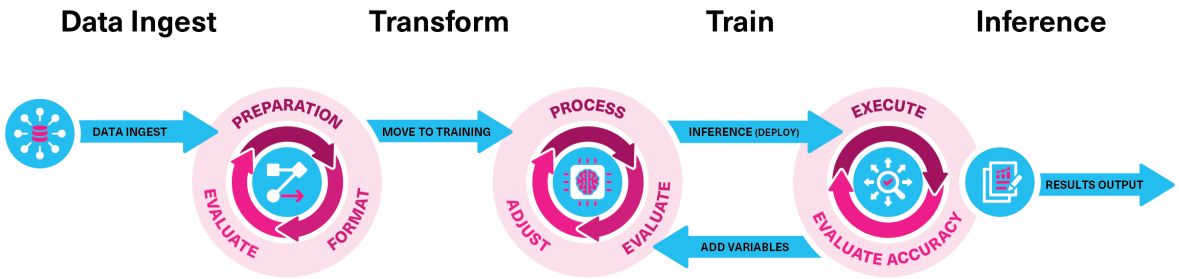
KIOXIA

Top 5 Reasons

Top 5 Reasons to Deploy E1.S SSDs for GPU-based AI Training and Inference Workloads

Artificial intelligence (AI) uses massive datasets to deliver responses which benefit many industries and applications. There are different AI stages that perform specific functions and have data workloads that put demands on system resources such as GPUs, CPUs, memory and storage. For the AI stages depicted below, more storage space and faster access to queries make it easier to accomplish each AI workload objective.



Source: KIOXIA Corporation[1]

AI workloads move enormous amounts of data, whether training large language models (LLMs) or running real-time inference at scale. While GPUs take on the computation, they depend on fast, reliable data storage to fulfill the varied AI workload requirements. SSDs based on the EDSFF E1.S standard provide benefits that make them a viable storage option for GPU-based AI systems. E1.S SSDs are scalable solutions that feature flexible thermal cooling designs, efficient power management options, compliance with PCIe® and NVMe™ specifications and support for the Open Compute Project® (OCP®) Datacenter NVMe SSD Specification to deliver high throughput, low latency and high reliability in a compact footprint with efficient, full power GPU utilization.

E1.S SSDs are well suited for AI training workloads where storage capacity is not the main concern. During the training stage, most of the data is temporary, however, shortening the data processing cycles to keep GPUs fully utilized is the objective. These SSDs are also a good fit for AI inference workloads where fast throughput of query responses is the goal. For these AI stages, the compactness of E1.S SSDs enable denser system configurations that free up more space for GPUs.

The top 5 reasons to deploy E1.S SSDs for GPU-based AI training and inference workloads include:

1. *Ability to keep up with GPU throughput*
2. *Flexible thermal and power design*
3. *Easy physical serviceability in Always-On environments*
4. *High density and scalable performance*
5. *Built for scalable AI infrastructures*

Ability to keep up with GPU throughput

GPUs and SSDs are the primary system resources that perform AI training and inference - GPUs compute - tiered memory and a storage hierarchy temporarily holds key data and moves it efficiently. GPUs can start training an LLM when massive datasets are fetched from high-capacity HDDs to QLC-based SSDs. The data is then prepared and cleaned for training. Key data is then pulled into faster SSDs (warm storage) so that active training datasets and checkpoints can be close to GPUs. The most important data (hot data) is then moved to high-bandwidth memory (HBM) that enables GPUs with extremely fast access to active data for immediate computation. When HBM runs out of LLM training data, less used data will be offloaded back to SSDs. The trained LLM then applies its learned knowledge to new, unseen data in real-world applications, comparing it with the information it learned during training and then generating an output response.

SSDs not only fill the storage capacity gap by enabling high-capacity storage of massive datasets, but they also deliver faster performance and lower latency than HDDs when responding to multiple queries and data transfers at various checkpoints for the process of dataset pretraining and retraining. If storage lags during these AI stages, it becomes a bottleneck where GPUs may become underutilized. To keep the many GPU cores busy executing algorithms, high capacity, high throughput and low latency SSDs are required.

