KIOXIA

WhitePaper February 2025

Understanding AI Storage Challenges and the SSD Advantage

Authors:
Jürgen Ahaus, General Manager, SSD Application Engineering, ***KIOXIA*** *Europe GmbH*
Aleksandar Relic, Chief Engineer, SSD Marketing and Engineering, ***KIOXIA*** *Europe GmbH*
Li-Fan Bartsch, Product Manager, SSD Marketing and Engineering, ***KIOXIA*** *Europe GmbH*
Niall O'Sullivan, Senior Manager, SSD Marketing and Engineering, ***KIOXIA*** *Europe GmbH*

Contents

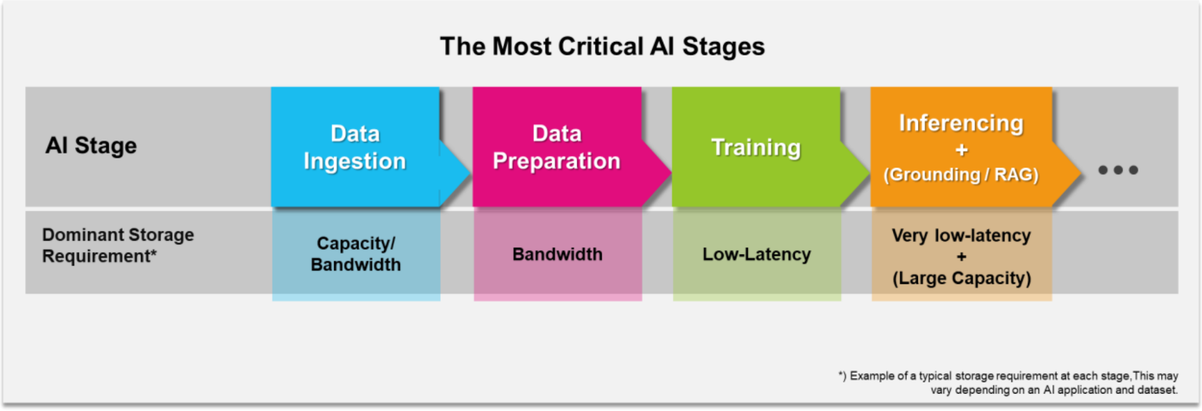
Introduction

In the last few years, we have witnessed a sudden breakthrough in the field of **artificial intelligence** (AI) and **machine learning** (ML), which came as a consequence of ground-breaking AI algorithms and a rapid increase in GPU performance. The progress in AI technology and a massive amount of training data has brought changes in **data storage**, resulting in a surging demand for specialized high-capacity, high-performance storage systems. It is a common perception today that a key objective and challenge for AI data storage systems is to ensure that the training efficiency is kept at a high level and the GPU infrastructure is used to its full capacity. This is true; however, this perspective is somewhat limited as the requirements and goals of AI data storage extend further. To offer a broader perspective, this paper explores additional key aspects and challenges associated with AI data storage and shows how **KIOXIA**'s advanced SSD technology can effectively address them.

Overview of the AI Processing Cycle

Training datasets are growing in size. A CPU’s dataset is typically measured in MegaBytes and GigaBytes, whereas a GPU’s dataset goes from Gigabytes to TeraBytes. If you are to believe Huang’s Law, which was first coined by IEEE Spectrum back in 2018[1], GPU performance will double every two years, which could halve the training time. In the future, with ever-growing AI models, the dataset will be measured in PetaBytes. Keeping the GPUs constantly saturated with data with no latency bottlenecks is a key challenge to keeping the efficiency at a high level.

In modern enterprise server environments, the AI data processing cycle involves multiple stages. Still, we'll focus on the most critical ones from a storage standpoint: data ingestion, data preparation, AI training, and AI inferencing (**Figure 1**). Large Language Models (LLMs) may also include grounding through Retrieval-Augmented Generation (RAG) in this scenario.



The Most Critical AI Stages

AI Stage	Data Ingestion	Data Preparation	Training	Inferencing + (Grounding / RAG)	• • •
Dominant Storage Requirement*	Capacity/ Bandwidth	Bandwidth	Low-Latency	Very low-latency + (Large Capacity)	

*) Example of a typical storage requirement at each stage,This may vary depending on an AI application and dataset.

***Figure 1**: The most critical AI stages from a storage standpoint are data ingestion, data preparation, AI training, and AI inferencing.*

As each stage of the AI workflow presents unique data management challenges - let’s explore what happens at each stage.

