



## Performance Brief

# Accelerating Vector Database Performance with Disk-based Indexes Using a PCIe® 5.0 Dell™ PowerEdge™ R7725xd Server / KIOXIA CM7 Series SSD Platform Over a PCIe 4.0 Generation Platform

## Introduction

Large Language Models (LLMs) are a type of AI designed to understand, generate and process multiple forms of output, including human-like text, by analyzing vast datasets using deep learning. They contain neural network components such as decoders and encoders that process and generate text. Encoders understand context by transforming input text into a numerical, semantic representation, while decoders take this representation to generate new, sequential and context-aware text. These models are generative and produce meaningful, coherent text based on user prompts. However, LLMs can often produce irrelevant or factually incorrect results known as hallucinations and can fabricate results that do not exist. It is also a challenge to keep LLMs trained on the most recent data, as retraining can be very expensive with respect to time and energy.

To alleviate hallucinations, Retrieval Augmented Generation (RAG) is a technique used to give LLMs access to specific, private or real-time data that wasn't included in the original training dataset. In a RAG pipeline, a vector database (DB) acts as the external knowledge base that provides LLMs with specific, relevant facts. Vector DBs are a specialized system designed to store, manage and query high-dimensional vector embeddings. These embeddings are numerical representations of complex data (like text, images or audio) in a multidimensional space where the distance between two vectors directly corresponds to the semantic or conceptual relationship between the original data points. A vector DB contains up-to-date and factually correct data that goes well beyond an LLM's training date and is used to alleviate hallucinations (known as grounding).

If an LLM were to use a vector DB without a vector index, it would then have to perform a brute force comparison of all the vectors stored with the initial query vector. This process can be very time consuming with large datasets and can make the interaction with the LLM seem delayed and unresponsive. Vector indexes organize these vectors in a way to allow for efficient lookup by narrowing down the search space, which decreases the overall time an LLM may take to generate a response. Vector indexes based on Approximate Nearest Neighbor (ANN) algorithms can achieve high accuracy while only using relevant vectors within this narrowed down search space.

Popular vector indexes, such as Hierarchical Navigable Small Worlds (HNSW), require huge memory footprints to store its graph and full-precision vectors of large datasets. However, alternative vector indexes are available such as Disk Approximate Nearest Neighbor<sup>1</sup> (DiskANN) algorithms that store compressed vectors in memory and offload the full-precision vectors and nearest neighbor information to disk. To support these disk-based algorithms, fast underlying storage is required alongside fast servers using the latest PCIe 5.0 interface.

**This performance brief presents** performance tests using DiskANN algorithms in a PCIe 5.0 Dell PowerEdge R7725xd server deployed with a KIOXIA CM7-R Series SSD compared with a PCIe 4.0 Dell PowerEdge R7525 server deployed with a KIOXIA CM6-R Series SSD. The SSDs held two prepared [Cohere™ datasets](#)<sup>2</sup> and a Large-scale Artificial Intelligence Open Network ([LAION](#)) dataset in a Milvus<sup>®</sup> 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. VectorDBBench software was used to perform similarity searches, and metrics including load duration, database throughput, database throughput per watt, percentage (%) of recall and 95<sup>th</sup> / 99<sup>th</sup> percentile latency were gathered.

**The test results show** that the PCIe 5.0 server/SSD platform provided higher performance, performance per watt and lower latencies with similar recall as compared to a PCIe 4.0 platform. The tests verified that supporting vector DBs for LLMs, the latest generation storage with PCIe 5.0 throughput and low latency is essential to process more data. Included is a brief description of the workload tests, test results, analysis, hardware and software configuration (Appendix A) and the configuration setup and test procedures (Appendix B).

