KIOXIA

Performance Brief

Accelerating Vector Database Performance through Disk-based Indexes with KIOXIA CD8P Series SSDs Deployed in a Supermicro® ASG-1115S-NE316R Server

Introduction

Artificial intelligence (AI) is becoming omnipresent and the list of applications, use cases and benefits are growing continually. Large language models (LLMs) are an important generative AI use case and include vast deep learning models pre-trained on large datasets. Storing these large datasets and the models' parameters can require a large amount of DRAM, which poses significant system design and cost challenges.

LLMs are trained on large amounts of text data consisting of words and sentences that derive semantic meaning and context to generate a meaningful response. However, oftentimes the model can generate irrelevant or factually incorrect responses. To prevent these incorrect responses (also known as AI hallucinations), AI developers have found ways to give the model a source of factual data that can be searched. This technique is known as Retrieval Augmented Generation (RAG) where indexing and searching of factual data is performed using a special type of database called a vector database (DB).

A vector DB stores original data (i.e., images, audio, text), as well as encoded data known as embeddings. Embedding models store vector data with high dimensionality, and Approximate Nearest Neighbor (ANN) algorithms are used to search for data points that are similar to the initial query data point. This approach can often be much quicker than an exhaustive linear search. Vector DB indexes can be created for vector datasets which allows for the similarity searches to complete even more quickly.

One of the most commonly used vector indexes, Hierarchical Navigable Small Worlds (HNSW), requires significant system memory that can become very expensive, as datasets and dimensionality grow. To counter the growing size of vector indexes, a new type of indexing was created that stores and searches the vector indexes using disk. More recent advances in vector DB indexing use Disk Approximate Nearest Neighbor[1] (DiskANN) algorithms that can offload the system memory footprint to disk without sacrificing performance. To support DiskANN algorithms, fast underlying storage is required.

This performance brief presents performance tests using DiskANN algorithms with a KIOXIA CD8P-R Series SSD - E3.S form factor, data center-class, read-intensive, and compliant with the PCIe® 5.0 specification and NVMe™ 2.0 protocol - that delivers high-read throughput and low latency in support of a broad range of applications and workloads. The KIOXIA CD8P-R Series SSD deployed in a Supermicro ASG-1115S-NE316R server was used to hold two prepared Cohere™ datasets[2] and a LAION dataset in a Milvus®[3] vector DB. The Cohere datasets included 1 million and 10 million vectors while the LAION dataset included 100 million vectors. All datasets utilized 768 dimensions per vector.

The tests determined whether the KIOXIA CD8P-R Series SSD using DiskANN algorithms can deliver similar or improved performance when compared with purely system memory-based indexes, such as HNSW. The similarity search tests were performed through VectorDBBench software to test for metrics such as database throughput as well as recall accuracy of retrieved results.

The test results show that a KIOXIA CD8P-R Series SSD using DiskANN algorithms was able to outperform or match memory-based indexes using ANN algorithms in queries per second (QPS), QPS per dollar and recall, while effectively lowering the system memory footprint. The results also demonstrate that in order to support LLM vector DB applications, using SSDs can minimize system memory use, which ultimately leads to lower overall system cost without sacrificing on performance. This enables DRAM resources to be freed for other tasks and is more efficient for processing large datasets, such as those used in data analytics, AI, and machine learning/deep learning applications.

The test results presented include a brief description of each workload test, a graphical depiction of the test results and an analysis. Appendix A covers the hardware and software configuration – Appendix B covers the configuration set-up and test procedures.

