Flash is Driving Scale in RAG-Based LLMs

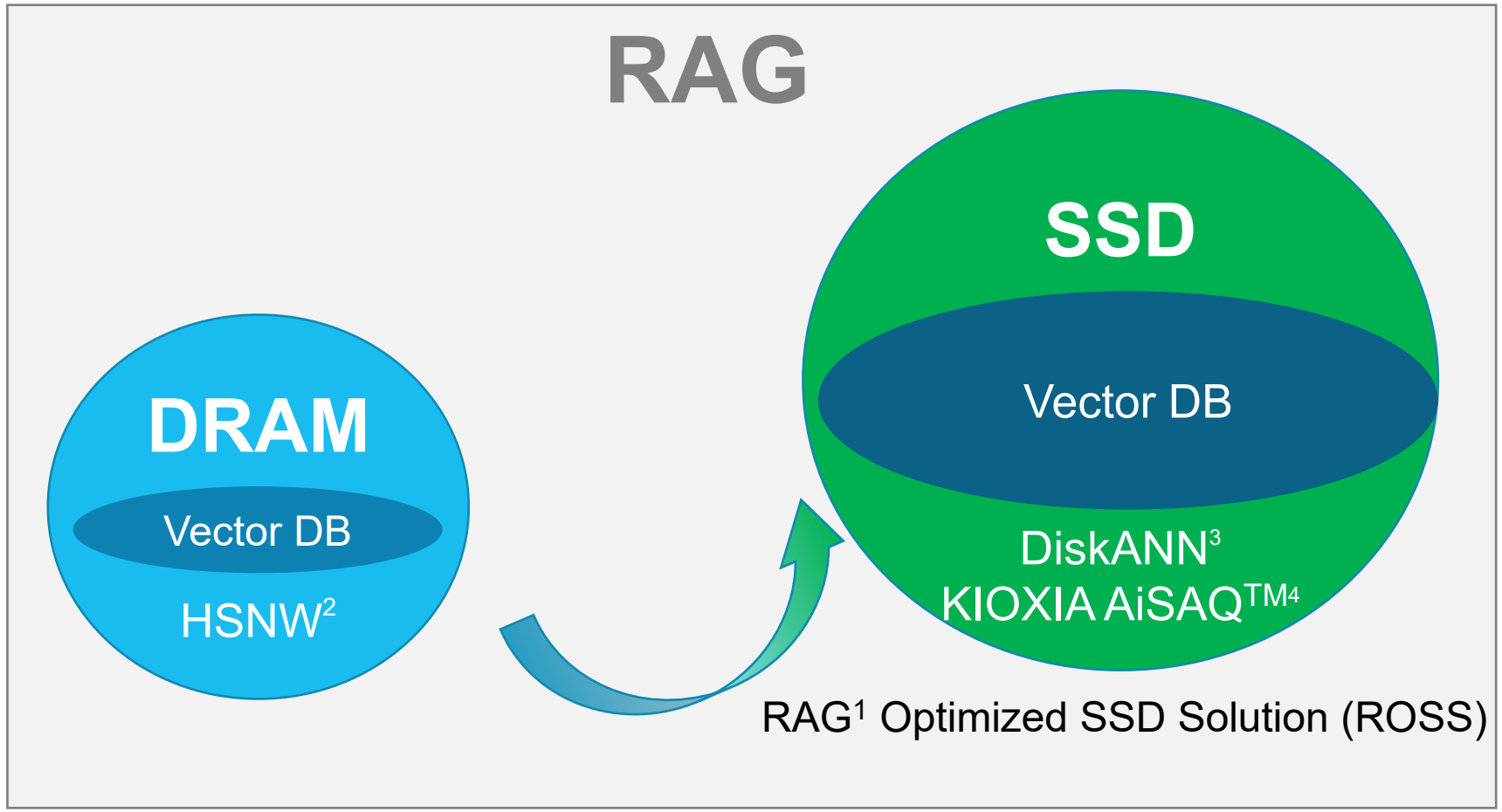
Assaf Sella

Vice President, Machine Learning R&D

KIOXIA

RAG Optimized SSD Solution (ROSS)

A concept that KIOXIA proposes: elevated utilization of SSD in RAG



ROSS Key Takeaway

- Moving vector DB from DRAM to SSD
- SSD based ANNS secures comparable performance to DRAM based solutions
- Scalable: size of vector DB is not dictated by the size of DRAM

