KIOXIA AiSAQ™ Technology

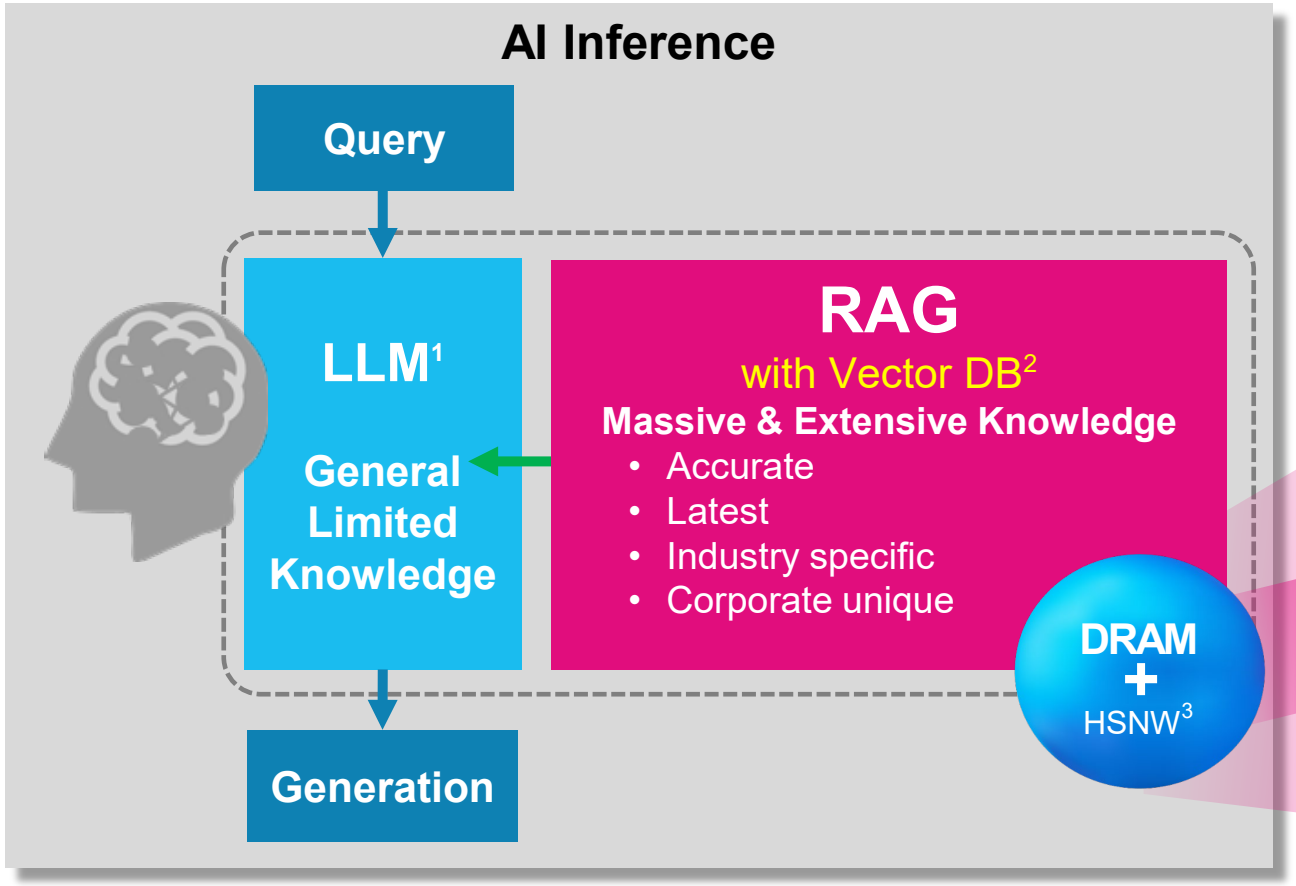
All-in-storage ANNS Algorithms Optimize VectorDB Usability within a RAG System