Test Results Snapshot

A KIOXIA CD8P-R Series SSD using disk-based indexes in a Supermicro® ASG-1115S-NE316R server delivered the following vector DB performance results when compared with system memory-based indexes:

Queries per Second (QPS)
(higher is better)

1M Vectors	10M Vectors	100M Vectors
+85%	+149%	+84%

Total System Memory Used
(lower is better)

1M Vectors	10M Vectors	100M Vectors
-60%	-72%	-76%

QPS per Dollar
(higher is better)

1M Vectors	10M Vectors	100M Vectors
+3x	+4x	+3x

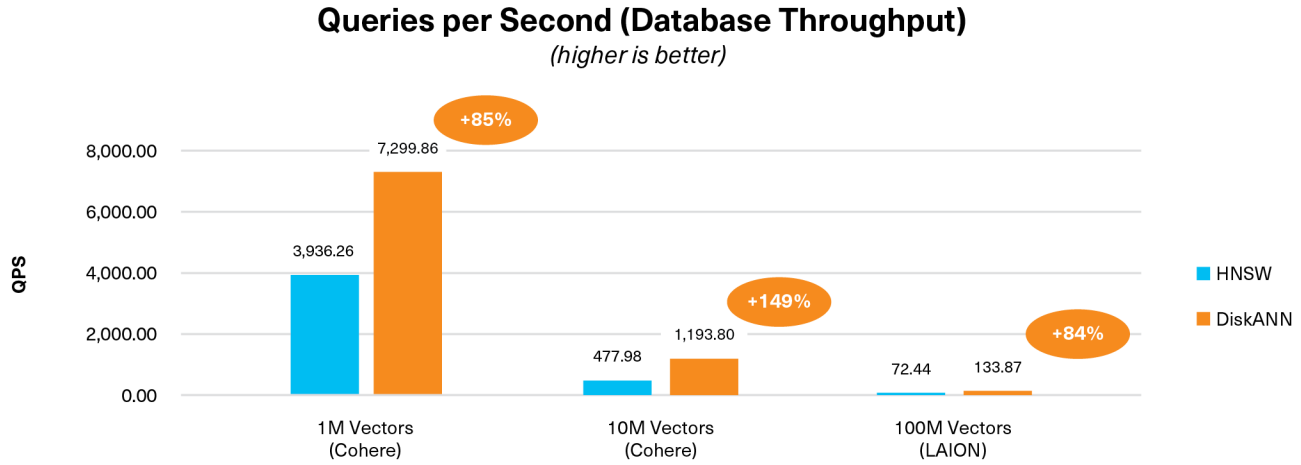
Recall
(higher is better)

1M Vectors	10M Vectors	100M Vectors
Similar	Similar	Similar

Test Results[4]

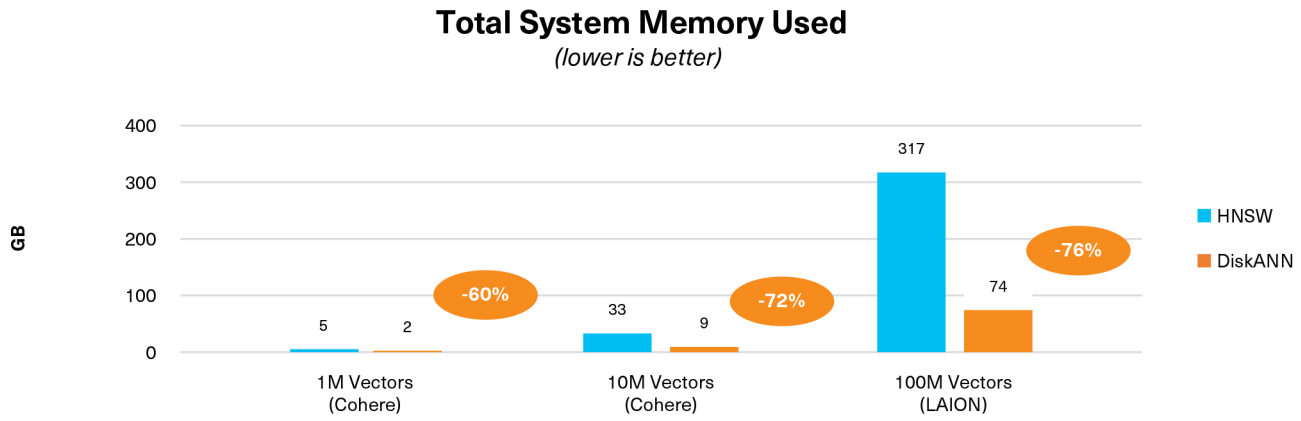
Metric 1: Queries per Second (Database Throughput)

