KIOXIA

Technical Brief

Why High-Capacity SSDs are Needed for AI

Delivering Storage and Performance Necessary for Tomorrow's AI Workloads

Artificial intelligence (AI) is technology that enables computer systems, using vast volumes of data, to perform tasks normally associated with human intelligence. Traditional AI responds to a set of inputs, while generative AI creates new data. As AI has quickly become an indispensable tool that is changing the world, its role will only grow more significant over time as it is poised to reshape and benefit many industries. See the addendum for a short list of key sectors using AI.

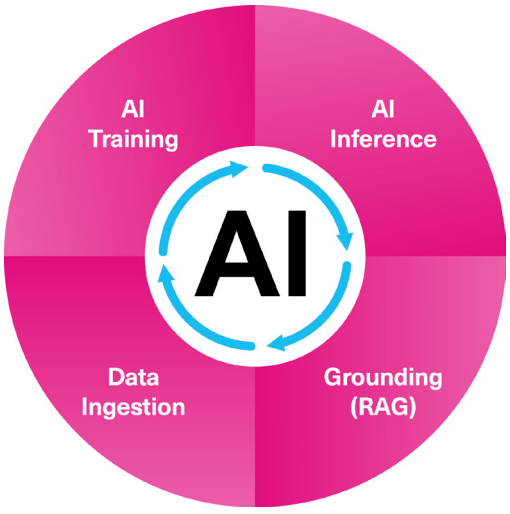
AI runs on systems as small as cell phones and as large as data centers that span multiple football fields. For these applications, the bigger the input datasets, the better the results (responses), and there is increasingly more data available to include in these input datasets. IT professionals are now concerned about building AI systems that will have enough storage to hold all this necessary data. The quest is to increase storage density by putting more data in a smaller space without exceeding the power budget of the overall system. Fortunately, SSDs are well suited to address this goal. With advances in flash memory technology that provides more bits per cell (such as quad-level cell or QLC), storage density can dramatically increase utilizing the same SSD form factor and power. It is in this arena where high-capacity SSDs excel.

Before AI, the sweet spots of SSD capacity deployed in application servers were 3.84 terabytes[1] (TB) and 7.68 TB. KIOXIA Corporation, in fact, has offered SSD capacities up to 30.72 TB for years, but it wasn't until the recent AI explosion that higher SSD capacities (such as 61.44 TB) started to be used on a larger scale. Recently, KIOXIA SSD capacities up to 122.88 TB have been announced to address the demands of AI and other data-intensive applications. AI is becoming pervasively used, so the need for even higher capacities will be required. Therefore, more recently, the company announced the industry's first PCIe® 5.0 245.76 TB NVMe™ SSD.

This technical brief introduces high-capacity SSDs for AI workloads that typically range from 61.44 TB to 245.76 TB. The next section addresses why AI requires so much storage.

Storage Requirements Associated with AI

In an AI system, there are process stages that perform different functions and have specific data workloads. These workloads put demands on core areas of the system, such as GPUs, CPUs, host memory and data storage. Today's data storage is in the form of SSDs that must fulfill the varied workload requirements depicted on the right:



Source: KIOXIA Corporation[2]

The following is a short list and description of the storage requirements for each AI workload stage:

AI Workload	*Function*	*Storage Requirements*
Data Ingestion	*Import vast amounts of raw data*	*High-sequential bandwidth*
Training	*Process data through AI algorithms* *Evaluate results* *Refine data for accuracy*	*Read-dominated training* *Write-dominated checkpointing*
Inference	*Model recognizes new patterns* *Extrapolate conclusions*	*Read-intensive dominated by high-volume small random reads*
Grounding (RAG)	*Model incorporates additional data specific to the use case* *Utilizes a vector database with a searchable repository*	*Mix of random reads and write*

Data Ingestion

The first stage of the AI workflow involves ingesting (data gathering). Ingested data can come from crawling the internet or from various internal sources such as sensors, databases and files. It's at this juncture that large language models (LLMs), an important area of generative AI, step in with their ability to transform workflows and train on large public datasets from natural and coding languages.

