KIOXIA

Pressmeddelande

KIOXIA AiSAQ™-tekniken utformad för att minska behoven av DRAM i generativa AI-system släpps som programvara med öppen källkod

Programvaruteknik förbättrar skalning av vektordatabaser och exakthet i arbetsflöden i RAG med hjälp av SSD-enheter.

Tyskland, Düsseldorf, 29 januari 2025 – Idag tillkännagav KIOXIA lanseringen av öppen källkod för den nya AiSAQ-tekniken (All-in-Storage ANNS with Product Quantization) En ny ANNS-algoritm (Approximate Nearest Neighbor) optimerad för SSD-enheter, ger KIOXIA AiSAQ™(1)-programvara skalbar prestanda för RAG (Retrieval-Augmented Generation) utan att placera indexdata i DRAM – och söker istället direkt på SSD-enheter.

Generativa AI-system kräver betydande beräknings-, minnes- och lagringsresurser. Även om de har potential att driva transformativa genombrott inom olika branscher blir deras implementering ofta mycket kostsam. RAG är en kritisk fas av AI som förfinar stora språkmodeller (LLM:er) med data som är specifika för företaget eller programmet.

En central komponent i RAG är en vektordatabas som ackumulerar och omvandlar specifik data till funktionsvektorer i databasen. RAG använder också en ANNS-algoritm, som identifierar vektorer som förbättrar modellen baserat på likhet mellan de ackumulerade vektorerna och målvektorerna. För att RAG ska vara effektivt måste det snabbt hämta den information som är mest relevant för en fråga. Traditionellt används ANNS-algoritmer i DRAM för att uppnå den höghastighetsprestanda som krävs för dessa sökningar.

KIOXIA AiSAQ™-tekniken ger en skalbar och effektiv ANNS-lösning för datamängder i miljardstorlek med försumbar minnesanvändning och snabba indexväxlingsmöjligheter.

Viktiga fördelar med KIOXIA AiSAQ-teknik™:

- Gör det möjligt för storskaliga databaser att fungera utan att förlita sig på begränsade DRAM-resurser, vilket förbättrar prestandan hos RAG-system.
- Eliminerar behovet av att ladda indexdata i DRAM, vilket gör att vektordatabasen kan startas direkt. Detta stöder sömlös växling mellan användarspecifika eller applikationsspecifika databaser på samma server för effektiv leverans av RAG-tjänster.
- Optimerad för molnsystem genom att lagra index i disaggregerad lagring för delning över flera servrar. Den här metoden justerar dynamiskt sökprestanda för vektordatabaser för specifika användare eller program och underlättar snabb migrering av sökinstanser mellan fysiska servrar.

"KIOXIA AiSAQ™-lösningen banar väg för nästan oändlig skalning av RAG-applikationer i generativa AI-system baserade på flashbaserade SSD-enheter i kärnan", säger Axel Stoermann, Chief Technology Officer & VP för KIOXIA Europe GmbH. ”Genom att använda SSD-baserad ANNS minskar vi beroendet av kostsam DRAM och matchar prestandabehoven hos ledande minneslösningar – vilket avsevärt förbättrar prestandaomfånget för storskaliga RAG-applikationer.”

KIOXIA visar sitt engagemang för att främja AI genom att bidra med sin innovativa KIOXIA AiSAQ-teknik till samhället som programvara med öppen källkod.

# # #

Anmärkningar:

1: KIOXIA AiSAQ: All-in-Storage ANNS with Product Quantization, en ny metod för att placera indexdata, är ett varumärke som tillhör KIOXIA.

Om KIOXIA

KIOXIA är en världsledande leverantör av minneslösningar, med fokus på utveckling, produktion och försäljning av flashminnen och SSD-enheter (Solid State Drive). I april 2017 knoppades föregångaren Toshiba Memory av från Toshiba Corporation, det företag som uppfann NAND-flashminnet 1987. KIOXIA har åtagit sig att lyfta världen med "minne" genom att erbjuda produkter, tjänster och system som skapar valmöjligheter för kunderna och minnesbaserat värde för samhället. KIOXIAs innovativa 3D-flashminnesteknik, BiCS FLASH™, formar framtiden för lagring i högdensitetsapplikationer, inklusive avancerade smartphones, PCs, fordonssystem, datacenter och generativa AI-system.

Läs mer på KIOXIA:s webbplats

Kontaktuppgifter för publicering:
KIOXIA Europe GmbH, Hansaallee 181, 40549 Düsseldorf, Tyskland
Tel: +49 (0)211 368 77-0
E-post: KIE-support@kioxia.com

Kontaktuppgifter för redaktionella förfrågningar:
Lena Hoffmann, KIOXIA Europe GmbH
Tel: +49 (0) 211 36877 382
E-mail: lena1.hoffmann@kioxia.com

Publicerat av:
Birgit Schöniger, Publitek
Tel: +49 (0)172 617 8431
E-mail: birgit.schoeniger@publitek.com
Webb: www.publitek.com