimate current recall r using model M and results in R
    if $r \geq \tau_R$ **then**
        Break and terminate worker threads
Return top-k results from R
function WORKERTHREAD(q, Local Job Queue)
    while Job Queue not empty **do**
        $P_i \leftarrow$ **Dequeue** next partition from Job Queue
        Compute distances between q and vectors in P_i
        Update local partial results R_j
        Signal Main Thread about new partial results

Algorithm Explanation: In Algorithm 2, the main steps are:

1. **Initialization**: The query vector q is distributed to the local memory of NUMA nodes with relevant partitions. Partitions are sorted using their centroid distance to q.
2. **Worker Threads Execution**: Each NUMA node has worker threads that process partitions assigned to that node. They compute distances between q and vectors in their local partitions, updating their local partial results.
3. **Main Thread Coordination**: The main thread periodically merges partial results from all worker threads. It uses a recall model M to estimate the current recall based on the results accumulated so far.
4. **Adaptive Termination**: If the estimated recall meets or exceeds the threshold τ_R, the main thread returns the top-k results and signals the worker threads to terminate the processing of remaining partitions.
5. **Continuation or Termination**: If the recall threshold is not met and all partitions have been processed, the algorithm returns the current results.

This adaptive approach ensures that the system processes only as much data as needed to meet the recall requirements, improving efficiency and reducing query latency.

Implementation Details We implemented Quake in 7,500 lines of C++ and provide a Python API for ease-of-use. We used primitives in Faiss [8], PyTorch [19], and SimSIMD [25] to enable high-performance management of inverted lists, efficient batch tensor operations, and AVX512 intrinsics for fast distance comparisons. We also used a high performance concurrent queue [24] to prevent contention during coordination of query processing. In addition, we developed a workload generator and evaluation framework in Python to create and evaluate vector search workloads. We will release Quake as an open-source system along with its evaluation tools, the Wikipedia-12M workload, and Workload Generator.

7 Experiments

We next evaluate Quake using a number of benchmarks and summarize our main findings:

1. Quake achieves the lowest search latency across all dynamic workloads in comparison to state-of-the-art graph indexes, with $1.5-8\times$ lower search latency than HNSW, DiskANN, and SVS while having $6-83\times$ lower update latency.
2. We also find that APS Scanning matches the nprobe of an oracle across recall targets on Sift1M, with only a 20% increase in latency relative to the oracle. In addition, APS exhibits stable recall on the Wikipedia-12M workload, with only a .005% average deviation from a recall target .90% over the course of the workload.
3. Quake's NUMA-aware intra-query parallelism exhibits linear scalability and saturates memory bandwidth on the MSTuring100M dataset with 100-million records. Quake achieves $16\times$ lower query latency when compared to a single-threaded version and $4\times$ lower latency compared to a non-NUMA aware configuration.

7.1 Workloads

We performed our evaluation on a diverse set of real-world and synthetic workloads, focusing on search and update latency, recall, and scalability.

Wikipedia-12M This dataset and workload trace are derived from monthly Wikipedia page additions and page-view [4] frequencies between April 2013 and December 2021. We consider only pages about people or those linking to people. The dataset begins with 1.6 million pages and grows to 12 million after 103 updates, and therefore the average update size is $\approx$ 100,000 vectors. Embeddings are generated by training DistMult [40] graph embeddings (via Marius [23, 35]) on the Wikipedia link structure, and use the inner product metric.

The workload simulates monthly inserts of new pages, followed by 1,000 search queries sampling page embeddings with probability proportional to their page views. This setting imitates evolving interest and periodic growth of the dataset.

OpenImages-13M Using the methodology described by SVS [5], we generate a workload of 13M images from the Open Images dataset [15]. Embeddings are produced using Clip [31] in an inner product metric space. The workload maintains a sliding window of 2 million resident vectors and inserts and deletes vectors based on class labels until all 13 million vectors have been indexed at least once. Each insert and delete affects roughly 110K vectors. After each insert and delete operation, we run 1,000 queries randomly sampled from the entire vector set. This scenario stresses both insertion and deletion performance as well as sustained query latency.

Workload Generator To test performance under different workload characteristics, we employ a configurable workload generator applicable to any vector dataset. The key parameters include: number of vectors per operation, operation count, operation mix (insert/delete/search ratio), and spatial skew. For skewed workloads, the generator forms clusters and samples queries and updates from these clusters, reflecting hot spots in the vector space.