Once the data is collected, it needs to be prepared (pre-processed), which may involve cleaning the data, handling missing values, transforming data types and/or potentially augmenting the dataset. These actions may further introduce a need for data provenance requiring metadata tagging and tracking to ensure data quality and reproducibility.

Data volume during data ingest can be immense, especially for deep learning models that depend on large datasets. Storage at this stage needs to be scalable and cost-effective since the raw data may consist of unstructured or semi-structured data that requires reformatting and pre-processing, potentially growing the dataset further.

Training

The next stage of the AI workflow trains the model on the abundance of data collected, and at these stages, data security starts to become more critical as sensitive information may be present in the dataset. Compute resources cannot start to train a model until the data is pre-processed and fetched from the SSDs. Once the data has been prepared and cleaned, the compute resources go through the training cycle and consume data in batches to train the model. A model may go through many iterations on the same dataset so that the algorithms can learn the features they need to create an accurate model.

High-capacity SSDs are required to respond to multiple queries very quickly, and with low latency. If the SSDs lag, they become a bottleneck and compute resources may be underutilized. To keep the many GPU cores always busy, high-capacity SSDs must deliver high sustained and burst throughput so that the compute resources can execute the AI algorithm on the data. This training process can be very lengthy in order to optimize the model for a specific task using a smaller, task-specific dataset.

Inference

The next stage of the AI workload is inference or when a trained AI model applies its learned knowledge to new, unseen data to make predictions, recommendations or decisions. It's essentially the execute stage where the AI model is used in real-world applications to process input data, compare it with the information it learned during training, and generate an output response.

Grounding (RAG)

Retrieval Augmented Generation is a complementary AI technique that can also take full advantage of the benefits of high-capacity SSDs. The technique allows an existing LLM to incorporate additional data specific to the use case and utilizes a vector database with a searchable repository to find relevant context without the need for retraining or tuning.

The size of the vector databases and associated indexes are a growing concern in many RAG implementations as extremely large DRAM footprints are required to store them. In some cases, vectorized data and associated indexes could be much larger than data footprint size, requiring terabytes of DRAM to store them and a potentially costly solution given today's DRAM prices.

As an alternative to DRAM-based vector databases and indexes, KIOXIA AiSAQ™ (**A**ll-**i**n-**S**torage **A**NNS with Product **Q**uantization) technology addresses the scalability concerns when performing vector searches on large datasets in RAG pipelines. AiSAQ technology offloads the indexes to high-capacity SSDs transforming DRAM-based scalability limitations into an SSD-based scalable architecture with competitive RAG performance when compared to DRAM. Comparable tests of AiSAQ technology also showed a ~**396x** reduction in DRAM utilization[3].

As AI becomes ubiquitous, massive volumes of data are being generated and utilized by vector databases to provide information that use cases can reference. The ability to lower the DRAM footprint with KIOXIA AiSAQ vector search engine technology and KIOXIA high-capacity SSDs can help pave the way for more sustainable AI solutions.

Benefits of High-Capacity SSDs for AI

In addition to more storage space and fast access to queries, high-capacity SSDs offer several key benefits:

High-Capacity SSD Benefits	AI Use Cases / Benefits
More Storage Capacity	Available in capacities ranging from 30.72 TB to 245.76 TB – a 245.76 TB SSD: • Doubles the density in the same SSD footprint (vs. 122.88 TB) • Stores larger LLMs in the ingest stage and scales vector databases when using RAG • Yields over 9.8 petabytes[1] (PB) (40 x 245.76 TB) in a 40-slot E3.L 2U rack-mount server • Yields over 2.9 PB (24 x 122.88 TB) in a 24-slot 2.5-inch 2U rack-mount server • Yields over 3.9 PB (32 x 122.88 TB) in a 32-slot E3.S 1U rack-mount server • Provides 2x capacity per watt (vs. 122.88 TB) Provides room for future data growth
High-Speed Data Access / Transfer	Utilizes PCIe® Gen5 performance and the NVMe™ 2.0 protocol Delivers fast access to data (low latency) Delivers improved performance per watt versus lower capacity SSDs Delivers improved density per watt versus lower capacity SSDs
Reliability and Security	Built with enterprise-grade components to ensure data integrity and longevity Available with security[4] and encryption capabilities such as SIE[5], SED[6] and FIPS 140-3 Level 2 SED validation[7] More cells to write to and erase from equate to longer SSD lifespans Dual-port design provides one drive for server and high availability storage arrays
Increased Efficiencies	Delivers fast application loading times Boots systems quickly Enables seamless multitasking between different tasks Provides high capacity/high performance within a 25-watt power envelope
Form Factor Options	Supports: • 2.5-inch SSDs: up to 122.88 TB for the most common servers today • E3.S SSDs: up to 122.88 TB for all flash (AF)-centric servers • E3.L SSDs: up to 245.76 TB for ultra-capacity density AF servers/storage Takes up less space than HDDs: • Contributes to more efficient capacity usage • Improves cubic feet per minute (CFM) airflow for cooling
Target Applications	Data-intensive Applications: • AI/ML • Data analytics • Video editing • Large-scale data processing • High-performance computing (HPC) • Content delivery networks (CDNs)