This metric measured the number of queries per second (QPS) that the Milvus® vector DB can achieve when querying the vector space for a similarity search. Higher database throughput indicates that the compute resources are working more efficiently to scan through more vectors in the index faster. High database throughput enables the results of the similarity search to return to the LLM quickly. The DiskANN index, in conjunction with a KIOXIA CD8P-R Series SSD, is able to outperform an HNSW index at all dataset sizes. The database throughput results are in QPS and the higher result is better.



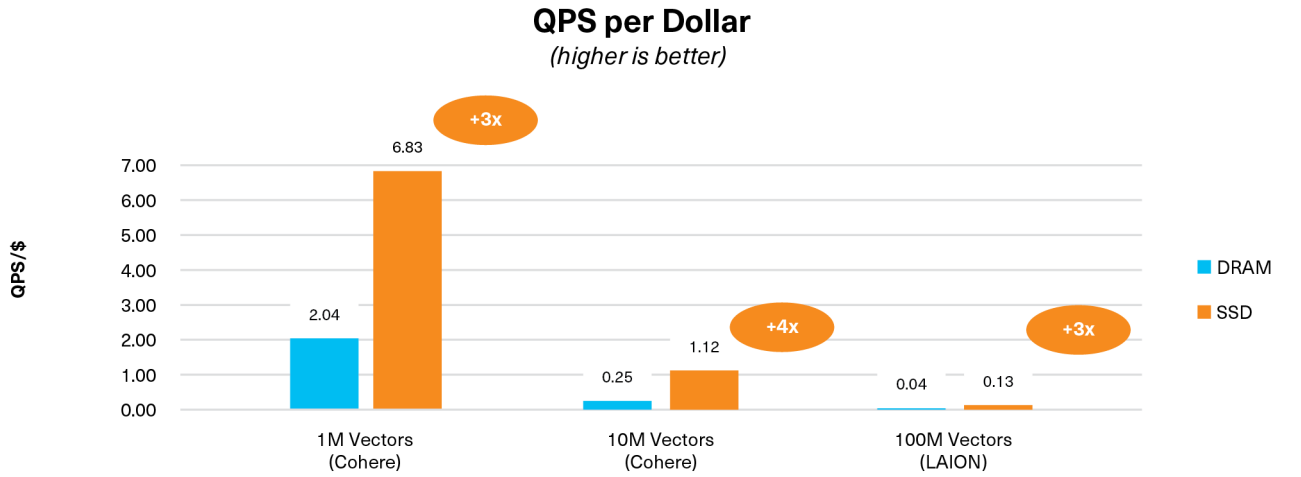
Metric 2: Total System Memory Used

This metric measured the amount of system memory that was utilized by the Milvus process when the vector DB tests were run. In many cases, the combined size of the initial vector dataset, as well as the vector index that is created against this dataset, can become incredibly large. With vector indexes that are stored into memory, it becomes increasingly more expensive to install enough DRAM into a single server to hold the necessary data, with some data being so large that it exceeds the maximum capacity capabilities of a single node, forcing horizontal scaling. Algorithms that use disk instead of DRAM can lower and offload the system memory footprint requirement to run RAG-based applications. This capability allows for more efficient vertical scaling within individual nodes, which can lead to more effective horizontal scaling. System purchases can be altered so that less DRAM is required to run large vector DB datasets by storing the indexes on fast SSDs with much higher available capacities. The results are in gigabytes[5] (GB) and the lower result is better.



