



Press Release

KIOXIA veröffentlicht AiSAQ-Technologie zur Reduzierung des DRAM-Bedarfs in generativen KI-Systemen als Open-Source-Software

Die Softwaretechnologie verbessert die Skalierung und Genauigkeit von Vektordatenbanken in RAG-Workflows durch die Nutzung von SSDs

Düsseldorf, 29. Januar 2025 – [KIOXIA](#) kündigte heute die Veröffentlichung seiner neuen „All-in-Storage ANNS with Product Quantization“-Technologie (AiSAQ) als [Open Source](#) an. Mit einem neuartigen Algorithmus für „Approximate Nearest Neighbor Search“ (ANNS), der für SSDs optimiert wurde, liefert KIOXIA AiSAQ¹ eine skalierbare Leistung für Retrieval-Augmented Generation (RAG), ohne Indexdaten im DRAM abzulegen – die Suche erfolgt stattdessen direkt auf SSDs.

Generative KI-Systeme erfordern erhebliche Rechen-, Arbeitsspeicher- und Storage-Ressourcen. Sie haben das Potenzial, transformative Durchbrüche in verschiedenen Branchen voranzutreiben, doch ihr Einsatz ist oft mit hohen Kosten verbunden. RAG ist eine kritische Phase bei der KI-Anpassung, in der große Sprachmodelle (LLMs) mit unternehmens- oder anwendungsspezifischen Daten verfeinert werden.

Eine zentrale Komponente von RAG ist eine Vektordatenbank, die spezifische Daten zusammenführt und deren Eigenschaften in Vektoren innerhalb der Datenbank umwandelt. RAG verwendet auch einen ANNS-Algorithmus, der Vektoren identifiziert, die das Modell auf Basis der Ähnlichkeit zwischen den akkumulierten und den Zielvektoren verbessern. Damit RAG effektiv ist, muss es schnell die Informationen abrufen, die für eine Abfrage am relevantesten sind.



Üblicherweise werden ANNS-Algorithmen im DRAM eingesetzt, um die für diese Suchvorgänge erforderliche hohe Leistung zu erreichen.

Die KIOXIA-AiSAQ-Technologie bietet eine skalierbare und effiziente ANNS-Lösung für Milliarden von Datensätzen mit minimalem Speicherbedarf und schnellem Index Switching.

Die wichtigsten Vorteile der KIOXIA-AiSAQ-Technologie:

- Ermöglicht den Betrieb großer Datenbanken, ohne auf limitierte DRAM-Ressourcen angewiesen zu sein, und verbessert dadurch die Leistung von RAG-Systemen.
- Eliminiert die Notwendigkeit, Indexdaten in den DRAM zu laden, sodass die Vektordatenbank sofort gestartet werden kann. Dies ermöglicht einen nahtlosen Wechsel zwischen benutzer- oder anwendungsspezifischen Datenbanken auf demselben Server für eine effiziente Bereitstellung von RAG-Services.
- Optimiert für Cloud-Systeme durch das Speichern von Indizes in disaggregiertem Speicher für die gemeinsame Nutzung durch mehrere Server. Dieser Ansatz passt die Suchleistung der Vektordatenbank dynamisch für bestimmte Benutzer oder Anwendungen an und erleichtert die schnelle Migration von Suchinstanzen zwischen physischen Servern.

„KIOXIA AiSAQ ebnet den Weg für eine nahezu unendliche Skalierung von RAG-Anwendungen in generativen KI-Systemen, die im Kern auf Flash-basierten SSDs aufbauen“, sagt Axel Störmann, VP and Chief Technology Officer of Memory and SSD products bei KIOXIA Europe. „Durch den Einsatz von SSD-basiertem ANNS reduzieren wir die Abhängigkeit von teurem DRAM und erfüllen gleichzeitig die Leistungsanforderungen führender In-Memory-Lösungen – was den Leistungsbereich großer RAG-Anwendungen erheblich erweitert.“

KIOXIA unterstreicht sein Engagement für die Weiterentwicklung von KI, indem es seine innovative [KIOXIA-AiSAQ-Technologie als Open-Source-Software](#) in die Community einbringt.

**Anmerkungen:**

¹ KIOXIA AiSAQ, All-in-Storage ANNS with Product Quantization, eine neuartige Methode für das Ablegen von Indexdaten, ist eine Marke von KIOXIA.

Alle anderen Firmennamen, Produktnamen und Dienstleistungsamen können Marken von Drittunternehmen sein.

Über KIOXIA

KIOXIA ist ein weltweit führender Anbieter von Speicherlösungen, der sich auf die Entwicklung, die Produktion und den Vertrieb von Flash-Speichern und Solid State Drives (SSDs) spezialisiert hat. Im April 2017 wurde das Vorgängerunternehmen Toshiba Memory aus der Toshiba Corporation ausgegliedert – dem Unternehmen, das 1987 den NAND-Flash-Speicher erfand. KIOXIA hat es sich zur Aufgabe gemacht, die Welt mit „Speicher“ zu verbessern, indem es Produkte, Dienstleistungen und Systeme anbietet, die den Kunden Auswahlmöglichkeiten und der Gesellschaft einen speicherbasierten Mehrwert bieten. KIOXIAs innovative BiCS FLASH™ 3D-Flash-Speichertechnologie gestaltet die Zukunft von Speicher in Anwendungen mit hoher Speicherdichte, darunter moderne Smartphones, PCs, Automobilsysteme, Rechenzentren und generative KI-Systeme.

Weitere Informationen finden Sie unter <https://www.kioxia.com/de>

Weitere Informationen:

KIOXIA Europe GmbH
Hansaallee 181
40549 Düsseldorf
Tel: +49 (0)211 368 77-0
E-Mail: KIE-support@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: lana1.hoffmann@Kioxia.com