### Test Results Snapshot

Using disk-based indexes, a PCIe 5.0 platform (Dell PowerEdge R7725xd Server and KIOXIA CM7-R Series SSD) delivered the following results versus a PCIe 4.0 platform:

**Load Duration**  
(lower is better)

1M Vectors	10M Vectors	100M Vectors
-1.4x	-1.7x	-1.4x

**Database Throughput**  
(higher is better)

1M Vectors	10M Vectors	100M Vectors
+6.2x	+4.5x	+2.5x

**Database Throughput / Watt**  
(higher is better)

1M Vectors	10M Vectors	100M Vectors
+2.9x	+2x	+1.1x

**Percentage of Recall**  
(higher is better)

1M Vectors	10M Vectors	100M Vectors
99.5%	99.5%	99.2%

**95<sup>th</sup> Percentile Latency**  
(lower is better)

1M Vectors	10M Vectors	100M Vectors
-2x	-1.6x	-1.8x

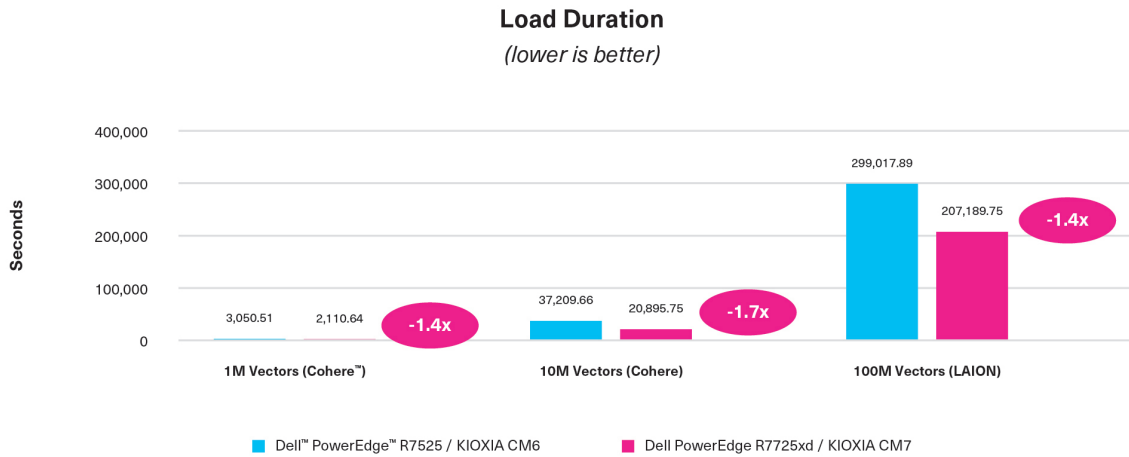
**99<sup>th</sup> Percentile Latency**  
(lower is better)