Supporting AI with KIOXIA High-Capacity SSDs[8]

AI Workload	Function	Storage Requirements	KIOXIA High-Capacity SSDs
Data Ingestion	Import vast amounts of raw data	High-sequential bandwidth	
Training	Process data through AI algorithms Evaluate results Refine data for accuracy	Read-dominated training Write-dominated checkpointing	
Inference	Model recognizes new patterns Extrapolate conclusion	Read-intensive dominated by high-volume small random reads	
Grounding (RAG)	Model incorporates additional data specific to the use case Utilizes a vector database with a searchable repository	Mix of random reads and writes	

Data Ingestion Stage: KIOXIA LC9 Series NVMe™ SSDs

KIOXIA LC9 Series NVMe SSDs are ideally suited for the data ingestion stage of the AI workflow and deliver very high-capacity density with unparalleled power and cost efficiency. It's at this stage of the AI process where storage needs to be scalable and cost-effective since the raw data may consist of unstructured or semi-structured data that requires reformatting and pre-processing, potentially growing the dataset further.

First generation KIOXIA LC9 Series NVMe SSDs received the FMS 2025 Best of Show award in the SSD Technology category. It features capacities from 30.72 TB up to 245.76 TB, BiCS FLASH™ generation 8, 2 terabit[9] (Tb) QLC die, and represent some of the highest capacity and most dense SSDs[10] in the industry. Available in 2.5-inch, E3.S and E3.L form factors.

KIOXIA LC9 Series NVMe SSD

Specifications

Sequential Read: *up to 12 GB/s**
Sequential Write: *up to 3.5 GB/s*
Random Read: *up to 1,000 KIOPS***
Random Write: *up to 45 KIOPS*

Supported High Capacities

30.72 TB – 61.44 TB – 122.88 TB – 245.76 TB

**GB/s = gigabytes per second*
***IOPS = input/output operations per second*

Training and Inference Stages: KIOXIA CD9P Series NVMe SSDs / KIOXIA CM9 Series NVMe SSDs

KIOXIA CD9P Series NVMe SSDs are good fits for the training stage of the AI workflow and deliver low latency and high bandwidth for read-dominated training and write-dominated checkpointing. These SSDs deliver high sustained and burst throughput so that the compute resources can do training on different pieces of data in parallel and optimize the model using a smaller, task-specific dataset.

These SSDs are also well positioned for the inference stage or execute phase of the AI workload process that applies learned knowledge to new, unseen data for making predictions or decisions. Additionally, KIOXIA CM9 Series E3.S SSDs are also a good match for inference workloads given their high-capacity, mixed-use endurance KV cache and suitable for context memory storage. This stage requires low latency and is also read-dominated requiring high bandwidth.

KIOXIA CD9P Series NVMe SSD

Specifications

Sequential Read: *up to 14.8 GB/s*
Sequential Write: *up to 7 GB/s*
Random Read: *up to 2,600 KIOPS*
Random Write: *up to 750 KIOPS*

Supported High Capacities

30.72 TB – 61.44 TB

KIOXIA CD9P Series NVMe SSDs feature capacities up to 61.44 TB (2.5-inch) and up to 30.72 TB (E3.S), BiCS FLASH generation 8 TLC die and CBA (CMOS directly Bonded to Array) technology and offered in read-intensive (1 DWPD[11]) and mixed-use (3 DWPD) endurances. Available in 2.5-inch and E3.S form factors.