Assaf Sella
Vice President, Machine Learning R&D

FMS
the Future of Memory and Storage

KIOXIA

KIOXIA AiSAQ™: Contributing to AI Industry by Enabling Truly Scalable RAG



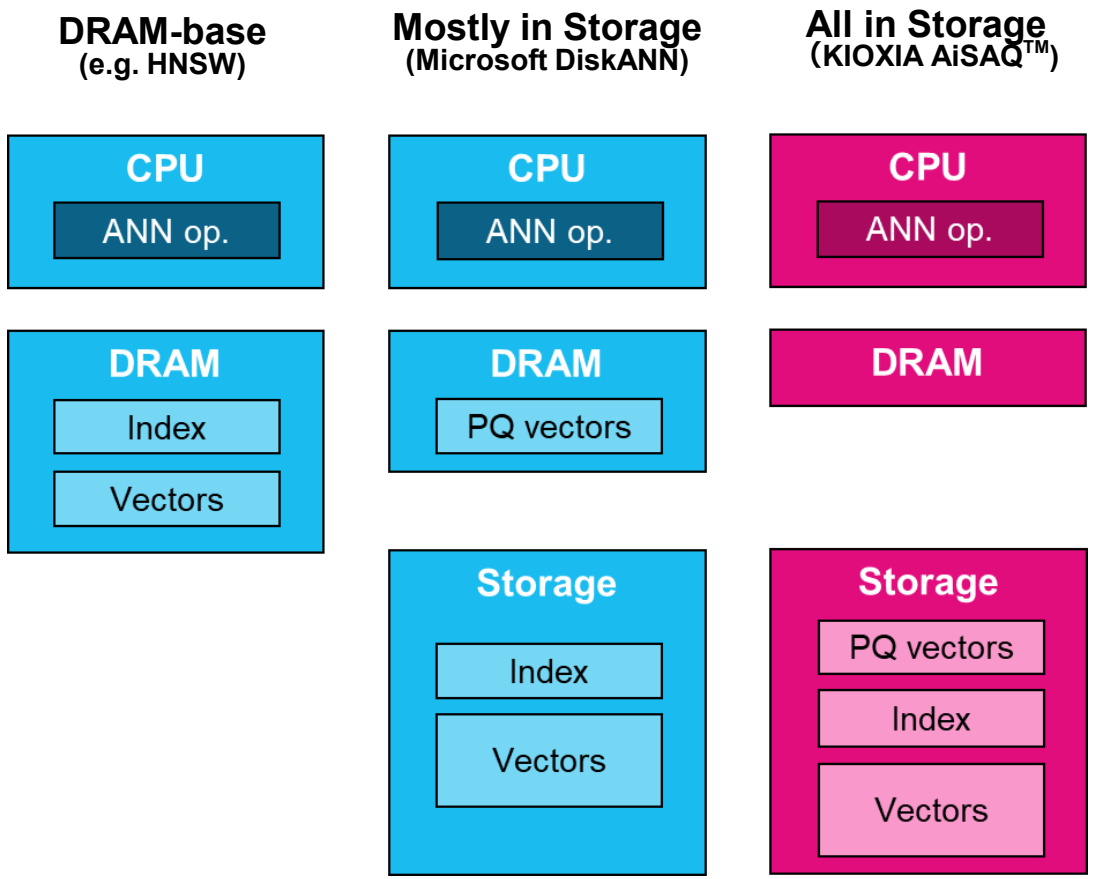
- AiSAQ removes DRAM limitation for scalable retrieval augmented generation (RAG)

- Need for RAG is getting bigger, so is need for bigger vector databases
- Open-source contribution for maximizing SSD utilization

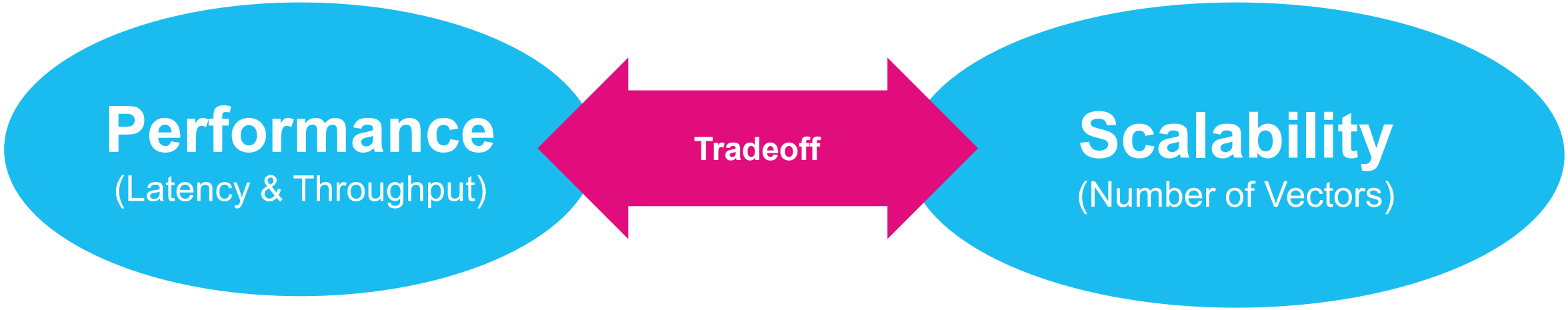
1. Large language model (LLM) 2. Data base (DB) 3. Hierarchical Navigable Small World (HNSW)

KIOXIA AiSAQ™ at a Glance

- **All-in-storage ANNS[1]:** Powered by a **near-zero DRAM** architecture with several unique algorithms to optimize performance while holding all data structures on SSD
- **High performance:** Innovative algorithms **optimize data-structure arrangement** on SSD media **for I/O[2] reduction**
- **High scalability**: economic scaling of vector DB, not limited by DRAM cost
- Strong supportability of **multi-tenant environment**
- **Open-source contribution** for the development of the generative AI