Image source: KIOXIA machine learning, R&D engineering team

KIOXIA

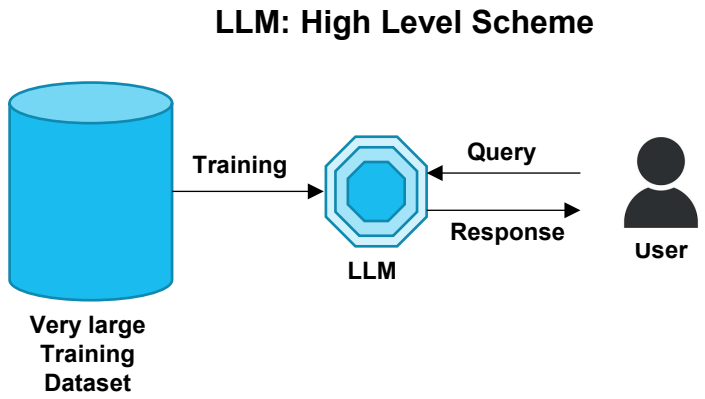
[1] Retrieval Augmented Generation (RAG) [2] Hierarchical Navigable Small World (HNSW)
[3] Approximate Nearest Neighbor (ANN) [4] All-in-Storage ANNS with Product Quantization (KIOXIA AiSAQ)

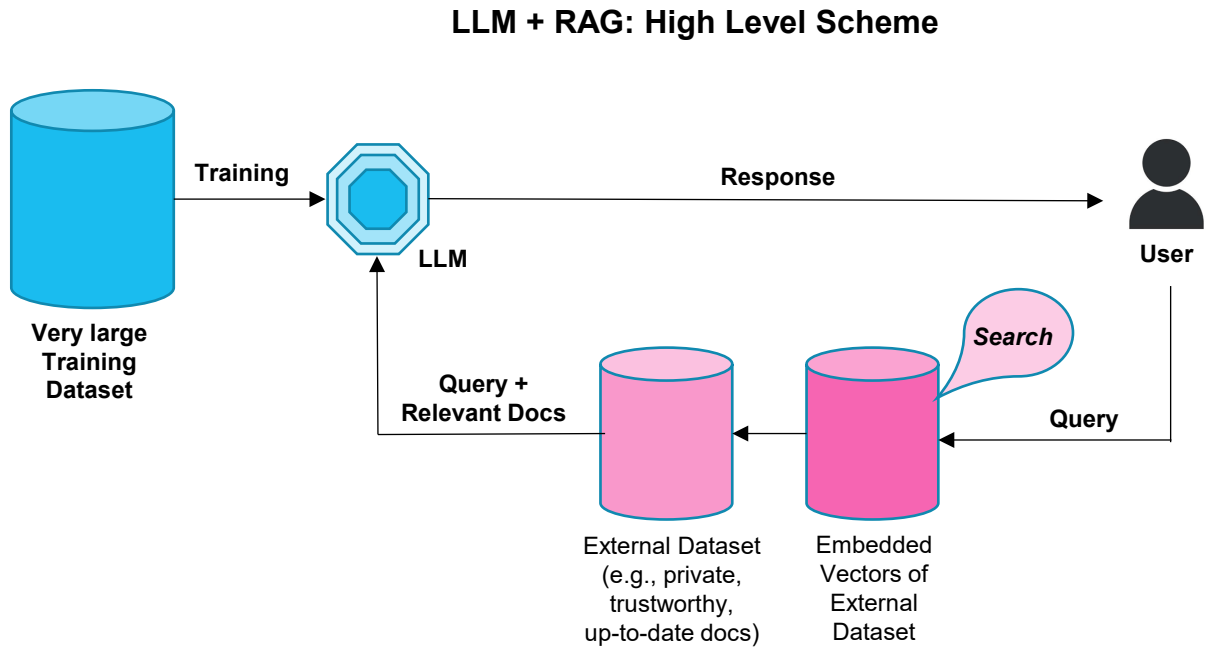
What is RAG (Retrieval Augmented Generation)?

- Large Language Models (LLMs) are one of the most important innovations in recent years, revolutionizing applications such as virtual assistants, chatbots, and dialogue systems

- However, LLMs have encountered significant challenges, including the issue of hallucinations, where they generate false or misleading information that lacks grounding in reality

- RAG enables LLMs to draw upon external knowledge sources to verify and ground their outputs





Images source: KIOXIA machine learning, R&D engineering team. Icon usage rights, no restrictions.