KIOXIA

KIOXIA CM9 Series NVMe™ SSDs feature capacities up to 61.44 TB (2.5-inch) and up to 30.72 TB (E3.S), BiCS FLASH™ generation 8 TLC die and CBA technology and offered in read-intensive (1 DWPD) and mixed-use (3 DWPD) endurances. These SSDs deliver efficient scaling of AI operations while reducing power usage and total cost of ownership across data center environments. Available in 2.5-inch and E3.S form factor.

Grounding (RAG): KIOXIA CM9 Series NVMe SSDs

KIOXIA CM9 Series NVMe SSDs are an excellent solution for RAG implementations where RAG-based vector databases can grow significantly larger than the size of the data. With up to 61.44 TB capacities, these SSDs are well suited where mixed workloads are prevalent and high IOPS, high bandwidth, low latency and short tail latency are required. As AI models grow in complexity and scale, the need for storage solutions that can sustain high throughput and low latency, and allow for better thermal efficiency, will become critical.

KIOXIA CM9 Series NVMe SSD

Specifications

Sequential Read:	*up to 14.8 GB/s*
Sequential Write:	*up to 11 GB/s*
Random Read:	*up to 3,400 KIOPS*
Random Write:	*up to 800 KIOPS*

Supported High Capacities

30.72 TB – 61.44 TB

KIOXIA CM9 Series NVMe SSDs feature capacities up to 61.44 TB (2.5-inch) and up to 30.72 TB (E3.S), BiCS FLASH generation 8 TLC die and CBA technology and offered in read-intensive (1 DWPD) and mixed-use (3 DWPD) endurances. These SSDs deliver efficient scaling of AI operations while reducing power usage and total cost of ownership across data center environments. Available in 2.5-inch and E3.S form factors.

Summary

AI is a burgeoning application that touches many key industries and sectors and requires very large datasets, which in turn has created a need for even higher storage capacities to store all this data. There are stages of AI that perform different functions and have specific data workloads that put demands on the system, such as GPUs, CPUs, host memory and data storage. In all phases of the AI process, more storage space and faster access to queries makes it easier to accomplish AI objectives, which places high dependence on high-capacity SSDs.

RAG is an AI technique that allows an LLM to incorporate additional data specific to the use case. It utilizes a vector database with a searchable repository that finds relevant context. The size of the vector databases and associated indexes are a growing concern in many RAG implementations as extremely large DRAM footprints are required to store them. For those applications where the vector databases and indexes are offloaded to high-capacity SSDs (via KIOXIA AiSAQ™ technology), RAG can eliminate DRAM-based scalability limitations and deliver performance comparable to DRAM.

For all AI workload stages described within, more storage space and faster access to queries is the path towards greater efficiency, placing high reliance on high-capacity SSDs to meet these objectives. KIOXIA has three high-capacity SSD series for AI workloads and RAG implementations that include KIOXIA LC9 Series NVMe SSDs, KIOXIA CD9P Series NVMe SSDs and KIOXIA CM9 Series NVMe SSDs.

- *The KIOXIA LC9 Series is ideal for the AI data ingestion stage delivering the highest capacity density with unparalleled power and cost efficiency. This AI stage values high bandwidth, large capacity and efficiencies associated with space, cost and power.*
- *The KIOXIA CD9P Series is well-positioned for the training and inference stages. These AI stages value low latency and high bandwidth for read-dominated training and write-dominated checkpointing.*
- *The KIOXIA CM9 Series is targeted for grounding implementations where RAG-based vector databases can grow significantly larger than the size of the data. For these use cases, mixed workloads are prevalent. This AI workload values high IOPS, high bandwidth and low latency.*

Visit the KIOXIA enterprise SSD home page and data center SSD home page as more information becomes available for these KIOXIA high-capacity SSD series.