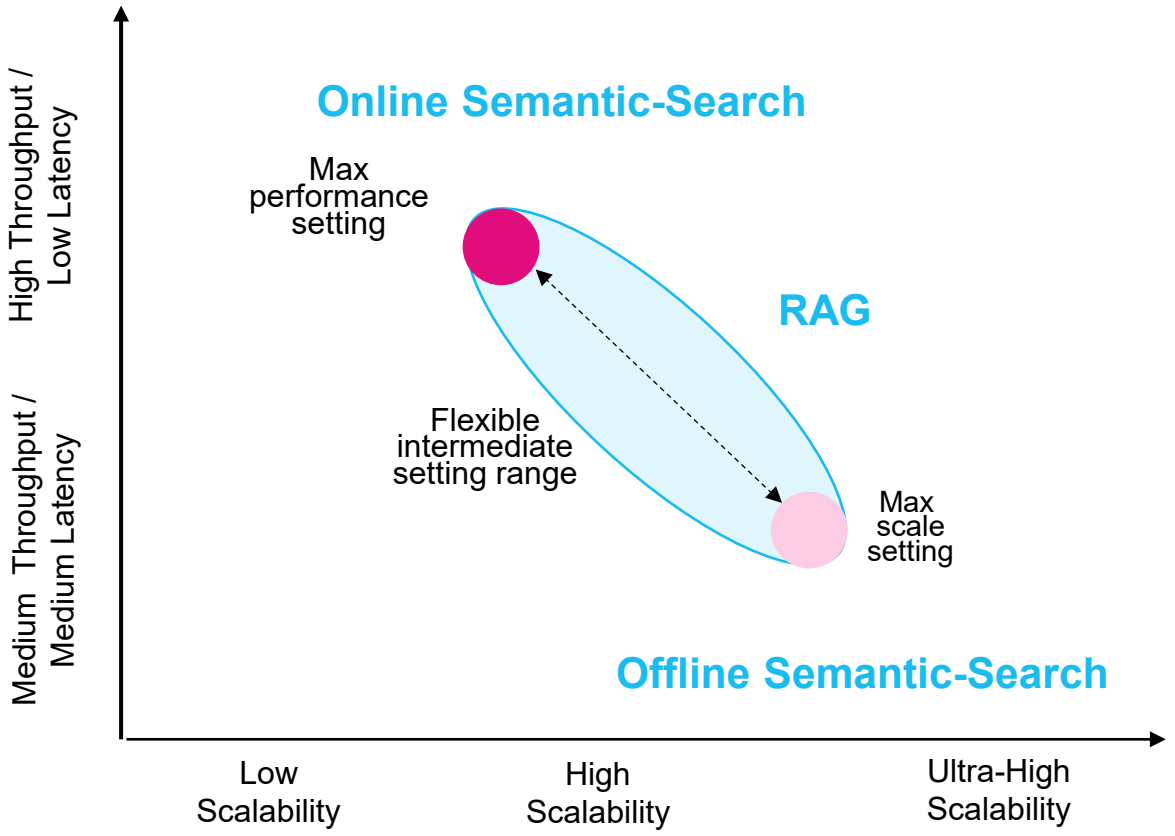
KIOXIA AiSAQ™ Provides Flexibility Between Performance and Scale



- **AiSAQ New Release:**
  - ✓ **Flexible setting** allows optimal balance between performance and scale

- **Tunable for selecting the optimal point** between maximum scale and maximum throughput
  - ✓ Meets or exceeds the **performance requirements** of various vector DB applications
  - ✓ Delivers significant **cost reduction** and enables **ultra-high-scale deployments**

The Range of KIOXIA AiSAQ™ - Flexibility Between Performance and Scale



	Max **Performance** Setting	Max **Scale** Setting
P95[1] Latency	**5** ms[2]	**< 30** ms
Throughput	**12,700** QPS[3]	**> 400 QPS**[4] (1 SSD) **> 2,200 QPS** (6 SSDs)
Scale tenants/server	**36,000**	**256,000**

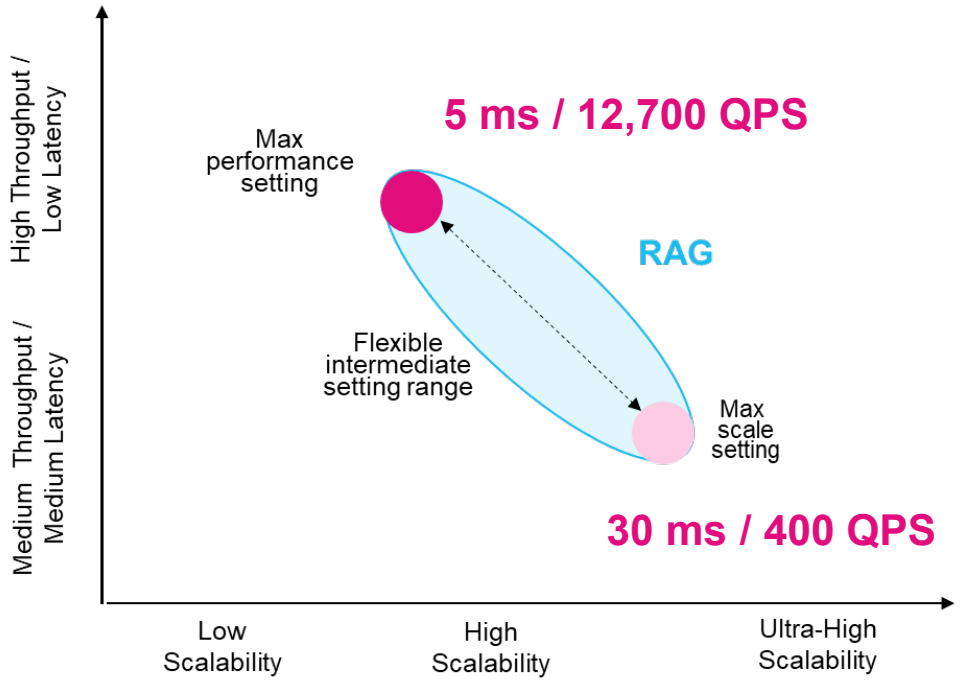
Example for 50M / 768 dimensions dataset (95% Recall@10)[5,6]