Effective ANNS is Key for RAG

- For RAG to be effective, it has to quickly retrieve the information elements most relevant to the query
- Approximate Nearest Neighbor Search (ANNS) algorithms are used
- ANNS provides efficient search while maintaining high recall

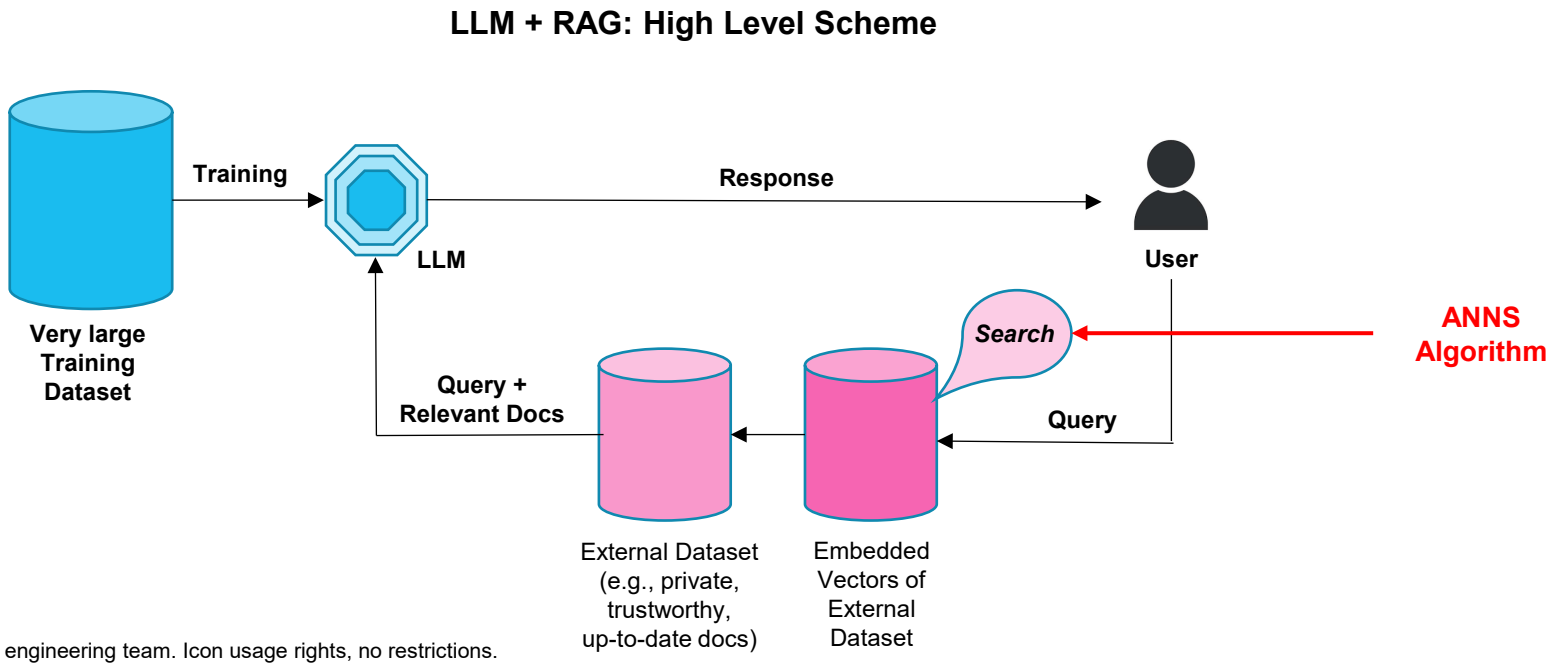
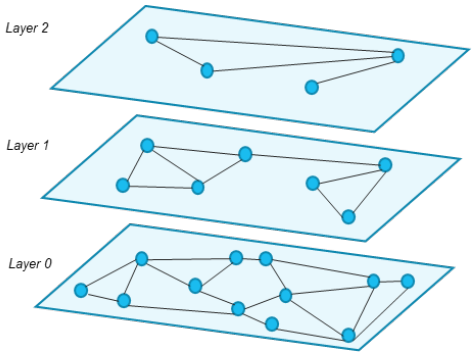


Image source: KIOXIA machine learning, R&D engineering team. Icon usage rights, no restrictions.