We construct two example workloads from a 10M vector subset of the MSTuring [2] dataset using L2 distance:

- **MSTuring-RO**: A pure search workload. We uniformly sample from 10,000 provided query vectors and execute 100 search operations, each querying 1,000 vectors. This setup tests query efficiency in a static setting.
- **MSTuring-IH**: A dynamic workload interleaving inserts and searches. Beginning with 1 million vectors, the dataset grows to 10 million as we process 1,000 operations with a 90% insert and 10% search ratio. This tests the ability to handle large-scale growth while maintaining query quality.

We also used the vector datasets Sift1M [14] and MSTuring100M [2] to conduct microbenchmarks.

7.2 Experimental Setup

Experiments are run on a machine with 80 cores, 500GB of memory with four NUMA nodes providing a total of 300 GB/s memory bandwidth. Search queries are processed one at a time and we report the average latency per query to reach a recall target of 90% recall for $k = 100$. Unless otherwise stated, all search-latency numbers use a single worker thread. Quake additionally reports a multi-worker (**MW**, 16 threads) For updates, both Quake and the baselines process updates in batches using 16 threads; we report the average update latency per vector (i.e., total batch update time divided by batch size). This setup simulates an online environment where queries arrive individually, and updates are applied in batches. We report maintenance time separately from update latency, as maintenance can be conducted in the background in online systems [39].

Baselines: We compare Quake against several state-of-the-art methods, including both partitioned and graph-based indexes:

- **Faiss-IVF** [8]: A popular inverted file (IVF) index in Faiss. It handles updates but does no maintenance.
- **DeDrift** [6]: An incremental maintenance strategy designed to reduce clustering drift by periodically reclustering large partitions together with small ones. We implement DeDrift's logic within Quake.
- **LIRE** [39]: Maintenance procedure used by SpFresh. LIRE incrementally splits large clusters and deletes small clusters after updates, followed by local reassignments. We implement LIRE's approach within Quake.
- **ScaNN** [11]: A state-of-the-art highly optimized partitioned index system. It uses an unpublished incremental maintenance procedure similar to LIRE.
- **Faiss-HNSW** [22]: A graph-based approach (HNSW) implemented in Faiss. It supports incremental inserts but not deletes. Thus, for workloads with deletions, we omit Faiss-HNSW from those comparisons.
- **DiskANN** [32]: System built around the Vamana [33] index with support for dynamic updates.
- **SVS** [5]: A recently released optimized implementation of the Vamana index with support for dynamic updates.

We configure the main parameters of Quake and the baselines as follows. We disable vector quantization/compression for all baselines, as not all baselines support it. For partitioned indexes we use $sqrt(|X_0|)$ partitions where $|X_0|$ is the initial number of vectors in the workload. For the graph indexes, we use a graph degree of 128. For LIRE and Quake, we set the partition refinement radius $r = 25$. For Quake we set $\tau = 1$ microsecond and use one iteration of k-means for refinement. All systems use 16 threads for updates and maintenance (if applicable). SCANN, DiskANN, and SVS perform maintenance eagerly during an update, therefore we do not measure maintenance time separately from update time. We consider maintenance after each operation for all methods. Throughout all experiments, indexes are tuned to achieve an average of 90% recall for $k = 100$ across the workloads.

7.3 End-to-End Evaluation

Comparison with Baselines Table 3 shows that Quake consistently achieves lower search and update latency across all workloads. On the Wikipedia-12M workload, where the dataset grows over time and partitions can become unbalanced, Quake-MW maintains a 0.53ms query latency while with a single worker Quake gets 3.28ms. In contrast, Faiss-IVF climbs to 57.4ms due to the lack of maintenance, DeDrift reaches 45.9ms despite its rebalancing efforts, LIRE is unable to meet the recall target even with 17ms latency and SCANN performs similarly with poor update latency due to over-eager maintenance applied during updates. Even the best-performing graph-based method, DiskANN, takes 4.19ms. Thus, Quake-MW is 8× faster than the strongest baseline on this workload derived from real-world access patterns.

