Efficient Serving of Large-scale Vector Search with Sharded Indexes

Zhenning Yang, Jiaheng Lu, Jeremy Flics
EECS 598 Systems for GenAI
UNIVERSITY OF MICHIGAN

Introduction

Motivation

- LLMs are great, BUT:
  - Hallucination: False plausible-sounding texts
  - Outdated training data: Non-step fine-tuning
- RAG connects LLMs to **external knowledge base** (**VectorDB**) to reduce hallucination.
- The knowledge base will continue to expand!

Vector Similarity Search

- KNN (K-Nearest Neighbors): Flat index
- ANN (Approximate NN) index: HNSW and IVF
- Memory intensive and hard to scale!

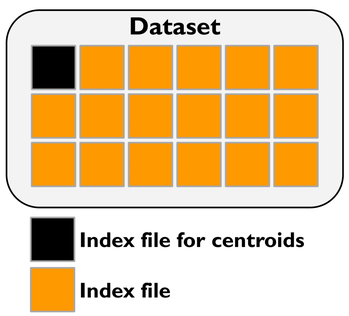
Opportunities

- Sharding and multi-tier memory systems.

Solutions

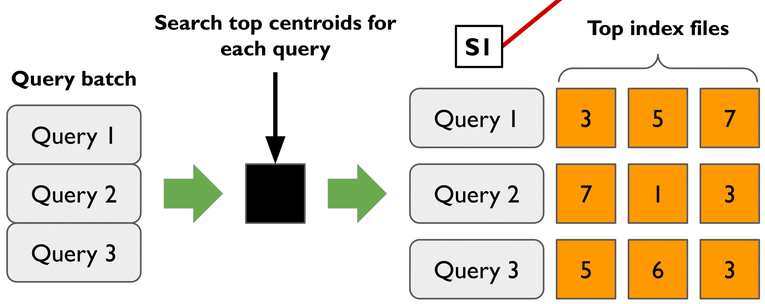
- Efficient batching
- Index placement

Vector Search Sharded Indexes

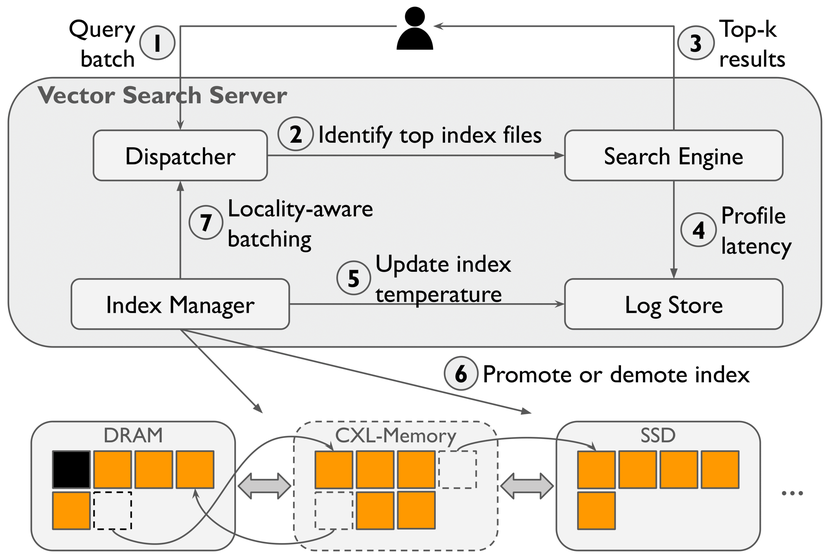


Index Large Datasets

1. Partition dataset into smaller clusters
2. Index clusters
3. Index centroids



Solution



Intra-batch Optimization

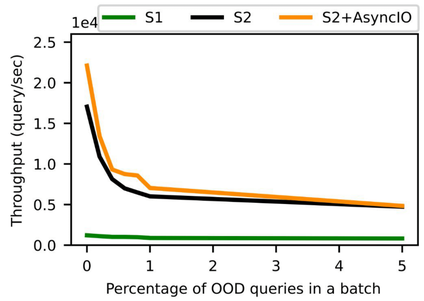
- Batch queries with shared-index
- Asynchronous IO
- Re-order index search-plan, prioritize large query batch

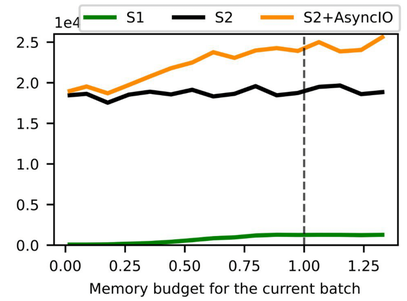
Inter-batch Optimization

- Promote or demote index files based on their temperature
  - "LRU" and "LFU x batch-size"
- Adjust search-plan, prioritize search on index files already in DRAM

Evaluation

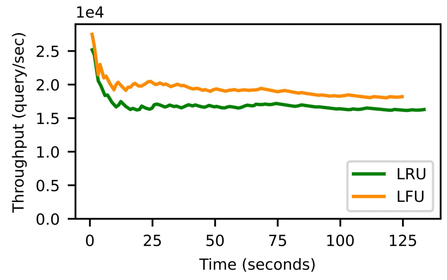
Intra-batch Throughput





Inter-batch Throughput

- S2 + AsyncIO
- Processed 100 requests
  - 10,000 queries per batch
  - Random % OOD queries

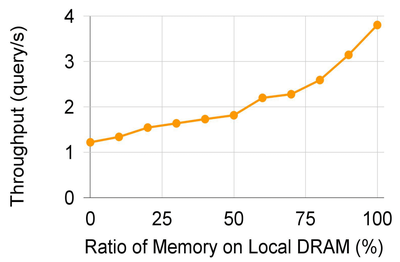


Conclusion

- Developed a prototype system for vector search in large datasets with sharded indexes.
- Optimized intra, inter-batch throughput through asyncIO and dynamic index file management.
- Evaluated system on different workloads under various constraints.

Future Work

- Experiment with RAG-LLMs and other applications.
- Improve QoS aware index query with CXL, already created monitor for B/W and latency.



References

- Douze, Matthijs, et al. "The faiss library." ArXiv, 2024
- Borgeaud, Sebastian, et al. "Improving Language Models by Retrieving from Trillions of Tokens." ArXiv, 2021
- Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs." IEEE Transactions on Big Data, 2019