Metric 3: QPS per Dollar

This metric measured[6] the amount of costs that can be saved by switching to a disk-based index and SSD as opposed to a memory-intensive index in system memory without sacrificing vector DB throughput performance. With more expensive DRAM costs, coupled with larger datasets, it becomes inefficient, and in some cases, infeasible, to store all of this content into system memory. With KIOXIA CD8P-R Series SSDs, large capacities are available to provide a cost-effective solution. The results are in queries per second per dollar (QPS/$) and the higher result is better.



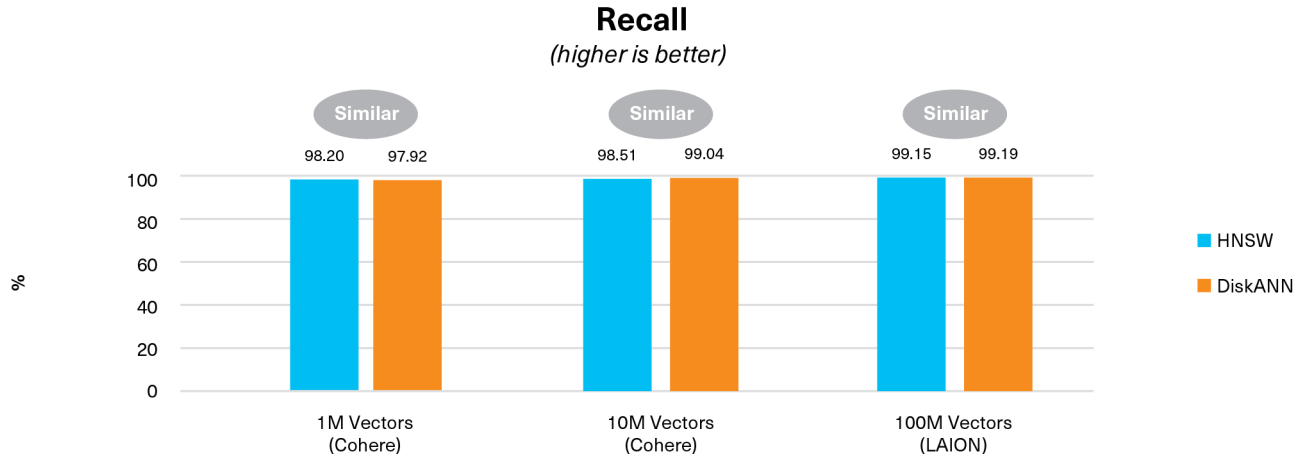
Metric 4: Recall

This metric measured the ability of the Milvus® vector DB to find all relevant cases (true positives) within a vector space when conducting a similarity search. The results are then calculated using the following formula:

$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

The recall shows what proportion of actual positives were identified correctly. From these calculations, the higher the recall, the more precise and similar the results will be when vector data is returned back to the Milvus vector DB.

The DiskANN index using the KIOXIA CD8P-R Series SSD was able to deliver similar recall when compared with the HNSW index. The results ensure that the similarity search returned a high percentage of relevant content without sacrificing performance. The results are in percentage (%) of recall and the higher result is better.



Analysis

Using DiskANN algorithms to offload the system memory footprint to the KIOXIA CD8P-R Series SSD enabled the DRAM resource to be utilized for other processes and tasks, delivering cost-effectiveness especially since system memory can be expensive. The KIOXIA CD8P-R Series SSD with DiskANN algorithms used 60% less system memory at 1 million vectors, 72% less at 10 million vectors and 76% less at 100 million vectors when compared with a system memory-based index.

For queries per second (database throughput), the KIOXIA CD8P-R Series SSD with DiskANN algorithms performed noticeably better at all dataset sizes. This means that IT personnel can substitute higher cost DRAM with less expensive SSDs without sacrificing performance. The KIOXIA CD8P-R Series SSD with DiskANN algorithms delivered 85% more queries per second at 1 million vectors, 149% more at 10 million vectors and 84% more at 100 million vectors when compared with a system memory-based index.