Source: KIOXIA engineering team

AiSAQ update release to open source on July 3, 2025

Images and/or graphics within this slide are the property of Kioxia Corporation (KIOXIA) and are reproduced with the permission of KIOXIA. 1. 95th percentile of latency indicates the latency value below which 95% of requests to a system are served. 2. Milliseconds (ms) 3. Queries Per Second (QPS)
4. Throughput using single SSD (CD8P-R 7.68TB) 5. Platform: Dell® PowerEdge™ R760xa, Dual processor Intel Xeon Gold 5418Y (2x24 Cores @ 2GHz), DRAM: 512GB, 16 x 32GB DDR5, 4400MT/s DIMM, Storage: 6 x CD8P-R 7.68TB drives. Dataset: WikiAll 50M, 768 dimensions/vector (3KiB/vector). AiSAQ-P parameters: R=56, Lbuild=200, PQ=128B, DiskPQ=768B, BW=4. AiSAQ-S parameters: R=60, Lbuild=200, PQ=128B, DiskPQ=768B, BW=1, VBW=1, Dcache=6MB. 6. The system is expected to retrieve at least 95% of the true nearest neighbors within the top 10 results returned for a given query. Dell and PowerEdge are trademarks of Dell Inc in the U.S. and/or other jurisdictions.

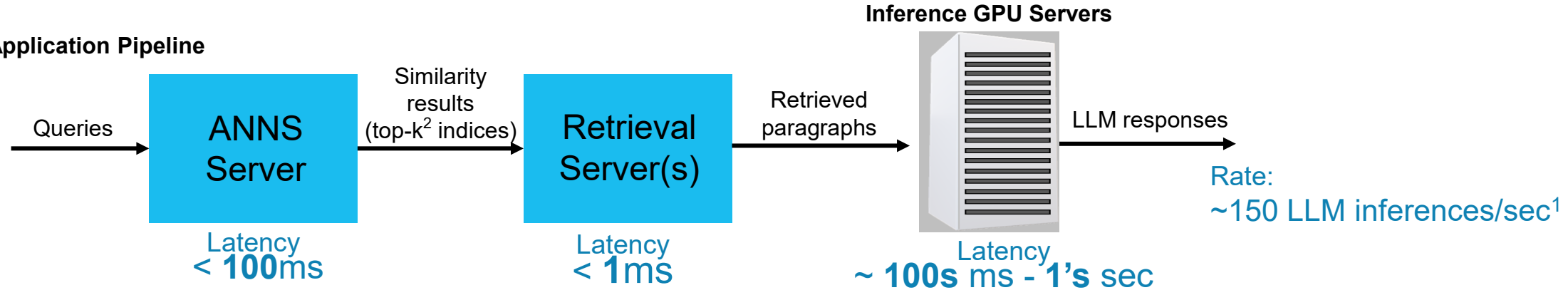
Optimization for RAG



KIOXIA AiSAQ™ allows flexible setting to meet the optimal work point in various deployment models