1M Vectors	10M Vectors	100M Vectors
-2x	-1.7x	-1.5x

## Test Results<sup>3</sup>

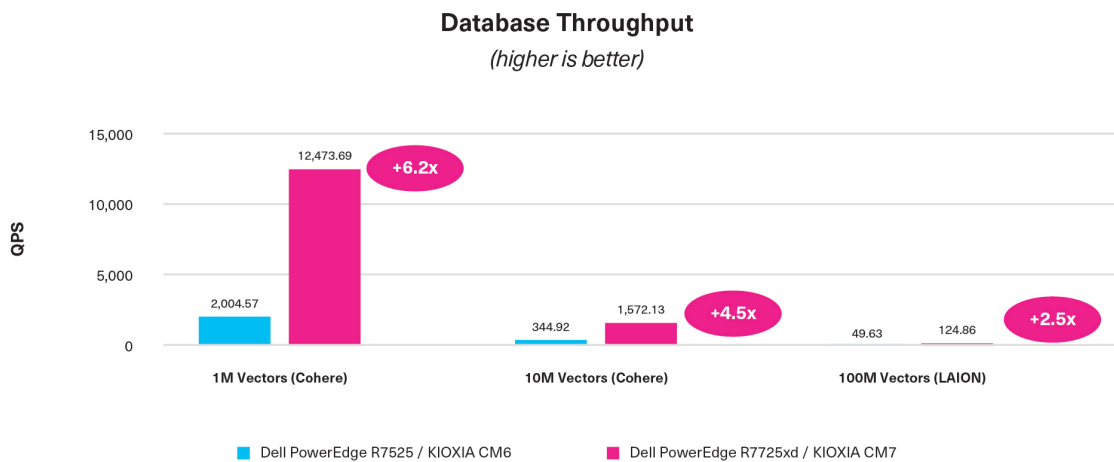
### Metric 1: Load Duration

This metric measured the total time required for the Milvus® vector DB to ingest data and construct the necessary search indexes which is crucial for evaluating how quickly the server/SSD platform can be prepared for production queries. Lower load durations equate to faster times to either construct the vector index initially or reconstruct vector indexes if new additional data needs to be considered. The results are in seconds, and the lower results for each dataset are better.



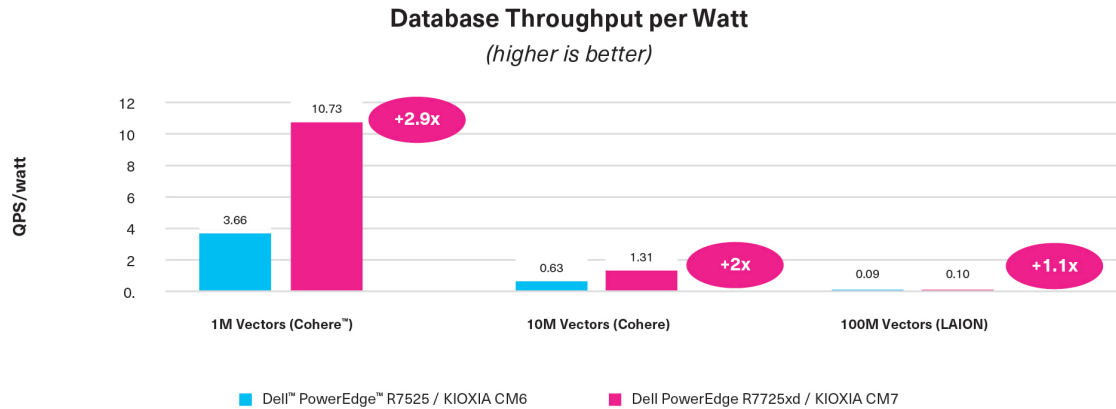
### Metric 2: Database Throughput

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 hardware resources are working more efficiently and can scan through more vectors in the index, returning the results of the similarity search faster. The results are in QPS, and the higher results for each dataset are better.



**Metric 3: Database Throughput per Watt**

This metric measured the energy efficiency of the compute hardware; specifically, how much performance a system delivers for every watt of power it consumes. In an era where energy costs and environmental concerns are paramount, understanding this metric is essential for businesses and data centers. The results show the database throughput in QPS achieved per watt drawn by each server/SSD platform. The results are in QPS/watt, and the higher results for each dataset are better.