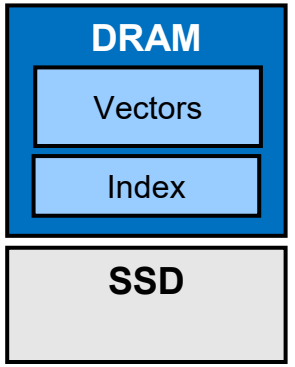
In-Memory ANNS Algorithms Can't Scale

- HNSW[1] is the leading in-memory ANNS algorithm
- Both vectors and index are stored in memory
- Example: 1B vectors RAG dataset with 768 dimensions requires over 1 tebibyte (TiB) DRAM*
- Scaling issue is specifically apparent in RAG applications
  - High dimensionality of the vectors embedding (768D - 1536D)
- High cost of DRAM limits the size of RAG dataset, and thus limits its grounding effectiveness

HNSW's Multi-Layer Graph-Based ANNS



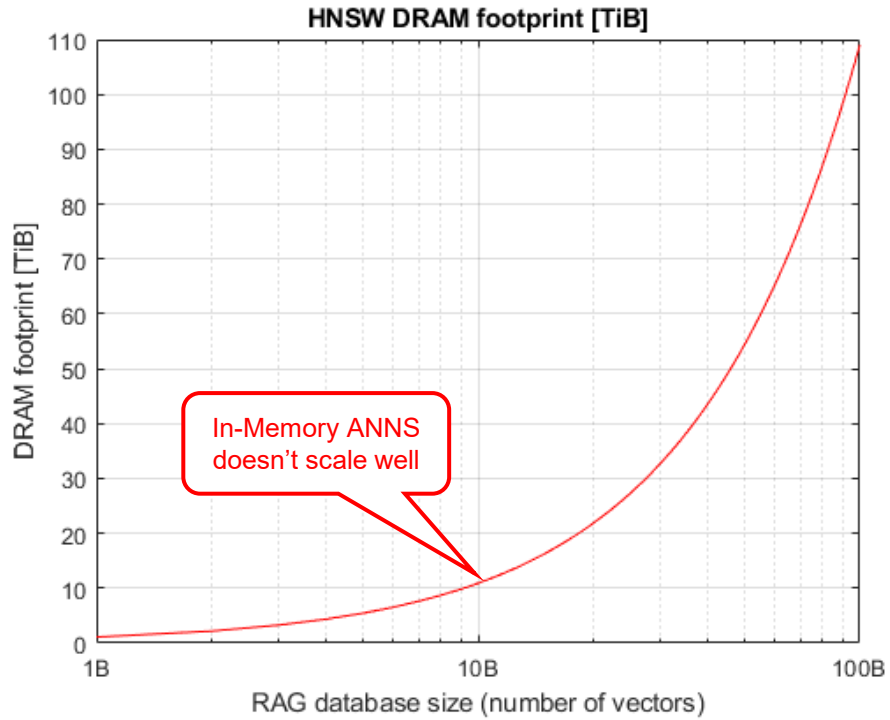
HNSW In-Memory ANNS



1B RAG (1.09TiB)

715 GiB

402 GiB



* Using PQ (Product Quantization) with 1B / dimension

[1] Hierarchical Navigable Small World (HNSW)

Images source: KIOXIA machine learning, R&D engineering team

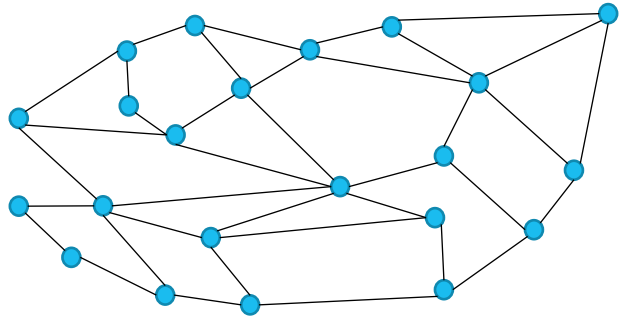
KIOXIA

One tebibyte equals 2^{40} or 1,099,511,627,776 bytes. KIOXIA defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 2^30 = 1,073,741,824 bytes.