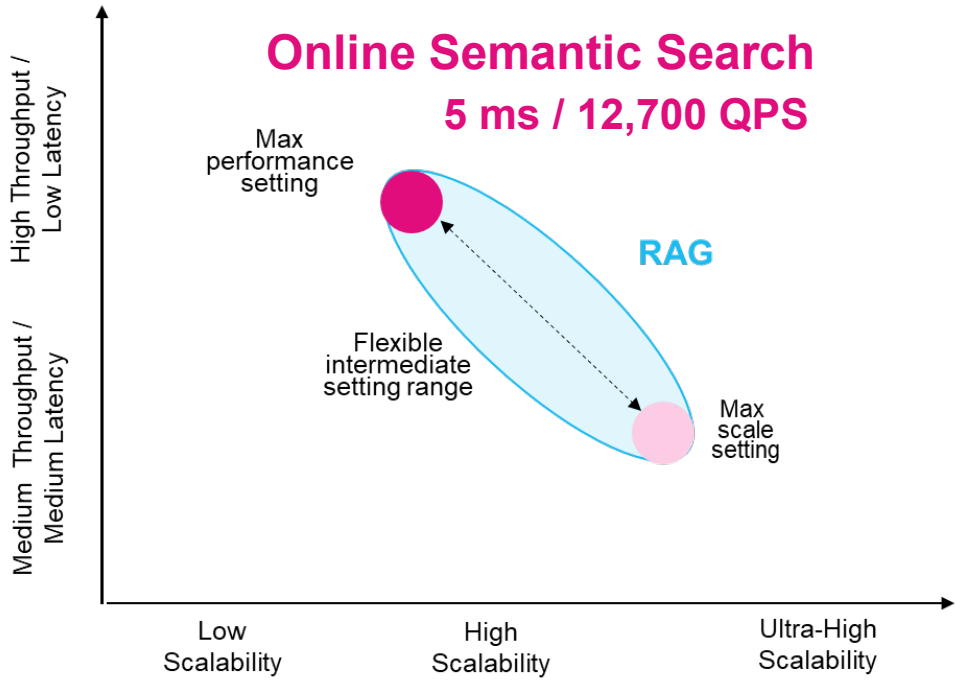
- **RAG**: Augments LLM by using external data sources
- **Requirements**
  - ✓ Latency/throughput requirement could be **high or low** depending on deployment model
  - ✓ **Dense multi-tenant** requires **high** throughput
  - ✓ **LLM inference latency** could be much **higher than ANNS** latency

RAG Application Pipeline



[1]Assuming NVIDIA® GB200 NVL72 rack (72 GPUs), 4 LLMs/GPU (100B model, 4b quantization), and 2 sec average inference latency per GPU (200 tokens/reply, inference rate: 100 tokens/sec)

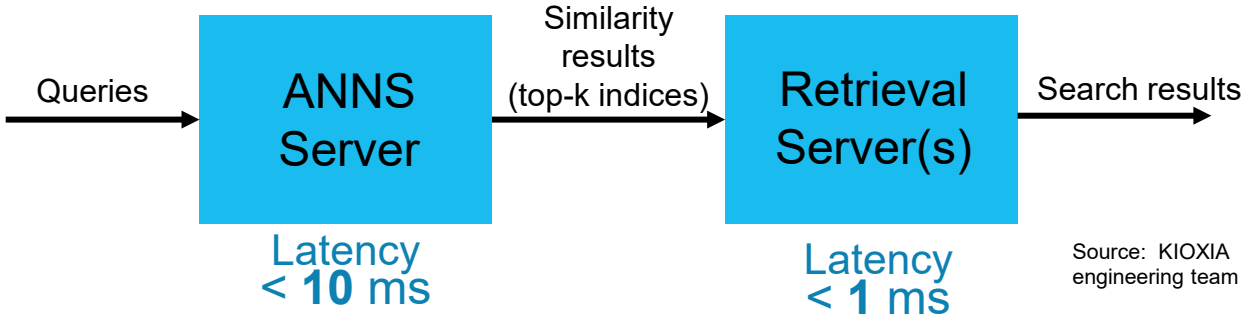
Optimization for Online Semantic Search



Max performance setting meets online semantic search apps that require real time response

- **Online sematic search**:
  - ✓ Find texts that semantically match user queries
  - ✓ Example use cases: e-commerce, web search
- **Requirements**
  - ✓ Latency for fast user experience: **typically 10ms**
  - ✓ Supports massive concurrent queries: **1,000s QPS**