On the OpenImages-13M workload, which includes both insertions and deletions, Quake-MW's search latency is 1.01ms. The best competing approach, DiskANN, records 7.99ms, making Quake 8× faster. Faiss-HNSW cannot support deletions, and both SVS's and DiskANN's delete consolidation is expensive, leading to multiple orders of magnitude slower update latency than partitioned indexes, illustrating that graph-based indexes struggle with dynamic operations. Quake and other partitioned index baselines are two orders of magnitude faster in comparison. Quake's continuous maintenance and adaptive parameter selection keep partitions balanced, achieving low latency and stable recall.

For the static, read-only, MSTuring10M-RO workload, Quake's maintenance improves the index structure even without data changes, adapting partitions to the query pattern. For Quake-MW, this yields a search latency of 2.28ms. However, the MSTuring10m dataset is especially challenging for partitioned indexes, as they need to scan roughly 10% of all partitions in order to meet the recall target. In contrast, the well-optimized SVS library exhibits a superior search latency of 1.18ms, demonstrating that in static settings, well-optimized graph indexes are strong competition.

On MSTuring10M-IH, where the dataset grows from one to ten million vectors, Quake-MW achieves an average search latency of 1.94ms. DiskANN, the second-best performer, has a latency of 2.9ms, making Quake 1.5× faster. The other baselines fail to maintain the recall target or suffer from high latency due to their static parameters and inability to prevent partition skew.

Overall, these results demonstrate that Quake's combination of adaptive partition scanning, incremental maintenance, and NUMA-aware parallelism consistently delivers low-latency queries at the desired recall. Systems without maintenance (Faiss-IVF) suffer from skew-induced latency increases, those tied to static search parameters (LIRE) struggle to maintain recall without incurring higher query times,

Method	Wikipedia-12M			OpenImages-13M			MSTuring10M-RO		MSTuring10M-IH		
	Search	Update	Maint.	Search	Update	Maint.	Search	Maint.	Search	Update	Maint.
Quake-MW	**.53 ms**	.004 ms	7.6 s	**1.01 ms**	.006 ms	1.71 s	2.28 ms	2.89 s	**1.94 ms**	.01 ms	.486 s
Quake-SW	3.28 ms	.004 ms	7.6 s	-	-	-	-	-	-	-	-
Faiss-IVF	57.4 ms	.002 ms	-	16.1 ms	.003 ms	-	44.1 ms	-	49.4 ms	.005 ms	-
DeDrift	45.9 ms	.01 ms	34.5 s	8.34 ms	.011 ms	3.43 s	-	-	69.0 ms	.01 ms	1.98 s
LIRE	15.3* ms	.01 ms	6.64 s	5.34 ms	.016 ms	1.90 s	-	-	32.7* ms	.009 ms	0.77 s
ScaNN	17.4 ms	.60 ms	-	14.8 ms	.07 ms	-	10.7 ms	-	24.1 ms	.037 ms	-
Faiss-HNSW	5.07 ms	.06 ms	-	-	-	-	6.83 ms	-	4.57 ms	.55 ms	-
DiskANN	4.19 ms	.11 ms	-	7.99* ms	.5 ms	-	4.17 ms	-	2.9 ms	.19 ms	-
SVS	7.11* ms	.197 ms	-	10.34 ms	.758ms	-	**1.18 ms**	-	7.61* ms	.095 ms	-

Table 3: Workload performance: showing single-query search latency to reach a recall of .9, mean single-update latency, and mean maintenance latency per operation. All methods run with one search thread except the Quake-MW row, which shows the effect of 16-thread intra-query parallelism. *Denotes the config is unable to meet the recall target with static query parameters.

and graph-based methods (Faiss-HNSW, DiskANN) face substantial overheads when handling updates and deletions. By integrating these components, Quake matches the low update cost of partitioned indexes while outperforming graph indexes in search latency in dynamic workloads. Our design represents a significant advancement to the state-of-the-art, providing stable, efficient performance across diverse, and evolving workloads.