Data Ingestion, as one of the first stages, collects and imports files from various sources into a single storage location like a database, data warehouse, or data lake. The data is then prepared for processing and analysis. This stage is capacity intensive, where AI and ML applications generate extensive high-throughput operations and encompass a broad range of storage workloads.

Data Preparation is the stage where data is cleaned, transformed, and engineered to create a structured and possibly enriched dataset. This dataset is later split into training, validation, and test sets to ensure successful model training and evaluation. The I/O workload can be

[1] Move Over, Moore’s Law: Make Way for Huang’s Law, IEEE Spectrum, Tekla S. Perry (Accessed October, 2024)

quite intensive due to the need to frequently access, modify, and rewrite data, especially for large datasets, making storage performance critical in this phase. Actual storage workload and access pattern may vary with the target application and input data format, but having storage that offers high write bandwidth is very beneficial at this stage.

Model Training moves from a data-intensive process to a more latency-sensitive one. In terms of storage requirements, this phase is one of the most demanding. The storage workload heavily depends on the model, dataset, and training process. However, AI training typically involves high data throughput for loading large datasets and frequent checkpointing. While high throughput might be critical for handling large volumes of data, it has to be borne in mind that the low latency storage helps improve the overall efficiency of the training process. For instance, low-latency storage reduces the time it takes to access and load a data chunk into memory - this is especially important when data needs to be accessed frequently during training. High write performance and low latency also reduce the time required for checkpointing leading to less GPU idle time.

Inferencing, either real-time or batch, is characterized by read-heavy operations requiring stable and very low latencies as the model tries to retrieve data from the storage for predictions. Furthermore, in the case of real-time inferencing (e.g., autonomous vehicles, fraud detection, etc.), the system requires the storage to consistently provide high-speed access to data, making low latency essential for keeping the AI model response within a predefined and short time frame.

LLM Grounding (RAG) is a process that uses LLMs with additional information that is not part of their current knowledge, which is usually use-case or application specific. Efficient grounding using the RAG method requires a low-latency response of the storage, as it needs the model to retrieve relevant documents or information in real time during the inferencing stage. Such retrieval tasks, possibly coming from multiple LLM instances, often involve accessing specific chunks of data (like retrieving relevant embeddings). Because of such access requirements, random read performance may often become important here to meet expected turnaround times when generating answers to end users.

Data Storage for Ingestion and Preparation

Raw data required for the training can be collected from various sources, like databases, user-generated content, different IoT and vehicle sensors. In a typical case, this step involves moving the data from its source to the centralized storage or location where it will be processed further. The data type and the actual data source generally determine whether the ingestion will be of batch or (real-time) streaming type.

Storage workloads may vary during the ingestion phase; in some cases, they may require many sequential writes, for instance. This scenario is visible in cases when data is ingested in large batches i.e., when importing structured datasets or other data in batch format. For cases when streaming data or real-time data arrives from multiple sources simultaneously or in small interleaved blocks, the storage workload may finally result in a random-access pattern. The ingestion stage is typically capacity intensive, as a lot of input data might need to be stored in the shortest possible time and then provided for further processing.

During data preparation, the data is processed to be made suitable for AI training. This typically involves cleaning, filtering, normalization, augmentation, labelling, etc. During this phase, the available storage bandwidth may play an important role.

For data ingestion and preparation stages, SSDs might be an optimal storage media choice as they offer **high capacities** and can efficiently handle **bandwidth**-intensive workloads, ensuring quick access to very large datasets. Regardless of the SSD form factor choice, whether it is E1.S or E3.S, or even a legacy 2.5-inch, SSD capacities are increasing considerably, making them ideal candidates for high-capacity and high-performance AI storage systems.

Which SSD Technology to Use

Today's dominant Flash technology used for SSDs is **TLC** (triple-level cell), which is complemented by the emerging **QLC** (quadruple-level-cell) technology. **KIOXIA**'s innovative **BiCS FLASH™** technology includes both types in its Flash memory line-up. However, despite the common misconception that QLC flash will quickly replace the TLC in higher-capacity SSDs, the truth is that the latter will remain present for use cases requiring short latencies, high performance, and a reasonably high level of endurance.