Online Semantic Search Application Pipeline

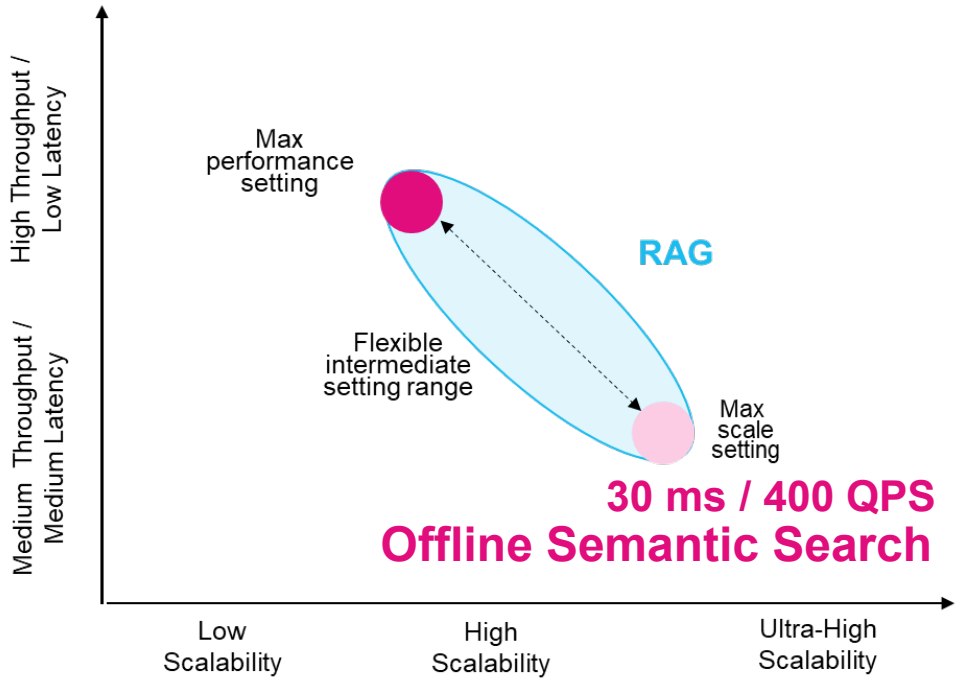


Source: KIOXIA engineering team

	Online Search Requirement	KIOXIA AiSAQ™ Max Performance
P95 Latency	10 ms	5 ms
Throughput	~ 1000's QPS	12,700 QPS

Example for 50M / 768 dimensions dataset (95% Recall@10)

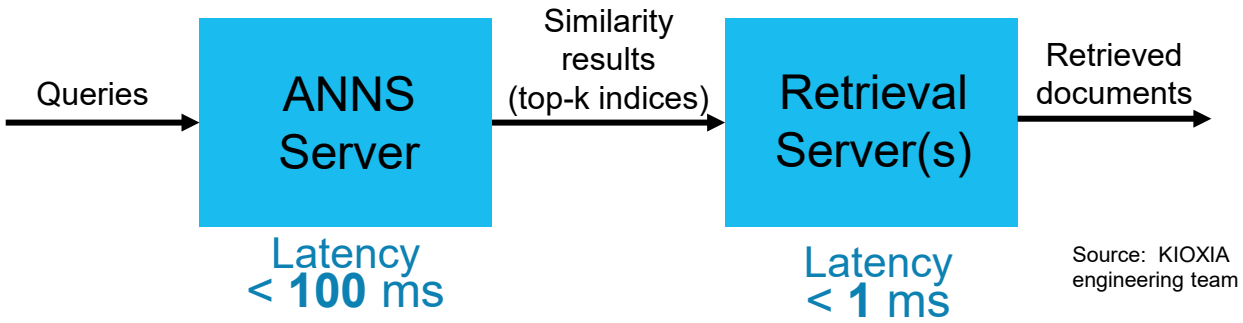
Optimization for Offline Semantic Search



Max scale setting exceeds the requirements of offline sematic search applications

- **Offline sematic search**:
  - ✓ Find texts that semantically match user queries
  - ✓ Example use cases: corporate wiki, legal research, engineering research, vector lake (ultra high scale datasets)
- **Requirements**
  - ✓ Medium latency : **typically <100msec**
    - Review time of retrieved documents by user is much longer than ANNS latency
  - ✓ Medium throughput: **100s QPS**
    - Throughput aligned with target number of users concurrently quiring the dataset

Offline Semantic Search Application Pipeline



Source: KIOXIA engineering team

	Offline Search Requirement	KIOXIA AiSAQ™ Max Scale
P95 Latency	100 ms	< 30 ms
Throughput	~ 100s QPS	> 400 QPS > 2,200 QPS

Example for 50M / 768 dimensions dataset (95% Recall@10)

Economic Index for High-Scale RAG and Offline Search Deployments

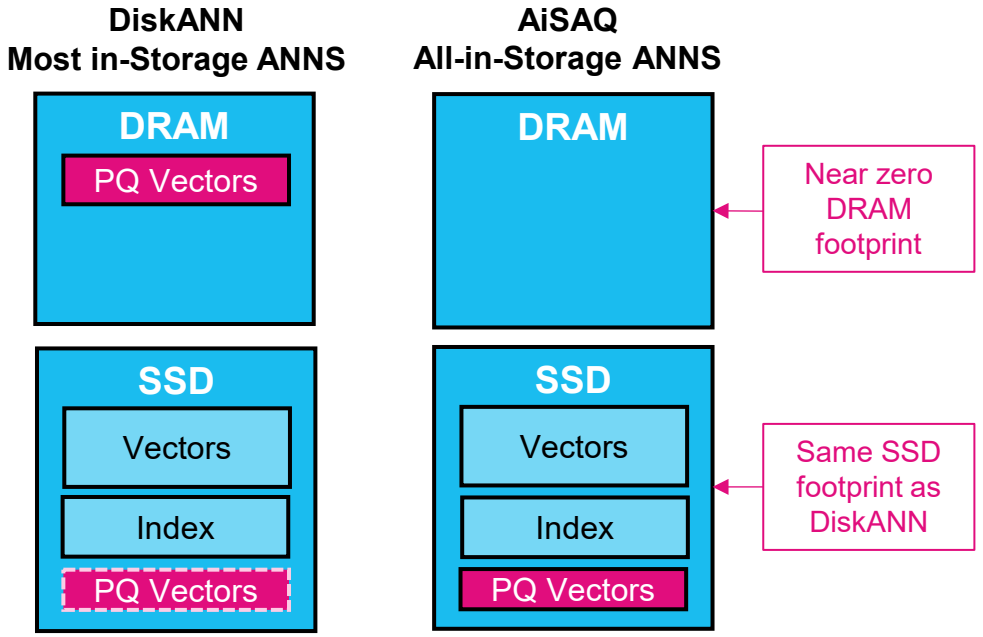
KIOXIA AiSAQ™ delivers significant cost reduction and enables reaching ultra-high scale deployments economically