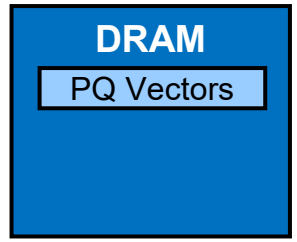
SSD-Based ANNS Algorithms

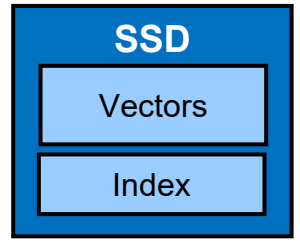
- Microsoft® DiskANN[1] is the leading SSD-based ANNS algorithm
- Index is stored in SSD, quantized vectors (PQ) are stored in memory
- DiskANN is optimized for SSD implementation
  - Optimized algorithm (re-ranking, Vamana graph), and SSD access (beamwith)
- To date, DiskANN was demonstrated only on large vector datasets with **low dimensionality** vectors (128 dimensions)

DiskANN Graph-Based ANNS



DiskANN
Hybrid In-Memory and SSD ANNS





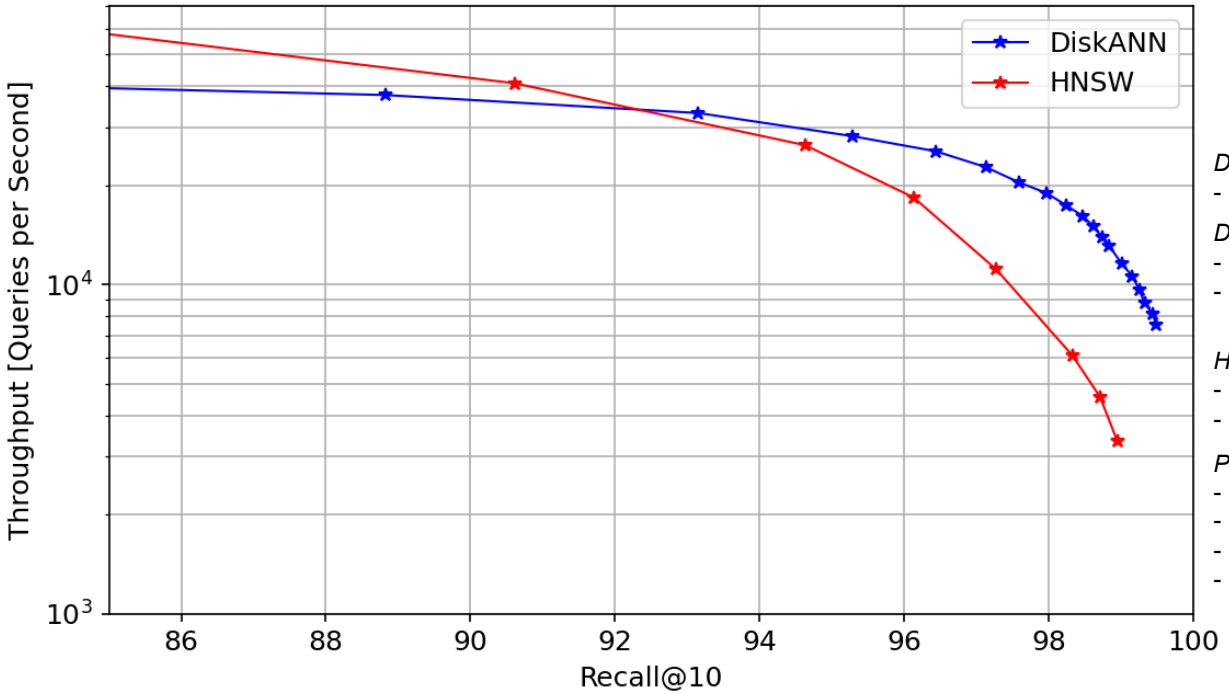
Images source: KIOXIA machine learning, R&D engineering team

[1] The DiskANN repository requests the following citation: @misc{diskann-github; authors = Simhadri, Harsha Vardhan and Krishnaswamy, Ravishankar and Srinivasa, Gopal and Subramanya, Suhas Jayaram and Antonijevic, Andrija and Pryce, Dax and Kaczynski, David and Williams, Shane and Gollapudi, Siddarth and Sivashankar, Varun and Karia, Neel and Singh, Aditi and Jaiswal, Shikhar and Mahapatro, Neelam and Adams, Philip and Tower, Bryan and Patel, Yash; title = DiskANN: Graph-structured Indices for Scalable, Fast, Fresh and Filtered Approximate Nearest Neighbor Search; urls = https://github.com/Microsoft/DiskANN}; version = 0.6.1; year = 2023;

In-Memory vs. SSD-based ANNS Performance

- Today KIOXIA is demonstrating DiskANN on large scale, high dimensionality RAG dataset
- DiskANN benchmarked with KIOXIA CD8P Series PCIe® NVMe™ Data Center SSD
- SSD-based DiskANN provides comparable performance to in-memory HNSW

DiskANN vs. HNSW on 50M RAG Dataset (768 Dimensions)