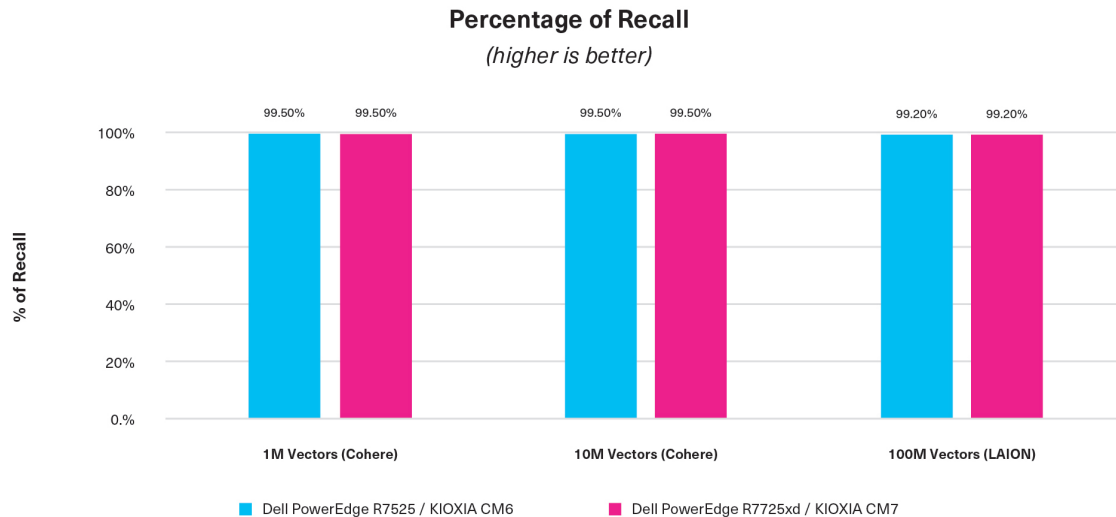


**Metric 4: Percentage of Recall**

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 calculated as follows:

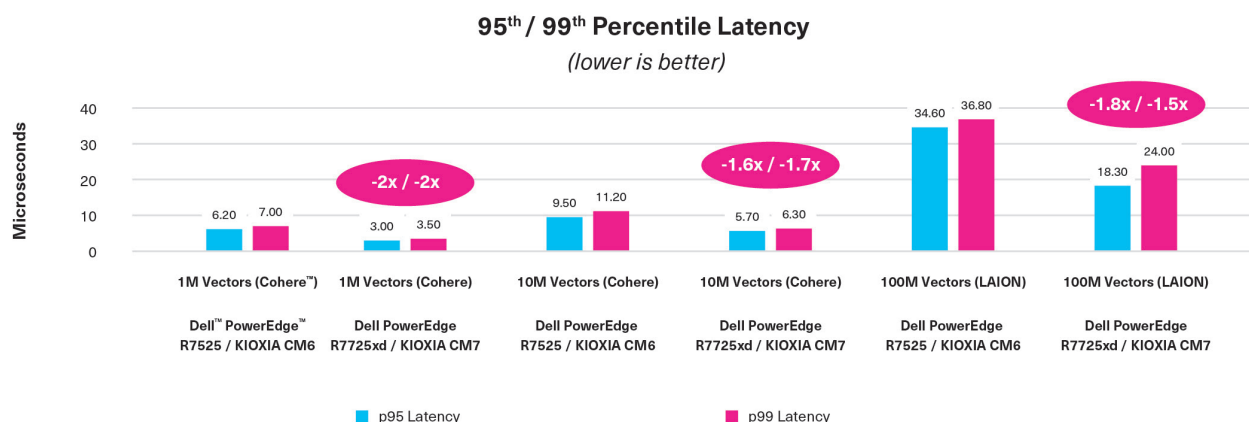
$$\frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}}$$

The recall shows the proportion of actual positives that were identified correctly. From the calculations, the higher the recall, the more precise and similar the results will be when vector data is returned to the vector DB. The results are in % of recall, and the higher results for each dataset are better.



**Metric 5: 95<sup>th</sup> / 99<sup>th</sup> Percentile Latency**

This metric measured the maximum time taken for the top 95% or 99% of vector search requests, respectively. These metrics are crucial for identifying tail latency, revealing how long the worst 5% or 1% of queries take, which often indicates system bottlenecks or performance spikes. The results are in microseconds, and the lower results for each dataset are better.