These performance results translate into cost savings as evident from the QPS per dollar results[6]. The KIOXIA CD8P-R Series SSD with DiskANN algorithms delivered 3x higher QPS per dollar at 1 million vectors, 4x higher QPS per dollar at 10 million vectors and 3x higher QPS per dollar at 100 million vectors.

The recall results demonstrate comparable accuracy percentages of the KIOXIA CD8P-R Series SSD and DiskANN index when compared with a system memory-based index. This enables IT personnel to save on operating costs and cost of ownership by using SSDs to gain accurate vector DB performance metrics instead of investing/reinvesting in relatively more expensive DRAM system memory.

Summary

The KIOXIA CD8P-R Series E3.S SSD and DiskANN index performed better than a system memory-based index at various vector DB sizes with regards to queries per second (database throughput) and also delivered comparable recall performance for accuracy. Lowering the overall system cost by utilizing fast storage devices as opposed to expensive DRAM to support a disk-based index versus a purely memory-based index was also demonstrated without sacrificing on performance. These drives are well-suited to support vector DBs that use DiskANN vector indexes as showcased by the test results. System design and associated purchases can be altered so that less DRAM is required to run large vector DB datasets by storing the indexes on fast SSDs.

KIOXIA CD8P-R Series SSD Product Info

The KIOXIA CD8P-R Series SSD product line is data center-class and compliant with the PCIe® 5.0 specification and the NVMe™ 2.0 protocol. These SSDs are available in 2.5-inch[7] and Enterprise and Datacenter Standard Form Factor (EDSFF) E3.S form factors.

For this performance brief, the KIOXIA CD8P-R Series E3.S SSD was used with 1 DWPD[8] of endurance and a 7.68 terabyte[5] (TB) capacity. The product series supports up to 30.72 TB capacity in the 2.5-inch format and up to 15.36 TB in the E3.S format with the following performance specifications[9] (single port (1x4) mode):

KIOXIA CD8P Series E3.S SSD[10]

Sequential Read:	up to 12,000 megabytes per second (MB/s); includes the 15.36 TB model
Sequential Write:	up to 5,500 MB/s; up to 5,300 MB/s for the 15.36 TB model
Random Read:	up to 2,000,000 input/output operations per second (IOPS); includes the 15.36 TB model
Random Write:	up to 200,000 IOPS; includes the 15.36 TB model

KIOXIA CD8P Series SSDs are also available for higher endurance mixed-use applications and include 3 DWPD endurance and capacities up to 12.8 TB.

Additional KIOXIA CD8P Series SSD information is available here.

Appendix A

Hardware/Software Test Configuration

Server Information	
Model	Supermicro® ASG-1115S-NE316R
No. of Servers	1
No. of CPU Sockets	1
CPU	AMD EPYC™ 9354P
No. of CPU Cores	32
CPU Frequency	3.25 GHz
Total Memory	384 GB[4] DDR5 DRAM
Memory Frequency	DDR5-4800

Operating System Information	
Operating System	Ubuntu®
Version	20.04.6 LTS
Kernel	5.8.0-50-generic

File System Information	
File System	XFS®
Mount Options	noatime, nodiratime

Vector Database Software Information	
Vector Database	Milvus®
Version	2.3.10

Test Software Information	
Test Software	VectorDBBench
Version	0.0.7
Vector Dimensions	768
Dataset #1	Cohere™
No. of Vectors (Dataset #1)	1M / 10M
Dataset #2	LAION
No. of Vectors (Dataset #2)	100M

HNSW Parameters	
M Vectors	30
efConstruction	360
ef	100

DiskANN Parameters	
search_list	100

SSD Information	
Model	KIOXIA CD8P-R Series
Interface	PCIe® 5.0 x4
Protocol	NVMe™ 2.0
No. of Drives	1
Form Factor	EDSFF E3.S
Capacity	7.68 TB
DWPD	1 (5 years)

Appendix B

Configuration Set-up/Test Procedures

Configuration Set-up

A Supermicro® ASG-1115S-NE316R server was setup with an Ubuntu® 20.04.6 LTS operating system.