Dataset:
- *WikiAll 50M vectors, 768 dimensions/vector (3KiB/vector)*

DiskANN:
- *Microsoft® DiskANN run under ann-benchmark*
- *Parameters: PQ=128B/vector, R=64, Lbuild=200, LSearch=10,…,250, beamwidth=4, num_nodes_to_cache=0*

HNSW:
- *Based on Facebook™ AI Similarity Search (FAISS) HNSW PQ Ver 1.8.0*
- *Parameters: PQ=768B/vector, M=48, efConst=200, efSearch=10,…,800*

Platform:
- *Supermicro®, AS-2125HS-TNR*
- *Dual processor AMD EPYC™ 9534 (2 x 64 Cores @ 2.45GHz)*
- *DRAM: 768GB, 24 x 32GB DDR5 4800MT/s DIMM*
- *Storage: 2 x CD8P-R 15.36TB drives, RAID 0*

KIOXIA CD8P
Gen-5 Data Center SSD

Image source: KIOXIA machine learning, R&D engineering team

SSDs are Driving Scale in RAG-Based LLMs

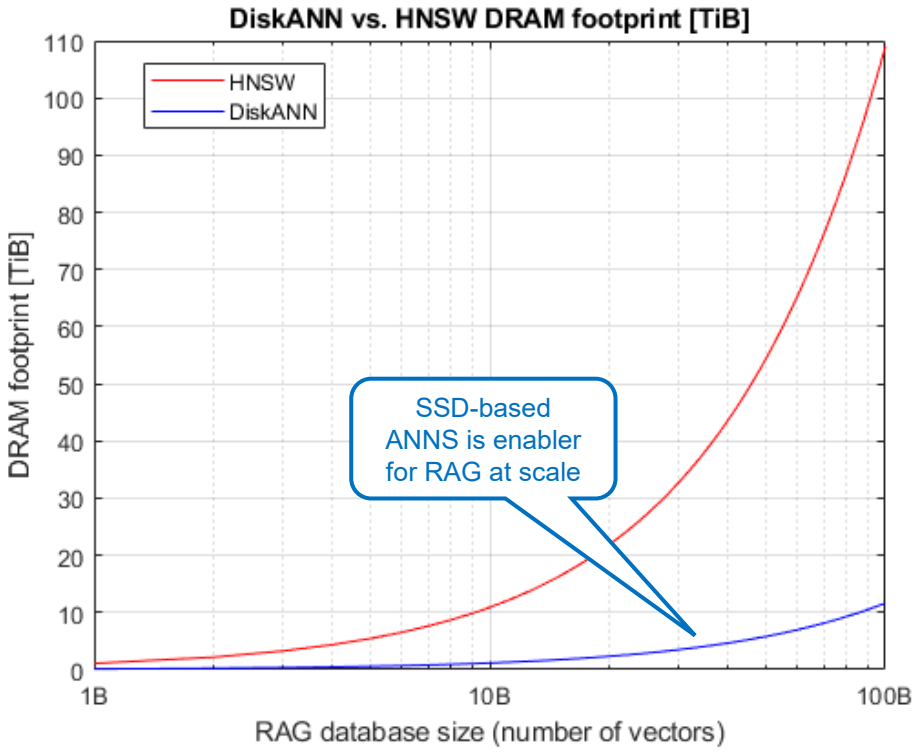
- DiskANN reduces HNSW DRAM footprint by 89% (!)
- SSD-based ANNS enable significant reduction in cost while providing comparable performance to the leading in-memory ANNS
- SSDs serve as enabler for RAG applications at very large scale

DiskANN vs. HNSW: DRAM Footprint Comparison

Vector Database	Vector DB Elements	HNSW PQ	DiskANN
RAG 50M/768D	Vectors (full precision)	143GiB	143GiB
	PQ Vectors (PQ=768B/128B)	35.8GiB	6GiB
	Index* (M=48/R=64)	20.1GiB	11.9GiB
	DRAM footprint	**55.9GiB**	**6GiB**
RAG 1B/768D	Vectors (full precision)	2.8TiB	2.8TiB
	PQ Vectors (PQ=768B/128B)	715.3GiB	119.2GiB
	Index (M=48/R=64)	402.3GiB	238.4GiB
	DRAM footprint	**1.09TiB**	**119.2GiB**
RAG 10B/768D	Vectors (full precision)	27.9TiB	27.9TiB
	PQ Vectors (PQ=768B/128B)	7.0TiB	1.2TiB
	Index (M=48/R=64)	3.9TiB	2.3TiB
	DRAM footprint	**10.9TiB**	**1.2TiB**