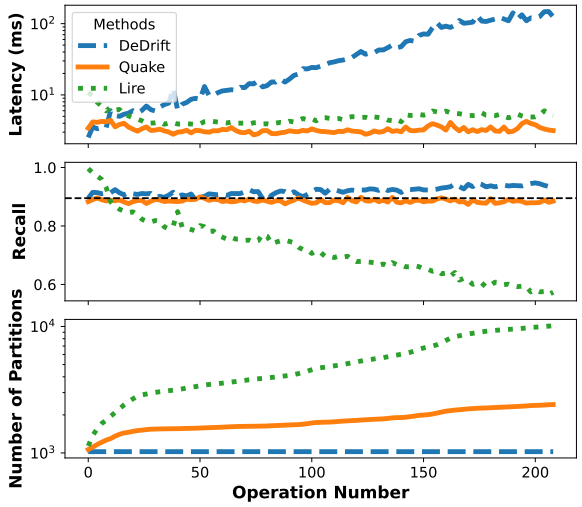


Figure 4: Comparison of single-threaded search latency, recall and number of partitions for Quake vs. maintenance approaches LIRE and DeDrift on Wikipedia-12M. Quake maintains stable latency and recall throughout the workload.

Comparison with Partitioned Index Maintenance Methods Here we perform a detailed comparison with LIRE and DeDrift, measuring the latency, recall, and number of partitions over time on the Wikipedia-12M workload. For a fair comparison, we disable NUMA to highlight the advantages of APS and maintenance in Quake. The results are shown in Figure 4. First looking at recall, we see that Quake maintains a stable recall of near .9, while LIRE's recall degrades over time as it uses a static nprobe. DeDrift's recall stays relatively constant, as it does not adjust the number of partitions and therefore does not need to adjust nprobe. However, when turning our attention to latency, we see that Quake has near-constant stable latency, even as the dataset grows, while DeDrift's latency increases significantly with time. In terms of the number of partitions, we see DeDrift stays constant while Quake and Lire increase by 2.5× and 10× respectively. LIRE uses significantly more partitions because it uses size thresholding to determine when to split, regardless of whether a given partition is hot or not. Quake on the other hand only splits partitions if their contribution to the cost model is high, allowing for more efficient maintenance. These results show that Quake's approach to maintenance is superior to existing methods for partitioned index maintenance in minimizing query latency and recall stability.

7.4 Wikipedia-12M Ablation

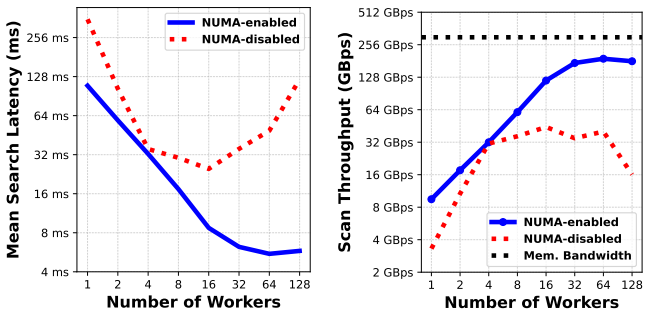
To quantify the contributions of Quake components, we disabled key features and measured the impact on Wikipedia-12M workload in Table 4. We see that disabling APS has little impact on the query latency, as Quake can achieve a low latency even in the static nprobe setting. However, APS provides significantly more recall stability, as evidenced by the increase in standard deviation when APS is disabled. Disabling NUMA parallelism, however, shows a 6× increase in query latency, demonstrating the benefit of parallelization of partition scans. We note that even with NUMA disabled, Quake achieves a lower latency than the best baseline. Finally, we disable maintenance and see a significant increase in latency, similar to the latency of Faiss-IVF; here partitions become extremely imbalanced due to the skew in the workload (see Figure 1a) causing queries to scan more vectors and therefore increasing latency. This further demonstrates the necessity for maintenance for dynamic workloads. In conclusion, each piece of Quake contributes to its performance in terms of both recall stability and minimal query latency.

Table 4: Ablation Study on Wikipedia-12M showing mean search latency and the standard deviation of recall.

Configuration	Search Latency	Recall Std.
Quake (Full)	.53 ms	.008 %
Quake w/o APS	.5 ms	.025 %
Quake w/o NUMA	3.28 ms	.005 %
Quake w/o NUMA/APS	3.18 ms	.025 %
Quake w/o Maint/NUMA/APS	45.2 ms	.014 %

7.5 Scalability

We tested Quake's parallel scalability by varying the number of threads. In Figure 5 we measure the mean search latency and scan throughput (bytes scanned / query latency) on MS-Turing100M to reach a recall of 0.9. Note that this dataset has 100 million vectors and is 10× larger than the datasets we compared against previously. We compare our NUMA-aware parallelism with one in which NUMA is disabled. For both configurations, we see near linear scalability up to around 8 workers, where the non-NUMA latency performs best (28ms). The NUMA configuration however further improves and at 64 workers achieves a latency of 6ms. Looking at the scan throughput, we see that NUMA achieves a peak throughput of 200GBps. We do not completely saturate memory bandwidth due to other overheads involved in query processing (topk sorting, memory allocations, coordination). In conclusion, NUMA-aware intra-query parallelism is an effective mechanism for decreasing query latency by utilizing the full memory capabilities of multicore machines.