**Analysis**

The PCIe® 5.0 platform comprised of a Dell PowerEdge R7725xd server and a KIOXIA CM7-R Series SSD demonstrated better performance at various vector database sizes with regards to lower times for the Milvus® vector DB to ingest data and construct the necessary search indexes. It delivered a higher database throughput and database throughput per watt, as well as lower latencies when compared to the PCIe 4.0 platform as follows:

Metric	PCIe 4.0 Platform	PCIe 5.0 Platform	PCIe 5.0 Gains
<b>Load Duration</b> (in seconds) <i>(lower is better)</i>			
• 1M dataset	3,050.51	2,110.64	1.4x
• 10M dataset	37,209.66	20,895.75	1.7x
• 100M dataset	299,017.89	207,189.75	1.4x
<b>Database Throughput</b> (in QPS) <i>(higher is better)</i>			
• 1M dataset	2,004.57	12,473.69	6.2x
• 10M dataset	344.92	1,572.13	4.5x
• 100M dataset	49.63	124.86	2.5x
<b>Database Throughput / Watt</b> (in QPS/watt) <i>(higher is better)</i>			
• 1M dataset	3.66	10.73	2.9x
• 10M dataset	0.63	1.31	2x
• 100M dataset	0.09	0.10	1.1x
<b>Percentage of Recall</b> (in % of Recall) <i>(higher is better)</i>			
• 1M dataset	99.5	99.5	0
• 10M dataset	99.5	99.5	0
• 100M dataset	99.2	99.2	0
<b>95<sup>th</sup> Percentile Latency</b> (in microseconds) <i>(lower is better)</i>			
• 1M dataset	6.2	3	2x
• 10M dataset	9.5	5.7	1.6x
• 100M dataset	34.6	18.3	1.8x
<b>99<sup>th</sup> Percentile Latency</b> (in microseconds) <i>(lower is better)</i>			
• 1M dataset	7	3.5	2x
• 10M dataset	11.2	6.3	1.7x
• 100M dataset	36.8	24	1.5x

## Summary

The PCIe® 5.0 server/SSD platform was able to provide faster times for the Milvus® vector DB to ingest data and construct the necessary search indexes, as well as higher database throughput and database throughput per watt, and lower latencies with similar and very high recall when compared to the PCIe 4.0 platform. The tests verified that supporting vector DBs for LLMs, the latest generation storage with PCIe 5.0 throughput and low latency is essential for processing more data.

## KIOXIA CM7 Series SSD Product Info

The KIOXIA CM7 Series enterprise NVMe™ SSDs support EDSFF E3.S and 2.5-inch form factors and are compliant with the PCIe 5.0 and NVMe 2.0 specifications. These SSDs are available in two configurations: KIOXIA CM7-R Series for read-intensive applications (1 DDPD<sup>4</sup>, up to 30.72 terabyte<sup>5</sup> (TB) capacities) and KIOXIA CM7-V Series for higher endurance, mixed-use applications (3 DDPD, up to 12.8 TB capacities). Additional features include a dual-port design for high availability applications, flash die failure protection and security options<sup>6</sup>. Additional KIOXIA CM7 Series SSD information is available [here](#).



*KIOXIA CM7 Series SSDs<sup>7</sup>*

## Dell™ PowerEdge™ 7725xd Product Info

The Dell PowerEdge 7725xd is a storage-dense, two-socket, 2U, 17<sup>th</sup> generation, air-cooled rack server designed for extreme storage density and high-performance workloads like data analytics, high-performance compute (HPC) and high bandwidth object-scale. It supports dual 5<sup>th</sup> generation AMD EPYC™ 9005 processors (up to 192 cores each) and 24 front-mounted Gen5 NVMe drives. Additional Dell PowerEdge 7725xd server information is available [here](#).



*Dell PowerEdge 7725xd server<sup>7</sup>*

## Appendix A

## Hardware/Software Test Configuration

Server Information		
Platform	PCIe® 5.0	PCIe 4.0
Vendor/Model	Dell™ PowerEdge™ R7725xd	Dell PowerEdge R7525
No. of Servers	1	1
BIOS Version	1.5.3	2.22.0

CPU Information		
Vendor/Model	AMD EPYC™ 9555	AMD EPYC 7552
No. of CPU Sockets	2	2
No. of CPU Cores	64	48
CPU Frequency	3,200 MT/s*	2,200 MT/s*

\*MT/s = megatransfers per second

Memory Information		
Memory Type	DDR5	DDR4
Memory Speed	6,400 MT/s*	2,933 MT/s*
Memory Size	32 GB^	16 GB^
No. of DIMMs	24	16
Total Memory	768 GB^	256 GB^

\*MT/s = megatransfers per second

^GB = gigabytes\*

SSD Information		
Model	KIOXIA CM7-R Series	KIOXIA CM6-R Series
Interface	PCIe 5.0 x4	PCIe 4.0 x4
No. of Drives	1	1
Form Factor	2.5-inch	2.5-inch
Capacity	3.84 TB	3.84 TB
DWPD	1 (5 years)	1 (5 years)

OS Information - Both Platforms		
Operating System	Ubuntu®	
Version	24.04.3 LTS	
Kernel	6.8.0-90-generic	

File System Information - Both Platforms		
File System	XFS®	
Mount Options	noatime, nodiratime	

Vector DB Information - Both Platforms		
Vector Database	Milvus®	
Milvus Version	2.6.9	
MinIO® Version	RELEASE.2024-12-18T13-15-44Z	
etcd® Version	3.5.25	

Load Generator Information - Both Platforms		
Load Generator	VectorDBBench	
Version	1.0.18	

## Appendix B

### Configuration Setup/Test Procedures

#### Configuration Setup

A Dell™ PowerEdge™ R7725xd server was set up with an Ubuntu® 24.04.3 LTS operating system.

One 3.84 TB KIOXIA CM7-R Series SSD was installed into the server.

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

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

The VectorDBBench load generator was installed and used to run a search performance test for various vector DB sizes. Two Cohere™ datasets and one LAION dataset were downloaded and inserted into the Milvus vector DB. The Cohere datasets consisted of 1 million and 10 million vectors, while the LAION dataset consisted of 100 million vectors, and all datasets utilized 768 dimensions per vector.

A DiskANN index was created to run the test metrics.

#### Test Procedures

The VectorDBBench Search Performance Test 768 Dim was run for 1, 10 and 100 million vectors utilizing the DiskANN index and the following metrics were gathered:

- *Load Duration (in seconds)*
- *Database Throughput (in QPS)*
- *Database Throughput per Watt (in QPS/watt)*
- *Percentage of Recall (in % of Recall)*
- *95<sup>th</sup> / 99<sup>th</sup> Percentile Latency (in microseconds)*

The tests above were then repeated with a DiskANN index for the PCIe® 4.0 platform consisting of one Dell PowerEdge R7525 using one KIOXIA CM6-R Series SSD.

The metrics of the two platforms were compared.

**NOTES:**

<sup>1</sup> 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;

<sup>2</sup> 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.

<sup>3</sup> Read and write speed may vary depending on various factors such as host devices, software (drivers, OS, etc.) and read/write conditions.

<sup>4</sup> 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.

<sup>5</sup> 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 1 Gbit = 2<sup>30</sup> bits = 1,073,741,824 bits, 1GB = 2<sup>30</sup> bytes = 1,073,741,824 bytes, 1TB = 2<sup>40</sup> bytes = 1,099,511,627,776 bytes and 1PB = 2<sup>50</sup> 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.

<sup>6</sup> Optional security feature compliant drives are not available in all countries due to export and local regulations.

<sup>7</sup> The product images shown are representations of design models and are not accurate product depictions.

**TRADEMARKS:**

AMD EPYC and combinations thereof are trademarks of Advanced Micro Devices, Inc. Cohere is a trademark of Cohere Communications, LLC. Dell and PowerEdge are trademarks of Dell Inc. or its subsidiaries. etcd is a registered trademark of The Linux Foundation in the U.S. and other countries. Milvus is a registered trademark of LF Projects, LLC. MinIO is a registered trademark of MinIO Corporation. 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. 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.

© 2026 KIOXIA America, Inc. All rights reserved.