	Advantage
Cost	4x – 7x more cost effective than DiskANN[1] (with TLC/QLC SSDs)
Scale	Reach ultra-high scale economically due to reduced cost and elimination of DRAM structure (PQ[2] vectors)

Search-media Cost for Aggregated 10B High-Dimensionality Dataset[3]

Vector DB Elements	HNSW PQ	DiskANN	AiSAQ Max Scale
PQ Vectors ($PQ=768B_{HNSW}/128B_{DiskANN/KIOXIA\ AiSAQ}$)	7.68TB	1.28TB	1.28TB
Index[4] (M=48/R=60)	4.32TB	10TB	10TB
DRAM footprint	10.9TiB	1.2TiB	~0
SSD footprint	12TB[5]	11.2TB	11.2TB
Search media cost: DRAM & TLC SSD	**100%**	**14%**	**3.5%**
Search media cost: DRAM & QLC SSD	**100%**	**12.6%**	**1.6%**

4-7x cost reduction vs. DiskANN



Source: KIOXIA engineering team. Cost analysis is based on assumptions regarding price per gigabyte for DRAM, TLC and QLC SSDs, and presented as a relative reference in percentage. Images and/or graphics within this slide are the property of Kioxia Corporation (KIOXIA) and are reproduced with the permission of KIOXIA.

1.Disk-based approximate nearest neighbor search algorithm developed by Microsoft (DiskANN) 2. Product Quantized (PQ) 3. Combined cost of DRAM and SSD used by ANNS. Dataset with 768 dimensions. 4. HNSW index size calculation based on 9M bytes per vector. 5. Definition of Capacity: Kioxia Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes, a terabyte (TB) as 1,000,000,000,000 bytes and a petabyte (PB) as 1,000,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1Gbit = 2^30 bits = 1,073,741,824 bits, 1GB = 2^30 bytes = 1,073,741,824 bytes, 1TB = 2^40 bytes = 1,099,511,627,776 bytes and 1PB = 2^40 bytes = 1,125,899,906,842,624 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

Advantages for Multi-Tenant Deployments

- Multi-tenancy is a common ANNS deployment: multiple users (tenants) run gen-AI applications on their private datasets
- Aggregated size of datasets can be huge: 1,000 tenants x 10M vectors/tenant → 10B vector dataset
- **KIOXIA AiSAQ™ provides highest tenant density with no cold-start delay for low latency & multi-tenant applications**

	DiskANN	AiSAQ Max Performance	Advantage
Number of tenants per server	Limited by DRAM capacity	Limited by SSD capacity	5x more active tenants in KIOXIA AiSAQ vs. DiskANN
Loading "cold tenants" from SSD to DRAM	"Cold" tenants must be loaded to DRAM prior to search	All tenants are always active	No "cold start" latency providing seamless tenant switching

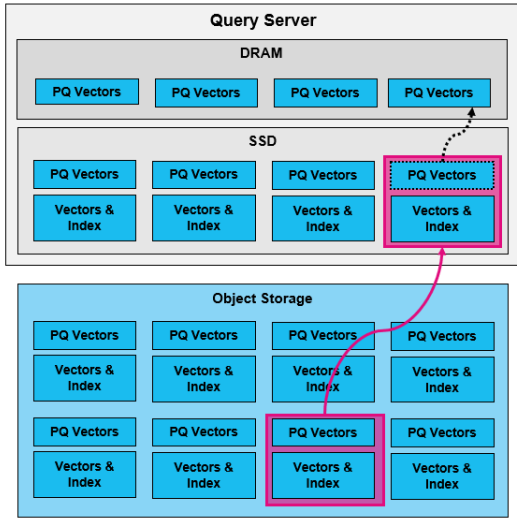
Active Tenants per Server for KIOXIA AiSAQ, DiskANN and HNSW[1]

Vector DB Elements	HNSW	DiskANN	KIOXIA AiSAQ Max Performance
Limited by	8 TiB DRAM		24 x 122.88 TB[3] SSD[2]
Number of active tenants[1]	**730**	**6,800**	**36,000**
Aggregated dataset size (vectors)	7.3B	68B	360B
DRAM footprint	7.9 TiB	7.9 TiB	~0
SSD footprint	8.8 TB	78.3 TB	2,949 TB
Relative cost / tenant	100%	12.6%	11.6%

[1] Active tenants with 10M vectors per tenant (768B/vector), and servers with up to 8TiB DRAM and 24 SSD slots
[2] Using KIOXIA LC9 High-Capacity SSD

Source: KIOXIA engineering team. Cost analysis is based on assumptions regarding price per gigabyte for DRAM, TLC and QLC SSDs, and presented as a relative reference in percentage.