(a) Mean Search Latency (b) Scan Throughput

Figure 5: MSTuring100M: Scaling the number of threads with and without NUMA.

7.6 Comparison with Early Termination Methods

Table 5 compares early-termination methods on SIFT1M, highlighting the tradeoff between query latency, tuning time, and recall. APS analytically estimates recall at query time, **eliminating offline tuning entirely**, while achieving latency within 20% of the oracle across all recall targets. FixedNProbe selects a static *nprobe* per target via an expensive offline grid search (up to 424 s), and SPANN similarly performs a grid search and tunes a centroid-distance threshold; both closely match recall targets but incur higher latency at 0.90 and 0.99 recall. LAET trains a per-query prediction model, incurring moderate tuning overhead (81–232 s), and matches recall targets with slightly higher latency compared to APS. Auncel is the most similar method to APS, as it aims to analytically estimate recall using partition intersection volumes, however it's volume estimation requires calibration, and it is a conservative method, overshooting recall. We tune Auncel by binary searching a geometric parameter (a), overshooting recall significantly (up to 8.1 pp) and increasing latency by up to 69% compared to APS. Finally, the Oracle, which scans the minimal amount of partitions per-query, serves as a practical lower bound on achievable latency, though with prohibitively high tuning cost. APS thus provides near-optimal performance without tuning overhead, matching or exceeding all baselines.

Table 5: Early-termination methods on SIFT1M with a partitioned index with 1000 partitions. Each row shows the average recall, nprobe, and mean per-query latency in milliseconds over 10000 queries after tuning for a specific recall target for $k = 100$. We also report the total tuning time in seconds, where APS needs *no* offline tuning.

Method	Target	Recall	*nprobe*	Latency	Tuning
APS	0.80	0.821	11.8	0.34ms	**0**
	0.90	0.912	20.2	0.48ms	**0**
	0.99	0.989	50.1	0.96ms	**0**
Auncel [42]	0.80	0.857	16.4	0.41ms	66.3s
	0.90	0.981	73.8	1.29ms	73.8s
	0.99	0.997	95.9	1.61ms	83.2s
SPANN [7]	0.80	0.816	11	0.31ms	173s
	0.90	0.902	19	0.43ms	183s
	0.99	0.990	70	1.07ms	259s
LAET [17]	0.80	0.813	10.5	0.29ms	81s
	0.90	0.905	18.2	0.42ms	104s
	0.99	0.990	58.3	1.03ms	232s
FixedNProbe	0.80	0.817	11	0.33ms	318s
	0.90	0.903	19	0.44ms	330s
	0.99	0.990	65	1.16ms	424s
Oracle	0.80	0.833	11.5	0.29ms	320s
	0.90	0.924	19.3	0.41ms	331s
	0.99	0.992	42.0	0.74ms	368s

7.7 Maintenance Ablation

To understand the effectiveness of the primary components of adaptive incremental maintenance (cost-model, partition refinement, and rejection), we replay a dynamic SIFT1M trace (30% inserts, 20% deletes, 50% queries) with different components disabled. All methods use a single-thread and search using APS with k=100 and a 0.90 recall target. We also in-

clude LIRE as a baseline. Table 6 reports cumulative times in seconds. For configurations with refinement we use a refinement radius of $r_f = 50$. The full Quake policy delivers the lowest search cost (86 s) while meeting the recall target. If we keep the cost model but skip refinement (NoRef), maintenance time decreases significantly from 21 s to 5 s, yet recall slips by 2.4 pp and the search time increase by 15.4s. This shows that while refinement is the dominant cost in maintenance, it is necessary for minimizing search latency. Disabling the cost model and instead using size-based thresholding (NoCost) shows why naïve size thresholds are inadequate: search time rises 8 % despite similar maintenance effort. The rejection mechanism is critical; once removed (NoRej), recall collapses to 0.66 even though search and maintenance appear cheap. LIRE, which relies solely on size thresholding, is 17 % slower in search latency, confirming that the cost model, rejection mechanism, and partition refinement are essential for maintaining both index performance and quality.