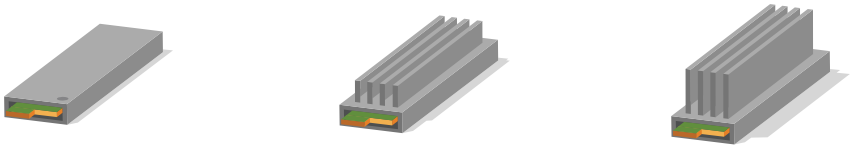
Given the nature of highly parallel and data hungry training and inference workloads, E1.S SSDs are a good solution for data storage. As an example, E1.S SSDs can deliver up to 12.5 gigabytes per second[2] (GB/s) of sequential read throughput and up to 5.8 GB/s[2] of sequential write throughput. This high-speed sequential performance helps to continually feed GPUs with data, reducing idle time while improving overall GPU efficiency.

When sensitive information present in a dataset needs to be protected, E1.S SSDs are available with a Self-Encrypting Drive[3] (SED) security option[4]. There is also support for the Trusted Computing Group® (TCG) Opal v2.0 Specification which is compliant with the OCP® v2.5 Security Project.

TCG Opal v2.0	OCP v2.5
This specification enhances data security on SSDs with a specific focus on SEDs where strong encryption of stored data is achieved through advanced cryptographic techniques.	*This specification was developed by the OCP cooperative community to help ensure that OCP-compliant data centers, cloud devices, appliances, etc., meet the highest security benchmarks.*

Flexible thermal and power design

GPU servers run hot especially in high-density configurations. To mitigate hot environments, improve interoperability across vendors and platforms, and provide the right balance of airflow, cooling and storage density, E1.S SSDs are available in multiple heatsink options which include 9.5 mm, 15 mm and 25 mm:



E1.S Heatsinks Specifications	9.5 mm	15 mm	25 mm
Thickness	9.5 mm	15 mm	25 mm
Width	33.75 mm	33.75 mm	33.75 mm
Length	118.75 mm	118.75 mm	118.75 mm

The innate thermal controls help maintain stable performance under intense workloads, preventing throttling and keeping the system consistent through long training or inference sequences. Liquid cooling is a recent industry standard and a new trend in AI storage technologies that will push E1.S SSD performance limits. An initial liquid-cooling design, exclusive for 9.5 mm heatsink options, uses an external cold plate with circulating liquid coolant that attaches to a modified E1.S 9.5 mm SSD enclosure. New liquid cooling updates in the latest SNIA® SFF-TA-1006 specification[5] include enclosure modifications to improve the compatibility with cold plates.

Easy physical serviceability in Always-On environments

AI systems are often in constant Always-On operations which make any type of downtime potentially expensive. E1.S SSDs can be hot swapped on the fly so that an entire server does not need to be taken down to replace a single SSD. Additionally, scheduled maintenances or upgrades can also be conducted transparently without interrupting active GPU workloads. This level of physical serviceability is significant for improving total cost of ownership.

High density and scalable performance

GPU servers tend to be densely packaged resulting in minimal real estate for local storage making E1.S SSDs ideal from a packaging perspective. E1.S SSDs enable larger capacity per compute rack unit when compared to traditional 2.5-inch or M.2 drives. The compact E1.S form factor allows for more drives per server or GPU node enabling higher collective bandwidth which is critical for AI queries that need fast access to large datasets.

Also, as defined by the EDSFF E1.S specification, with PCB width dimensions increased to 33.75 mm (from 22 mm in traditional M.2 SSD form factors), more space is available on the PCB for additional flash memory chips, which in turn provide more capacity per drive.

Built for scalable AI infrastructures

The E1.S form factor applied to SSDs has garnered strong industry support and standardization under EDSFF specifications which allows for smooth integration across various platforms and vendors. It is designed to scale from single-node GPU systems to multi-rack AI clusters and will continue to evolve with future PCIe® and system design standards.

Final Thoughts

E1.S SSDs are a viable option for GPU-based training and inference AI platforms by providing the following storage requirements:

AI Workload	Function	Storage Requirements
Train	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

E1.S SSDs are helping to bridge the gap between storage and GPU compute by providing a combination of speed, efficiency and serviceability. They are ideal for AI systems that need to process large amounts of data with minimal delay.

Notes:

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

[2] Based on KIOXIA XD8 Series SSD specifications provided by KIOXIA Corporation and accurate as of this publication data but subject to change. Read and write speed may vary depending on various factors such as host devices, software (drivers, OS, etc.) and read/write conditions.

[3] SED (Self-Encrypting Drive) encrypts/decrypts data written to and retrieved from an SSD via a password-protected alphanumeric key, continuously encrypting and decrypting the data. Optional security compliant drives are not available in all countries due to export and local regulations.

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

[5] Source: "SNIA® SFF-TA-1006 specification for Enterprise and Datacenter 1U Short Device Form Factor (E1.S)," Rev. 2.0, December 19, 2025, Section 5.4.