89% reduction in DRAM footprint

DRAM ANNS not scalable



* HNSW index size calculation based on 9M bytes per vector

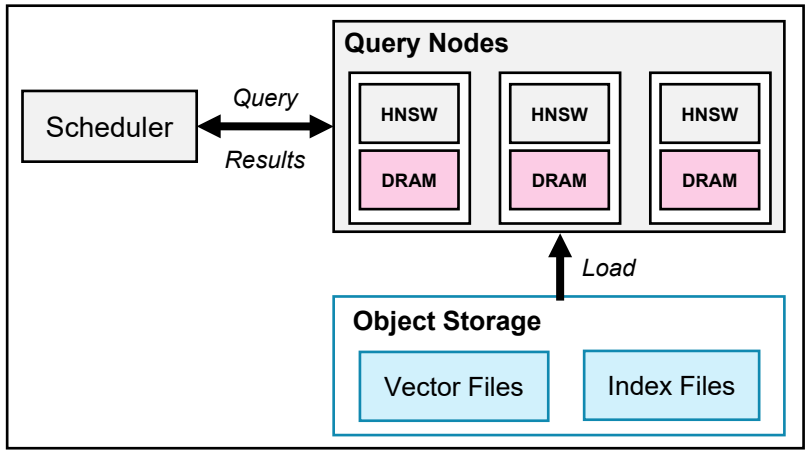
Images source: KIOXIA machine learning, R&D engineering team

One GiB as 1024^3 (1,073,741,824) bytes. One tebibyte equals 2^{40} or 1,099,511,627,776 bytes. KIOXIA defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 2^30 = 1,073,741,824 bytes.

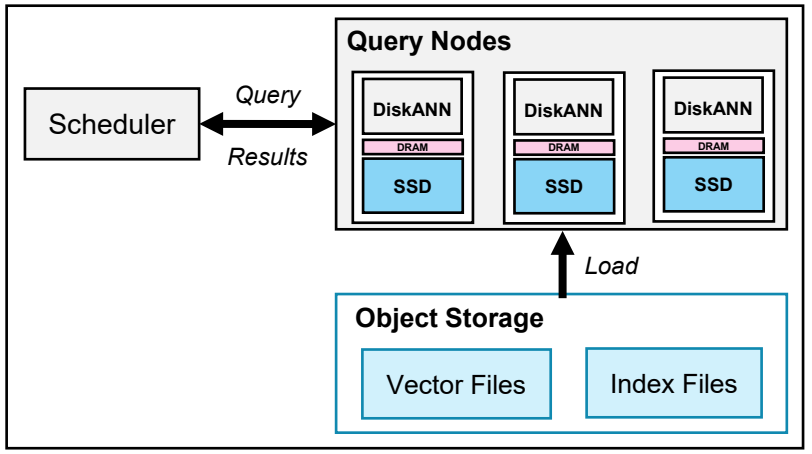
SSDs are Driving Scale in RAG-Based LLMs: Vector DB Architecture

- Vector databases load vector and index data from object storage to query nodes
- Query nodes use SSD as the search media for large scale RAG applications
- With DiskANN same architecture achieves comparable throughput to HNSW with significant reduction in query nodes' cost