One 7.68 TB KIOXIA CD8P-R Series E3.S SSD was installed into the server.

The SSD was setup with an XFS® file system and mounted with noatime and nodiratime flags.

The Milvus® vector DB was downloaded and configured so that raw data, vectors and indexes would be stored onto the SSD.

The VectorDBBench load generator was installed to run a search performance test for various vector DB sizes.

The Cohere™ and LAION datasets were downloaded and inserted into the Milvus vector DB. The Cohere datasets consisted of 1 million and 10 million vectors with 768 dimensions, while the LAION dataset consisted of 100 million vectors with 768 dimensions.

Test Procedures

The VectorDBBench benchmark test was run utilizing an HNSW index and the following metrics were gathered:

- *Total System Memory Used (in GB)*
- *Queries per Second (in QPS)*
- *QPS per Dollar (in QPS/$)*
- *Recall (in %)*

The tests above were then repeated with a DiskANN index.

The metrics were compared between the two indexes.

NOTES:

[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;

[2] Cohere dataset is part of the Cohere platform used to generate high-quality vector DB embeddings. The Cohere Embed API endpoint is used to generate language embeddings, and then those embeddings are indexed into the Pinecone™ vector DB for fast and scalable vector searches.

[3] Milvus is an open-source vector database built to power embedding similarity search and AI applications, and makes unstructured data searches more accessible.

[4] Read and write speed may vary depending on the host device, read and write conditions, and file size.

[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.

[6] The QPS per Dollar metric was calculated by dividing the QPS metric reported in Metric 1 from the pricing available on public sources as of this publication date for the total amount of DRAM or SSDs used in the server solution for HNSW and DiskANN algorithms, respectively.

[7] 2.5-inch indicates the form factor of the SSD and not the drive's physical size.

[8] DWPD: Drive Write(s) Per Day: One full drive write per day means the drive can be written and re-written to full capacity once a day, every day, for the specified lifetime. Actual results may vary due to system configuration, usage, and other factors.

[9] The KIOXIA CD8P-R Series E3.S SSD performance specifications provided by KIOXIA Corporation and accurate as of this publication. Specifications are subject to change.

[10] The product image shown is a representation of the design model and not an accurate product depiction.

TRADEMARKS:

AMD EPYC and combinations thereof are trademarks of Advanced Micro Devices, Inc. Cohere is a trademark of Cohere Communications, LLC. Milvus is a registered trademark of LF Projects, LLC. NVMe is a registered or unregistered trademark of NVM Express, Inc. in the United States and other countries. PCIe is a registered trademark of PCI-SIG. Pinecone is a trademark of Pinecone, LLC. Supermicro is a registered trademark of Super Micro Computer, Inc. or its subsidiaries in the United States and other countries. Ubuntu is a registered trademark of Canonical Ltd. XFS is a registered trademark of Silicon Graphics International Corporation or its subsidiaries in the United States and/or other countries. All other company names, product names and service names may be trademarks of third-party companies.

DISCLAIMERS:

KIOXIA America, Inc. may make changes to specifications and product descriptions at any time. The information presented in this performance brief is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of KIOXIA America, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and KIOXIA America, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may render inaccuracies for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. KIOXIA America, Inc. assumes no obligation to update or otherwise correct or revise this information.

KIOXIA America, Inc. makes no representations or warranties with respect to the contents herein and assumes no responsibility for any inaccuracies, errors or omissions that may appear in this information.

KIOXIA America, Inc. specifically disclaims any implied warranties of merchantability or fitness for any particular purpose. In no event will KIOXIA America, Inc. be liable to any person for any direct, indirect, special or other consequential damages arising from the use of any information contained herein, even if KIOXIA America, Inc. are advised of the possibility of such damages.