Table 6: Maintenance ablation on the SIFT1M workload. Times are cumulative (in seconds) over the course of the workload. Recall is averaged over all queries.

Maintenance Variant	Search	Update	Maint.	Recall
Quake (Full)	86.3s	21.7s	21.4s	0.905
NoRef	101.7s	22.2s	5.2s	0.881
NoRef+NoRej	85.5s	21.0s	1.0s	0.730
NoRej	84.2s	19.6s	18.5s	0.662
NoCost	93.5s	20.0s	20.4s	0.901
NoCost+NoRef	100.7s	21.2s	0.8s	0.879
LIRE	100.5s	21.2s	11.9s	0.900

8 Conclusion

Experimental results show that Quake reduces query latency compared to baseline approaches under dynamic and skewed workloads, without requiring manual tuning. It achieves high recall, matching the performance of an oracle for setting the query parameter nprobe. Compared to existing partitioned indexes like Faiss and SCANN, Quake reduces query latency by A) adaptively maintaining index partitions and B) maximizing memory bandwidth during query processing. Compared to graph indexes like SVS, HNSW, and DiskANN, Quake offers more efficient indexing and updates while matching or reducing query latency. In summary, our evaluation shows Quake minimizes query latency while meeting recall targets on dynamic workloads with skewed access patterns.

References

[1] Qdrant - Vector Database. https://qdrant.tech/.

[2] Billion-scale approximate nearest neighbor search challenge: Neurips'21 competition track. https://big-ann-benchmarks.com/, 2021.

[3] Vector database for vector search | pinecone. https://www.pinecone.io, 2024. Accessed on December 4, 2023.

[4] Wikipedia:pageview statistics. https://en.wikipedia.org/wiki/Wikipedia:Pageview_statistics, 2024.

[5] Cecilia Aguerrebere, Mark Hildebrand, Ishwar Singh Bhati, Theodore Willke, and Mariano Tepper. Locally-adaptive quantization for streaming vector search. *arXiv preprint arXiv:2402.02044*, 2024.

[6] Dmitry Baranchuk, Matthijs Douze, Yash Upadhyay, and I. Zeki Yalniz. DeDrift: Robust Similarity Search under Content Drift, August 2023. arXiv:2308.02752 [cs].

[7] Qi Chen, Bing Zhao, Haidong Wang, Mingqin Li, Chuanjie Liu, Zengzhong Li, Mao Yang, and Jingdong Wang. SPANN: Highly-efficient Billion-scale Approximate Nearest Neighbor Search.

[8] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2024.

[9] Mihajlo Grbovic and Haibin Cheng. Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 311–320, 2018.

[10] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, 2020.

[11] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3887–3896. PMLR, November 2020. ISSN: 2640-3498.

[12] Malay Haldar, Mustafa Abdool, Prashant Ramanathan, Tao Xu, Shulin Yang, Huizhong Duan, Qing Zhang, Nick Barrow-Williams, Bradley C Turnbull, Brendan M Collins, et al. Applying deep learning to airbnb search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1927–1935, 2019.

[13] Helia Hashemi, Aasish Pappu, Mi Tian, Praveen Chandar, Mounia Lalmas, and Benjamin Carterette. Neural instant search for music and podcast. In *Proceedings*

of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 2984–2992, 2021.

[14] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

[15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, March 2020.

[16] Viktor Leis, Peter Boncz, Alfons Kemper, and Thomas Neumann. Morsel-driven parallelism: a numa-aware query evaluation framework for the many-core age. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, page 743–754, New York, NY, USA, 2014. Association for Computing Machinery.

[17] Conglong Li, Minjia Zhang, David G Andersen, and Yuxiong He. Improving approximate nearest neighbor search through learned adaptive early termination. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2539–2554, 2020.

[18] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2010.

[19] LibTorch: PyTorch C++ API. https://pytorch.org/cppdocs.

[20] David C Liu, Stephanie Rogers, Raymond Shiau, Dmitry Kislyuk, Kevin C Ma, Zhigang Zhong, Jenny Liu, and Yushi Jing. Related pins at pinterest: The evolution of a real-world recommender system. In *Proceedings of the 26th international conference on world wide web companion*, pages 583–592, 2017.

[21] Zhuoran Liu, Leqi Zou, Xuan Zou, Caihua Wang, Biao Zhang, Da Tang, Bolin Zhu, Yijie Zhu, Peng Wu, Ke Wang, and Youlong Cheng. Monolith: Real time recommendation system with collisionless embedding table. In *5th Workshop on Online Recommender Systems and User Modeling (ORSUM2022), in conjunction with the 16th ACM Conference on Recommender Systems*, 2022.

[22] Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836, April 2020.

[23] Jason Mohoney, Roger Waleffe, Henry Xu, Theodoros Rekatsinas, and Shivaram Venkataraman. Marius: Learning massive graph embeddings on a single machine. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pages 533–549, 2021.

[24] moodycamel::ConcurrentQueue. https://github.com/cameron314/concurrentqueue.

[25] SimSIMD. https://github.com/ashvardanian/SimSIMD.

[26] Jiongkang Ni, Xiaoliang Xu, Yuxiang Wang, Can Li, Jiajie Yao, Shihai Xiao, and Xuecang Zhang. DiskANN++: Efficient Page-based Search over Isomorphic Mapped Graph Index using Query-sensitivity Entry Vertex, November 2023. arXiv:2310.00402 [cs].

[27] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1933–1942, 2017.

[28] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. Pinnersage: Multi-modal user embedding framework for recommendations at pinterest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2311–2320, 2020.

[29] Iraklis Psaroudakis, Tobias Scheuer, Norman May, Abdelkader Sellami, and Anastasia Ailamaki. Adaptive numa-aware data placement and task scheduling for analytical workloads in main-memory column-stores. *Proc. VLDB Endow.*, 10(2):37–48, October 2016.

[30] An Qin, Mengbai Xiao, Yongwei Wu, Xinjie Huang, and Xiaodong Zhang. Mixer: efficiently understanding and retrieving visual content at web-scale. *Proceedings of the VLDB Endowment*, 14(12):2906–2917, 2021.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[32] Aditi Singh, Suhas Jayaram Subramanya, Ravishankar Krishnaswamy, and Harsha Vardhan Simhadri. Freshdiskann: A fast and accurate graph-based ann index for streaming similarity search. *arXiv preprint arXiv:2105.09613*, 2021.

[33] Suhas Jayaram Subramanya, Devvrit, Rohan Kadekodi, Ravishankar Krishaswamy, and Harsha Vardhan Simhadri. *DiskANN: fast accurate billion-point nearest neighbor search on a single node*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[34] Philip Sun, David Simcha, Dave Dopson, Ruiqi Guo, and Sanjiv Kumar. Soar: Improved indexing for approximate nearest neighbor search. In *Neural Information Processing Systems*, 2023.

[35] Roger Waleffe, Jason Mohoney, Theodoros Rekatsinas, and Shivaram Venkataraman. Mariusgnn: Resource-efficient out-of-core training of graph neural networks. In *ACM SIGOPS European Conference on Computer Systems (EuroSys)*, 2023.

[36] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627, 2021.

[37] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 839–848, 2018.

[38] Chuangxian Wei, Bin Wu, Sheng Wang, Renjie Lou, Chaoqun Zhan, Feifei Li, and Yuanzhe Cai. Analyticdb-v: A hybrid analytical engine towards query fusion for structured and unstructured data. *Proceedings of the VLDB Endowment*, 13(12):3152–3165, 2020.

[39] Yuming Xu, Hengyu Liang, Jin Li, Shuotao Xu, Qi Chen, Qianxi Zhang, Cheng Li, Ziyue Yang, Fan Yang, Yuqing Yang, Peng Cheng, and Mao Yang. SPFresh: Incremental In-Place Update for Billion-Scale Vector Search. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 545–561, New York, NY, USA, October 2023. Association for Computing Machinery.

[40] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[41] Qianxi Zhang, Shuotao Xu, Qi Chen, Guoxin Sui, Jiadong Xie, Zhizhen Cai, Yaoqi Chen, Yinxuan He, Yuqing Yang, Fan Yang, et al. {VBASE}: Unifying online vector similarity search and relational queries via relaxed monotonicity. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 377–395, 2023.

[42] Zili Zhang, Chao Jin, Linpeng Tang, Xuanzhe Liu, and Xin Jin. Fast, approximate vector queries on very large unstructured datasets. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 995–1011, 2023.