



Press Release

KIOXIA schafft 4,8 Milliarden hochdimensionale Vektorsuchen auf einem Server

Die cuVS-Bibliothek von NVIDIA beschleunigt in Verbindung mit der AiSAQ™-Technologie von KIOXIA die Indexerstellung durch GPUs um das 7,8-Fache

Düsseldorf, 19. März 2026 – [KIOXIA Europe GmbH](#) hat die erfolgreiche Demonstration der KIOXIA Corporation bekannt gegeben, die zeigt, dass hochdimensionale Vektorsuchen auf 4,8 Milliarden Vektoren auf einem einzigen Server skaliert werden können. Grundlage dafür ist die Open-Source-Technologie KIOXIA AiSAQ™ für Approximate Nearest Neighbor Search (ANNS). Darüber hinaus konnte das Unternehmen durch GPU-Beschleunigung mit Hilfe der [cuVS-Bibliothek](#) von NVIDIA die Zeiten für die Indexerstellung signifikant verkürzen. Diese beiden Erfolge stellen einen wichtigen Fortschritt für RAG (Retrieval Augmented Generation)-Suchlösungen dar. Die weitere Entwicklung zielt nun darauf ab, auch größere Deployments jenseits von 4,8 Milliarden Vektoren zu unterstützen.

Die Indexerstellung in großen Vektordatenbanken stellt für die Branche einen zentralen Engpass dar. In Zusammenarbeit mit NVIDIA konnte KIOXIA eine bis zu 20-fache Verbesserung der Indexierungszeiten bei Verwendung der AiSAQ™-Technologie für hochdimensionale Vektoren mit 1.024 Dimensionen erzielen und die Ende-zu-Ende Build-Zeiten um das bis zu 7,8-Fache reduzieren. Konkret bedeutet das eine Reduktion von 28,4 Tagen auf 1,4 Tage für die Indexerstellung mit vier [NVIDIA Hopper GPUs](#) und eine Verringerung der Gesamtdauer von 31 auf 4 Tage.¹

KI-Anwendungen können nun auf größere Mengen vektorisierter Informationen zugreifen, die auf SSDs gespeichert sind und mehrere zehn Milliarden Vektoren umfassen, da DRAM allein selbst im Milliardenbereich unpraktikabel wird. KIOXIA AiSAQ™ ermöglicht eine hochskalierbare Speicherarchitektur, die Milliarden von Suchvorgängen unterstützt, die Latenzanforderungen von RAG-Anwendungen erfüllt und durch GPU-beschleunigte Indexerstellung auf einem einzigen Server große Bereitstellungen realisierbar macht, etwa

in einer Milvus-VektorDB-Umgebung.

AiSAQ™: Quelloffene Skalierung für RAG

[Erstmals im vergangenen Jahr vorgestellt](#), adressiert die Open-Source-Software KIOXIA AiSAQ™ Skalierungsherausforderungen bei RAG-Anwendungen, indem Vektorsuchen direkt von SSDs durchgeführt werden können und der DRAM-Bedarf reduziert wird. Die Technologie bietet hohe Skalierbarkeit, eignet sich für Multi-Tenant-Umgebungen ebenso wie für monolithische Index-Deployments im großen Maßstab. KIOXIA AiSAQ™ nutzt einen innovativen Global-Index-Algorithmus, der hybride Cluster-Bildung und Graph-Suche kombiniert, um effiziente Vektorsuchen im Extremmaßstab zu ermöglichen. Mit flexiblen Anpassungsoptionen zur Balance zwischen Leistung und Skalierbarkeit großer Vektorvolumina macht die Software hochskalierte Deployments leichter zugänglich und einfacher erweiterbar.

„Vektordatenbanken bilden das Rückgrat für KI-Anwendungen, die Intention, Kontext und Ähnlichkeit in riesigen, unstrukturierten Datensätzen in Echtzeit erkennen müssen“, erklärt Jason Hardy, Vice President, Storage Technologies bei NVIDIA. „Durch die Nutzung GPU-beschleunigter Indexierung mit der cuVS-Bibliothek von NVIDIA unterstützt KIOXIA hochdimensionale Vektordatenbanken, die sowohl skalierbar als auch effizient in der Indexerstellung sind.“

„Die Skalierung von Vektorsuchen im Milliardenbereich auf einem einzigen Server ist ein Meilenstein für die Branche“, betont Axel Störmann, Chief Technology Officer & Vice President bei KIOXIA Europe. „Mit der SSD-basierten KIOXIA-AiSAQ™-Vektorsuche und GPU-beschleunigter Indexierung durch NVIDIA cuVS reduzieren wir die Zeiten für die Indexerstellung von Wochen auf Tage – ein entscheidender Schritt, um RAG-Systeme im großen Maßstab effektiv zu betreiben.“

KIOXIA setzt sich weiterhin dafür ein, KI-Lösungen mit speichergetriebener Architektur voranzutreiben, die intelligente Datenverarbeitung im großen Maßstab unterstützen, und entwickelt KIOXIA AiSAQ™ kontinuierlich weiter, um künftig sogar Billionen-Vektor-Deployments zu ermöglichen.

Die Open-Source-Software KIOXIA AiSAQ™ steht hier zum Download bereit:
<https://github.com/kioxia-jp/aisaq-diskann>

###

Anmerkungen:

1: Insgesamt wurden 19,66 TB an Vektordaten für diesen Benchmark verarbeitet. Die Performance oder Benchmark-Ergebnisse können je nach Host Device, Lese- und Schreibkonditionen, Datengrößen und anderen Faktoren variieren.

KIOXIA AiSAQ ist eine geschützte Marke von KIOXIA.

Die folgenden Marken, Services und/oder Unternehmensnamen – NVIDIA, NVIDIA Corporation – wurden von der KIOXIA Europe GmbH oder verbundenen Unternehmen der KIOXIA-Gruppe weder verwendet, eingetragen, erstellt noch gehalten. Sie können jedoch von Dritten in verschiedenen Gerichtsbarkeiten beantragt, eingetragen oder erstellt worden sein und/oder Eigentum Dritter sein und sind daher vor unbefugter Nutzung geschützt.

Über KIOXIA Europe GmbH

KIOXIA Europe GmbH ist die europäische Tochtergesellschaft der KIOXIA Corporation, einem weltweit führenden Anbieter von Flashspeichern und Solid State Drives (SSDs). Von der Erfindung des NAND-Flashspeichers bis hin zum heutigen, renommierten BiCS FLASH™ 3D Flashspeicher ist KIOXIA ein Pionier auf dem Gebiet innovativer Speicherlösungen und -dienste, die das Leben der Menschen bereichern und der Gesellschaft neue Perspektiven eröffnen. KIOXIAs innovative BiCS FLASH 3D™ Flashspeicher-Technologie prägt die Zukunft von Speichern in Anwendungen mit hoher Speicherdichte, darunter moderne Smartphones, PCs, Fahrzeugsysteme, Rechenzentren und GenAI-Systeme.

Weitere Informationen auf der [KIOXIA-Website](#).

Weitere Informationen:

KIOXIA Europe GmbH
Hansaallee 183
40549 Düsseldorf
Tel: +49 (0)211 368 77-0
E-Mail: KIE-support@eu.kioxia.com

Pressekontakte:

Hanna Greve, PR-COM GmbH
Tel: +49 (0) 89 59997 756
E-Mail: hanna.greve@pr-com.de

Lena Hoffmann, KIOXIA Europe GmbH

Tel: +49 (0) 211 36877 382

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