Results of the Big ANN: NeurIPS'23 Competition

Harsha Vardhan Simhadri (Microsoft) Matthijs Douze (Meta AI) Dmitry Baranchuk (Yandex) Ben Landrum (UMD) Martin Aumüller (ITU) George Williams (Ansible AI) Edo Liberty (Pinecone) Mazin Karjikar (UMD) Frank Liu (Zilliz) Laxman Dhulipala (UMD)

Participating Teams: Shanghai Jiao Tong University, Fudan University, Baidu, and others.

◎ Introduction

Approximate Nearest Neighbor (ANN) search is critical for LLMs (RAG), computer vision, and recommendation systems. While previous challenges focused on scaling standard dense vector indexing, the 2023 Big ANN Challenge[1] addressed **practical, complex variants** of ANN search encountered in real-world applications.

Key Goals

- Move beyond standard dense indexing.
- Address diverse data distributions and types.
- Evaluate on constrained hardware (16GB RAM).
- Promote open-source contributions.

Tracks & Datasets

1. Filtered Search

Dataset: YFCC 100M[2] (CLIP embeddings + Tags)

Search for nearest neighbors that *also* match specific metadata tags (e.g., "camera model", "country").

Query: Image_Emb + ["freight", "country_GB"]

2. Out-Of-Distribution (OOD)

Dataset: Yandex Text-to-Image 10M [4]

Database (Images) and Query (Text) vectors have different distributions in the shared space. Standard indices often fail to provide high recall.

3. Sparse Search

Dataset: MSMARCO[3] (SPLADE model)

High-dimensional vectors (>30k dim) with few non-zero elements (~120). Optimized for inverted indices and specialized graphs.

4. Streaming

Dataset: MS Turing[4] (10M subset)

Indices must handle a "runbook" of Insertions, Deletions, and Searches under strict memory (8GB) and time limits.

Evaluation Protocol

Hardware	Metric
Azure D8lds_v5 8 vCPUs, 16GB RAM	10-recall@10 & QPS (Throughput)

"Ranking based on highest throughput (QPS) achieving at least 90% recall."

Results & Winners

Filtered Track — Winner: ParlayANN

- **ParlayANN** achieved 11x baseline speed.
- **Strategy:** Inverted index for tags + Vamana graph for dense vectors.
- Used efficient bit-vector intersections for low-cardinality tags.

OOD Track — Winners: RoarANN & PyANNS

- **Challenge:** Database query distribution mismatch.
- **RoarANN (MysteryANN):** Used a bipartite graph between base and query samples to adapt the index structure.
- **PyANNS:** Relied on highly optimized Vamana graph + quantization (VNNI instructions).

Sparse Track — Winners: PyANNS & GrassRMA

- **Baseline:** Linscan (inverted index).
- **Winning Approach:** Graph-based indices outperformed inverted indices.
- **PyANNS:** Quantized HNSW. Refinement step used full vectors to recover accuracy.
- **GrassRMA:** Optimized memory access patterns for sparse data in graphs.

Streaming Track — Winner: PyANNS

Scenario: 4:4:1 ratio of Insert:Delete:Search.

- **Winner Strategy:** DiskANN[4] with 8-bit scalar quantization.
- Quantization allowed deeper graph search within the time limit, maintaining high recall despite deletions.

***Note: Original competition results were corrected post-event due to a caching error in recall calculation.**

Discussion & Future Work

The 2023 challenge demonstrated a significant leap in performance over industry baselines, driven by specialized data structures rather than raw scaling.

Key Trends

- **Graph Indices** are dominant, even for sparse/OOD.
- **Quantization** is essential for efficiency on constrained hardware.
- **Hybrid Indices** (Graph + Inverted) needed for Filtered search.

Impact

- Established benchmarks for "ragged" real-world data.
- Highlighted need for "deletions" support in streaming indices.

References:

[1] Simhadri et al. "Results of the NeurIPS'21 Challenge on Billion-Sclae Approximate Nearest Neighbor Search." NeurIPS Competition and Demos, PMLR, 2021.

[2] Thomee et al. "YFCC100M: The new data in multimedia research" Comm. ACM, 2016.

[3] Nguyen et al. "MS MARCO: A Human Generated Machine Reading Comprehension Dataset.", 2016.

[4] Simhadri et al. "DiskANN: Graph-structured Indices for Scalable, Fast, Fresh and Filtered ANN Search", https://github.com/microsoft/DiskANN, 2023.