Addendum

Key Sectors Using AI

Sector	AI Use Cases / Benefits
Agriculture	Enables precision farming Monitors crop health Predicts yields Uses drones and AI-based sensors to improve productivity
Autonomous Vehicles	Optimizes automotive logistics Coordinates sensors Enhances safety and transportation efficiency networks
Cybersecurity	Enhances threat detection and responses Analyzes large volumes of data to determine potential threats Automates responses to cyberattacks
Education	Personalizes learning by adapting materials to the learner's pace and style Automates administrative tasks Provides tutoring systems
Energy	Optimizes energy distribution in smart grids Aids in preventative equipment maintenance Predicts weather conditions to enhance the efficiency of renewable energy sources
Entertainment	Provides curated personalized media and entertainment content Assists in movie and music production Analyzes viewer preferences to influence future productions
Finance	Detects fraud Performs risk management Enables algorithmic trading Enhances customer service Identifies financial patterns Predicts market trends
Healthcare	Enhances diagnostic accuracy Personalizes treatment plans Manages patient data Assists in early health detections Assists in robotic surgeries Acts as virtual health assistants
Legal	Automates document analysis Assists in legal research Streamlines case management enabling legal professionals to focus on strategic decision-making
Manufacturing	Drives robotic automation that performs assembly line tasks Enhances quality control Predicts equipment failures through predictive maintenance
Public	Enhances public safety through surveillance systems Streamlines administrative processes Improves public service delivery through predictive analysis
Real Estate	Determines property valuations Predicts market trends Automates property management tasks Enhances virtual tours Enhances lead generation
Retail	Assists customers with personal recommendations Optimizes inventory management Automates customer interactions Analyzes consumer behavior and buying trends
Telecommunications	Optimizes network management Enhances customer service via virtual assistants Improves system reliability through predictive maintenance

FOOTNOTES:

[1] Definition of capacity: KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes, a terabyte (TB) as 1,000,000,000,000 bytes and a petabyte (PB) as 1,000,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1Gbit = 2^{30} bits = 1,073,741,824 bits, 1GB = 2^{30} bytes = 1,073,741,824 bytes, 1TB = 2^{40} bytes = 1,099,511,627,776 bytes and 1PB = 2^{40} bytes = 1,125,899,906,842,624 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

[2] KIOXIA + AI infographic, published March 2025, V2.0.

[3] Based on testing performed by KIOXIA America, Inc., and completed on October 1, 2024.

[4] Optional security feature compliant drives are not available in all countries due to export and local regulations.

[5] SIE: Sanitize Instant Erase is compatible with the Sanitize device feature set, which is the standard prescribed by NVM Express™, Inc., first introduced in the NVMe v1.3 specification, and improved in the NVMe v1.4 specification, and by the T10 (SAS) and T13 (SATA) committees of American National Standards Association (ANSI).

[6] A Self-Encrypting Drive (SED) encrypts/decrypts data written to and retrieved from an SSD via a password-protected alphanumeric key, continuously encrypting and decrypting the data. In support of the SED security option (TCG-Opal/Ruby), there are a limited number of features not supported.

[7] FIPS 140-3 Level 2 SED validation validates that an SSD's cryptographic module is in compliance with the FIPS 140-3 standard developed by the National Institute of Standards and Technology (NIST) through its rigorous Cryptographic Module Validation Program (CMVP) certification process.

[8] All KIOXIA high-capacity SSD models listed are on the PCI-SIG® PCIe® 5.0 Integrators List as PCIe 5.0 compliant, https://pcisig.com/developers/integrators-list?field_version_value%5B%5D=5&combine=KIOXIA. The product images shown are representations of design models and not accurate product depictions.

[9] The flash memory capacity is calculated as 1 terabit (1 Tb) = 1,099,511,627,776 (2^{40} bits), and 1 terabyte (1 TB) = 1,099,511,627,776 (2^{40} bytes).

[10] As of this publication date and based on a KIOXIA Corporation survey.

[11] DWPD: Drive Write(s) Per Day. One full drive write per day means the drive can be written and re-written to full capacity once a day, every day, for the specified lifetime. Actual results may vary due to system configuration, usage, and other factors.

TRADEMARKS:

DISCLAIMERS: