KIOXIA

Press Release

KIOXIA Achieves 4.8 Billion High-Dimensional Vector Search Database on a Single Server, with 7.8x Index Build Time Acceleration via GPUs

Leveraging the NVIDIA cuVS Library with KIOXIA AiSAQ™ Technology to Index Vectors of 1024 Dimensions with Minimal DRAM Use

Germany, Düsseldorf, 16 March 2026 – KIOXIA Europe GmbH today announced KIOXIA Corporation's successful demonstration of achieving high-dimensional vector search scaling to 4.8 billion vectors on a single server with its open-source KIOXIA AiSAQ™ approximate nearest neighbor search (ANNS) technology. Additionally, KIOXIA demonstrated a significant reduction in index build time by leveraging GPU acceleration through the NVIDIA cuVS. These two achievements mark a significant advancement for retrieval augmented generation (RAG) search solutions. Continued development is underway to support larger-scale deployments beyond 4.8 billion vectors.

Index build time on a massive-scale vector database is a crucial pain point for the industry. In collaboration with NVIDIA, KIOXIA demonstrated up to 20x improvement in KIOXIA AiSAQ™ index build time for high-dimensional vectors of 1024 dimensions, and up to 7.8x improvement in end-to-end build times. This 20x improvement represents a reduction from 28.4 days using CPU to 1.4 days using four NVIDIA Hopper GPUs to build the index, and a reduction from 31 days to 4 days in end-to-end testing.[1]

AI applications may now rely on larger volumes of vectorized information reaching tens of billions of vectors and beyond stored on SSDs, while DRAM alone becomes impractical even at a billion scale. KIOXIA enables a highly scalable storage architecture with KIOXIA AiSAQ™ technology by achieving billion-scale search, exceeding RAG application latency requirements using a single query server in a Milvus vectorDB environment powered by GPU acceleration on index builds that make large scale deployments practical.

"Vector databases provide a backbone for applications that need to understand intent, context, and similarity across massive, unstructured datasets in real time," said Jason Hardy, Vice President, Storage Technologies, NVIDIA. "By leveraging GPU-accelerated indexing with the NVIDIA cuVS library, KIOXIA supports high-dimensional vector databases that can scale and build indexes with unprecedented efficiency."

[First announced last year](#), KIOXIA AiSAQ™ open-source software technology addresses RAG scalability challenges by enabling vector search directly from SSDs, with reduced DRAM usage. KIOXIA AiSAQ™ technology provides high scalability, making it well-suited for both multi-tenant environments and large-scale monolithic index deployments. The technology leverages an innovative Global Index algorithm that combines hybrid clustering and graph search to deliver efficient vector search at extreme scale. With flexible tuning options to balance performance and high-volume vector scalability, KIOXIA AiSAQ™ software makes large-scale deployments more accessible and easier to expand.

"Achieving scalable vector search at the multi-billion scale on a single server is a major milestone for the industry," said Axel Stoermann, Chief Technology Officer & Vice President, KIOXIA Europe GmbH. "With KIOXIA AiSAQ™ SSD-based vector search and GPU-accelerated indexing with the NVIDIA cuVS, we are reducing index build times from weeks to days – a critical step toward making large-scale RAG systems operational."

KIOXIA remains committed to advancing storage-driven AI solutions that support intelligent data processing at scale and continues to evolve KIOXIA AiSAQ™ toward trillion-vector deployments.

Link to download KIOXIA AiSAQ™ open-source software: https://github.com/kioxia-jp/aisaq-diskann.

Notes:

1: A total of 19.66 TB of vector data was processed for this benchmark. Performance or benchmark results may vary depending on the host device, read and write conditions, data sizes and other factors.

KIOXIA AiSAQ is a trademark of KIOXIA.

Company names, product names, and service names may be trademarks of third-party companies.

About KIOXIA Europe GmbH

KIOXIA Europe GmbH is the European-based subsidiary of KIOXIA Corporation, a leading worldwide supplier of flash memory and solid-state drives (SSDs). From the invention of NAND flash memory to today's renowned BiCS FLASH™ 3D flash memory KIOXIA continues to pioneer innovative memory solutions and services that enrich people's lives and expand society's horizons. The company's innovative BiCS FLASH™ 3D flash memory technology is shaping the future of storage in high-density applications, including advanced smartphones, PCs, automotive systems, data centers and generative AI systems.

Visit our KIOXIA website

Contact details for publication:
KIOXIA Europe GmbH, Hansaallee 183, 40549 Düsseldorf, Germany
Tel: +49 (0)211 368 77-0
E-mail:KIE-support@eu.kioxia.com

Contact details for editorial enquiries:

Lena Hoffmann, KIOXIA Europe GmbH

Tel: +49 (0) 211 36877 382

E-mail: lena.hoffmann@eu.kioxia.com

Issued by:

Birgit Schöniger, Pretzl GmbH

Tel: +49 (0)172 617 8431

E-mail: birgit.schoeniger@pretzl.com

Web: www.pretzl.com