Simplified Vector Database Architecture for Small Scale RAG



Simplified Vector Database Architecture for Large Scale RAG

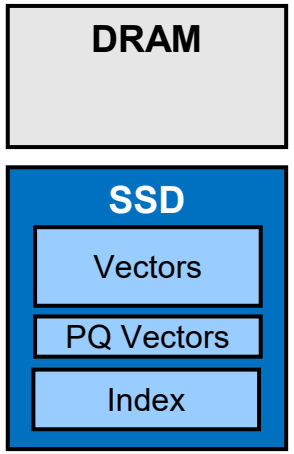


Images source: KIOXIA machine learning, R&D engineering team

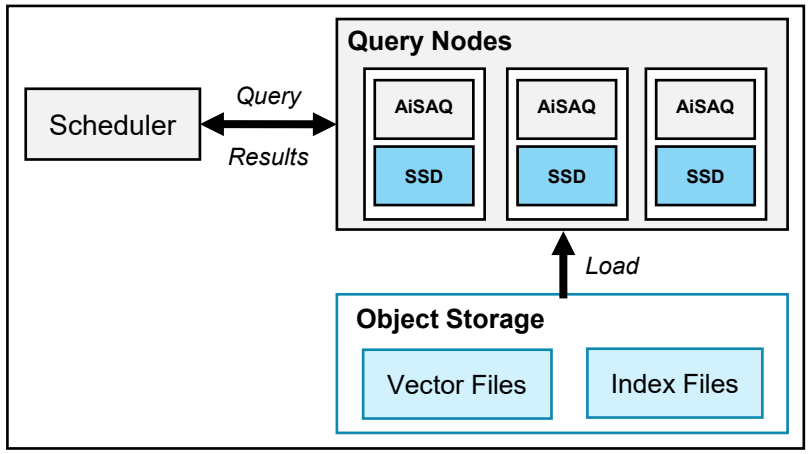
SSDs are Driving Scale in RAG-Based LLMs Further …

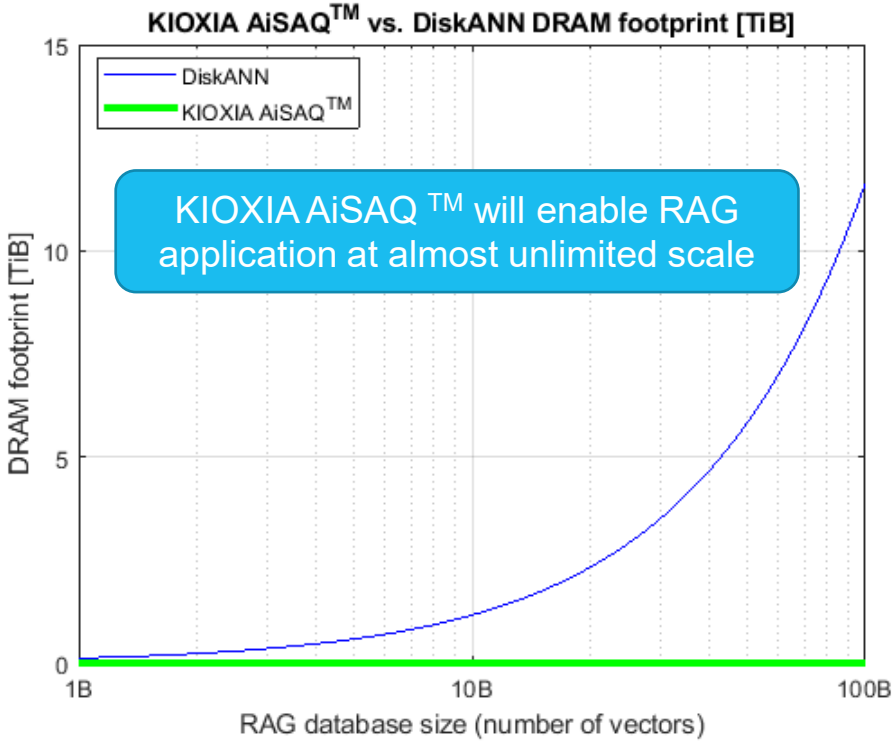
- KIOXIA promotes ROSS (RAG Optimized SSD Solution) to further expand scale with disk-based ANNS solutions
- KIOXIA AiSAQ™ (All-in-Storage ANNS with Product Quantization[1]) is the first KIOXIA development effort for ROSS
- KIOXIA AiSAQ™ enables transition from “most in storage” (e.g. DiskANN) to all-in-storage, almost zero DRAM architectures

KIOXIA AiSAQ
All-in-Storage ANNS



Simplified Vector Database Architecture
with All-in-Storage ANNS





[1] KIOXIA AiSAQ: All-in-Storage ANNS with Product Quantization, a novel method of index data placement, is a trademark of KIOXIA.

Images source: KIOXIA machine learning, R&D engineering team

Summary

- RAG and ANNS are a key components in modern LLM solutions
- In-memory ANNS can’t scale economically and limits RAG size and grounding effectiveness
- SSD-based ANNS provides comparable performance to the leading in-memory ANNS with significantly lower cost, **enabling RAG applications at very large scale**
- SSD-based ANNS solution is available as Microsoft® open-source DiskANN, and can be seamlessly integrated in existing vector DB architectures
- KIOXIA continues to develop disk-based ANNS solutions
- KIOXIA AiSAQ™ will enable RAG applications at almost unlimited scale

KIOXIA AiSAQ: All-in-Storage ANNS with Product Quantization, a novel method of index data placement, is a trademark of KIOXIA.

KIOXIA