KIOXIA

KIOXIA + AI

Powering AI Insight: Importance of Storage for Inference and Grounding

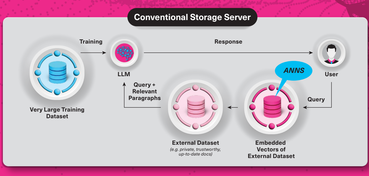
Storage is critical for each phase of AI

AI has varied storage requirements for each phase. Here is a deeper look at the inference and grounding (RAG) phases, storage needs, and solutions.

Inference and Grounding Requirements

Phase Description	Storage Requirements
Inference: A query is made on a pre-trained model and it predicts a response with newly generated information in real-time	Low latency, high throughput storage to load the model and keep the GPU busy.
Grounding / RAG: Continually improve the AI model without retraining. More vectors increase results accuracy. Uses a scalable database that can hold many millions of vectors, which can be stored in DRAM and/or SSDs.	Capacity to hold and scale the vector database. Low latency, random access and high IOPS to service retrieval requests.

Deeper Dive into Retrieval Augmented Generation (RAG)



- Retrieval Augmented Generation (RAG) is an approach that improves LLMs by giving them access to up to date or domain specific knowledge during inference, without the need to retrain
- RAG reduces hallucinations and improves accuracy by grounding answers in real-world data and trustworthy knowledge bases
- More vectors in the vector database increases accuracy of the results
- Approximate Nearest Neighbor Search (ANNS) is used and attempts to find vectors nearest to the query vector, and responds quickly

How to Improve RAG and Make AI More Effective?

Increase the vector database size and make it scalable

- Scale the database
- Use high capacity SSDs
- No DRAM slot limitation

Make vector database available to multiple servers

- Common storage available to multiple servers

Optimize storage use

- Tune storage for vector database performance or capacity

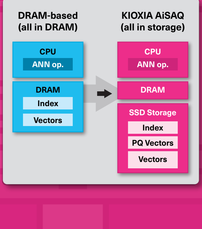
Reduce vector database load time

- Database stored on SSDs can be directly queried during inference
- No waiting for the database to load into DRAM

KIOXIA SSD Solutions for AI

	KIOXIA LC Series	KIOXIA CM Series	KIOXIA CD and XD Series
			
Features	• High Capacity • Dual-port	• Highest Performance • Lowest Latency • Dual-port	• High Performance • Low Latency
Recommended AI Phases	• Ingest • RAG	• Training • Inference	• Ingest • Training

KIOXIA AiSAQ™ Software for RAG



- Allows vector database components to be stored on SSDs
- Scalable with more SSDs
- No capacity limitation, like DRAM
- Accessible by multiple servers
- No need to load vector database into DRAM
- Open-source software available to all
- Integrated into Milvus open-source vector database