



Comunicato stampa

KIOXIA raggiunge un database di ricerca vettoriale ad alta dimensionalità da 4,8 miliardi di vettori su un singolo server, con un'accelerazione fino a 7,8x dei tempi di costruzione dell'indice tramite GPU

Sfruttare la libreria NVIDIA cuVS con la tecnologia AiSAQ™ di KIOXIA per indicizzare vettori a 1024 dimensioni con un uso minimo della DRAM

Germania, Düsseldorf, 16 marzo 2026 – [KIOXIA Europe GmbH](#) ha annunciato oggi che KIOXIA Corporation ha dimostrato con successo la scalabilità della ricerca vettoriale ad alta dimensionalità fino a 4,8 miliardi di vettori su un singolo server, utilizzando la tecnologia open source KIOXIA AiSAQ™ per la ricerca approssimata dei vicini più prossimi (ANNS). Inoltre, KIOXIA ha dimostrato una significativa riduzione del tempo di compilazione dell'indice sfruttando l'accelerazione della GPU tramite il [NVIDIA cuVS](#). Questi due risultati rappresentano un significativo progresso per le soluzioni di ricerca a generazione aumentata da recupero (RAG). Sono in corso ulteriori sviluppi per supportare implementazioni su scala ancora più ampia, oltre i 4,8 miliardi di vettori.

Il tempo di costruzione dell'indice su un database vettoriale su larga scala rappresenta un punto critico cruciale per il settore. In collaborazione con NVIDIA, KIOXIA ha dimostrato un miglioramento fino a 20 volte nel tempo di costruzione dell'indice AiSAQ™ di KIOXIA per vettori ad alta dimensionalità di 1024 dimensioni, e fino a 7,8 volte il miglioramento nei tempi complessivi end-to-end. Questo miglioramento di 20 volte rappresenta una riduzione da 28,4 giorni usando la CPU a 1,4 giorni usando quattro [GPU NVIDIA Hopper](#) per costruire l'indice e una riduzione da 31 giorni a 4 giorni nei test end-to-end.¹



Le applicazioni di IA possono ora basarsi su volumi più grandi di informazioni vettorializzate, raggiungendo decine di miliardi di vettori e oltre, memorizzati su SSD, mentre l'uso esclusivo della DRAM diventa impraticabile già su scala di miliardi. KIOXIA abilita un'architettura di archiviazione altamente scalabile con la tecnologia AiSAQ™ di KIOXIA ottenendo ricerche su scala miliardarie, superando i requisiti di latenza delle applicazioni RAG utilizzando un singolo server di query in un ambiente Milvus vectorDB alimentato dall'accelerazione GPU su build di indici che rendono pratiche le implementazioni su larga scala.

"I database vettoriali costituiscono la spina dorsale per applicazioni che devono comprendere intenti, contesto e somiglianza attraverso enormi dataset non strutturati in tempo reale", ha dichiarato Jason Hardy, Vicepresidente Storage Technologies di NVIDIA. "Sfruttando l'indicizzazione accelerata dalla GPU con la libreria NVIDIA cuVS, KIOXIA supporta database vettoriali ad alta dimensione che possono scalare e costruire indici con un'efficienza senza precedenti".

[Annunciata per la prima volta lo scorso anno](#), la tecnologia software open-source AiSAQ™ di KIOXIA affronta le sfide di scalabilità RAG abilitando la ricerca vettoriale direttamente dagli SSD, con un uso ridotto delle DRAM. La tecnologia AiSAQ™ offre un'elevata scalabilità, rendendola adatta sia per ambienti multi-tenant sia per implementazioni di indici monolitici su larga scala. La tecnologia sfrutta un innovativo algoritmo di Global Index che combina clustering ibrido e ricerca grafica per offrire una ricerca vettoriale efficiente su scala estrema. Con opzioni di tuning flessibili per bilanciare prestazioni e scalabilità vettoriale ad alto volume, il software AiSAQ™ di KIOXIA rende le implementazioni su larga scala più accessibili e più facili da espandere.

"Raggiungere la ricerca vettoriale scalabile su scala di miliardi su un unico server rappresenta un traguardo fondamentale per il settore", ha dichiarato Axel Stoermann, Chief Technology e Vice Presidente di KIOXIA Europe GmbH. "Con la ricerca vettoriale basata su SSD AiSAQ™ di KIOXIA e l'indicizzazione accelerata tramite GPU con NVIDIA cuVS, stiamo riducendo i tempi di compilazione dell'indice da settimane a giorni; un passo cruciale per rendere operativi i sistemi RAG su larga scala".



KIOXIA ribadisce il proprio impegno nello sviluppo di soluzioni AI basate sullo storage per supportare l'elaborazione intelligente dei dati su larga scala e continua a evolvere la tecnologia KIOXIA AiSAQ™ verso implementazioni con trilioni di vettori.

Segui il link per scaricare il software open-source KIOXIA AiSAQ™: <https://github.com/kioxia-jp/aisaq-diskann>.

###

Note:

1: Per questo benchmark sono stati processati in totale 19,66 TB di dati vettoriali. Le prestazioni o i risultati dei benchmark possono variare a seconda del dispositivo host, delle condizioni di lettura e scrittura, delle dimensioni dei dati e di altri fattori.

KIOXIA AiSAQ: è un marchio di KIOXIA.

I nomi di aziende, prodotti e servizi possono essere marchi di proprietà di terze parti.

Informazioni su KIOXIA Europe GmbH

KIOXIA Europe GmbH è la filiale europea di KIOXIA Corporation, leader mondiale nella fornitura di memorie flash e unità a stato solido (SSD). Dall'invenzione della memoria flash NAND all'attuale rinomata memoria flash 3D BiCS FLASH™, KIOXIA continua a essere pioniera di soluzioni e servizi di memoria innovativi che arricchiscono la vita delle persone e ampliano gli orizzonti della società. L'innovativa tecnologia di memoria flash 3D BiCS FLASH™ sta plasmando il futuro dell'archiviazione nelle applicazioni ad alta densità, tra cui smartphone avanzati, PC, sistemi automobilistici, data center e sistemi di intelligenza artificiale generativa.

Visita il nostro [sito web KIOXIA](#)

Contatti per la pubblicazione:

KIOXIA Europe GmbH, Hansaallee 183, 40549 Düsseldorf, Germania

Tel: +49 (0)211 368 77-0

E-mail: KIE-support@eu.kioxia.com



Dati di contatto per richieste redazionali:

Lena Hoffmann, KIOXIA Europe GmbH

Tel: +49 (0) 211 36877 382

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

Pubblicato da:

Birgit Schöniger, Pretzl GmbH

Tel: +49 (0)172 617 8431

E-mail: birgit.schoeniger@pretzl.com

Web: www.pretzl.com