DiskANN: Cold Start Latency of Non-Active Tenant

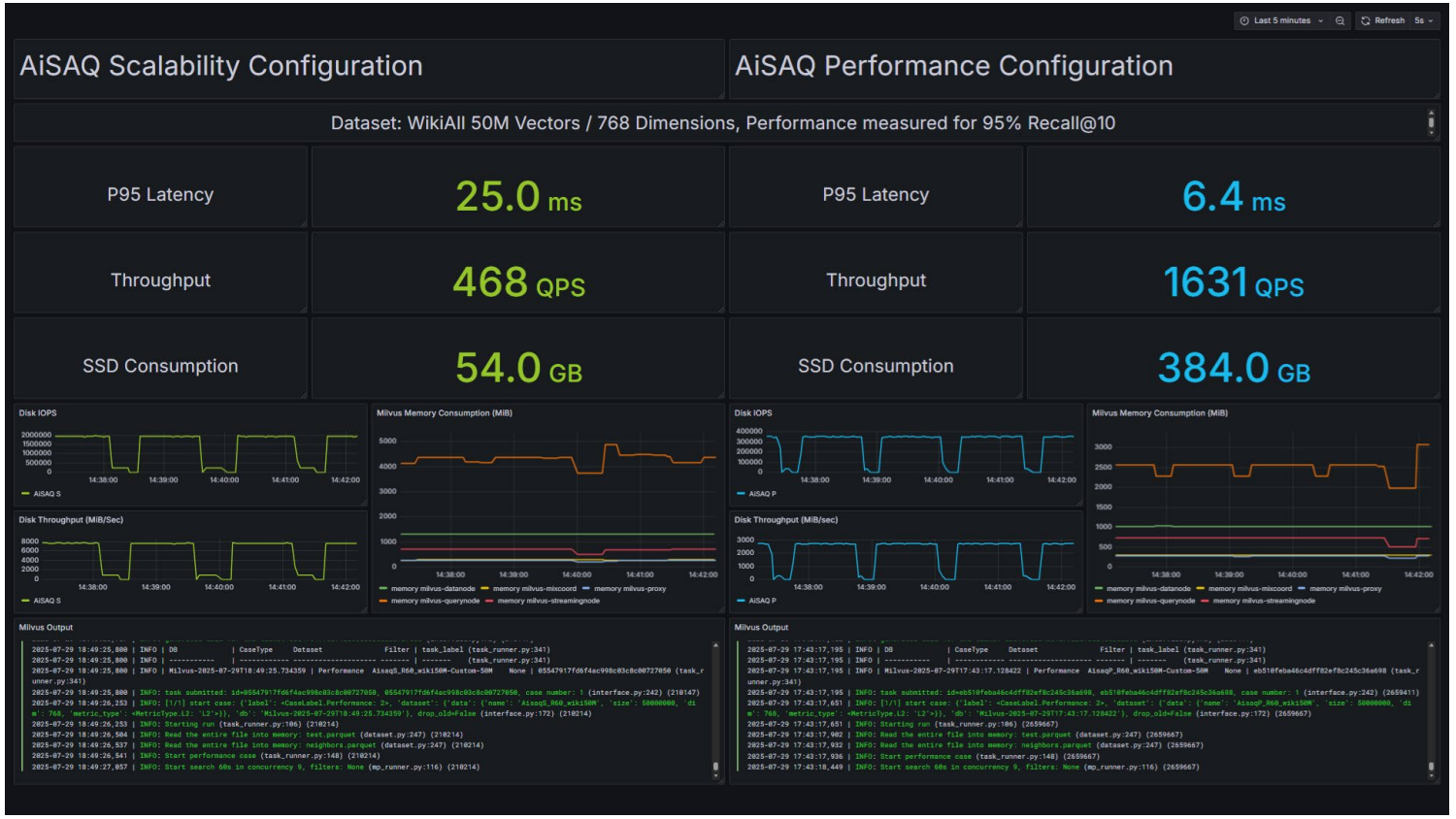


Integration with Milvus®

- Milvus is a leading open-source vector database designed to manage large-scale embedding datasets
- Popular choice for many companies with gen-AI applications requiring fast and accurate semantic matching at scale

Image used with permission from Milvus

- KIOXIA integrated KIOXIA AiSAQ™ into Milvus
- Integrated AiSAQ meets or exceeds target ANNS performance requirements



Summary

- KIOXIA AiSAQ™ is an **open-source "all-in-storage" ANNS** algorithm
- **Utilization of SSD** by AiSAQ delivers **multiple advantages for expanding vector DB application needs**
  - ✓ High scalability
  - ✓ High performance (low latency and high QPS)
  - ✓ Flexibly offers the optimal balance between scalability and performance
  - ✓ Various advantages for multi-tenancy deployments
  - ✓ Excellent solution for both RAG and semantic search
- KIOXIA integrated **AiSAQ in Milvus®** – a leading open-source vector DB

Visit KIOXIA booth #307 for the demo of integrated solution

KIOXIA