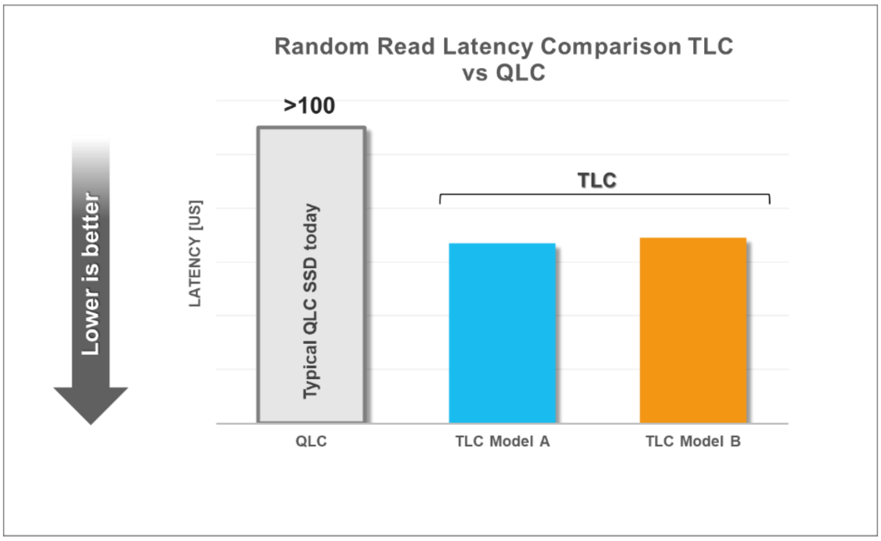


***Figure 2**: QLC SSDs meet the needs where extremely large capacities are required, but TLC SSDs will remain present for use cases requiring short latencies, high performance, and a reasonably high level of endurance.*

At present, the QLC drives may usually be used where extremely large capacities are required (i.e., SSDs with > 60 TB). However, in cases where latency plays an important role, TLC drives still expose much lower latencies than their existing QLC counterparts (**Figure 2**).

Furthermore, it has to be taken into account that existing QLC drives sometimes require countermeasures at the remote AI data storage side to compensate for their lower performance and lower endurance. Currently, some QLC-based storage solutions require a few percent of storage class memory (SCM) in terms of total capacity (**Figure 3**). This requirement can contribute to the overall storage costs if we take into account the costs of SCM, which is multiple times more expensive than QLC or TLC in terms of price-per-gigabyte.

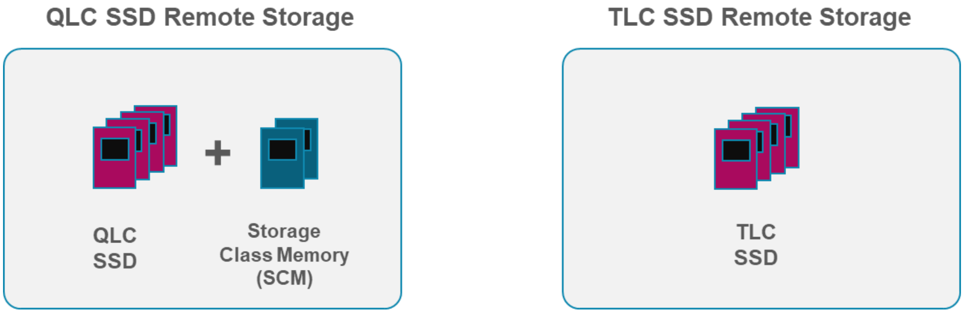


***Figure 3**: A typical QLC SSD setup may require expensive storage class memory (SCM) for caching to enhance endurance and compensate for low IOPS performance.*

Therefore, all these factors must be carefully considered before selecting an appropriate SSD technology for the AI data storage system, as each of those SSD technologies has its unique advantages.

KIOXIA CM7, **KIOXIA CD8P** and **KIOXIA XD8** series are all TLC-based **NVMe**™ SSDs, making them a good and future-safe choice for an entire spectrum of very demanding AI workloads, which tend to generate challenging access patterns and impose latency and/or performance-critical requirements. In order to satisfy higher capacity demands, the **KIOXIA CM7** and **KIOXIA CD8P** series currently goes up to 30 TB in 2.5-inch form factor.

Influence of Power Efficiency

In some cases, the power efficiency of AI data storage systems may have an impact on the overall training capacity. A research document published by Meta and Stanford University shows that in the case of particular Deep Learning Recommendation Models (DLRMs) in Meta's data centers, the **storage and pre-processing** can consume more power than the actual GPUs during the training[2]. The paper further suggests that this situation may limit the training capacity because of the fixed power budgets of data centers.

From that standpoint, the efficiency of data storage and that of SSDs must also be considered when designing storage for AI systems.

To address the mentioned challenges, **KIOXIA** is focusing on optimizing the power efficiency of SSDs, which maintain moderate power consumption even with their high-performance **PCIe® 5.0** capabilities. This helps reduce overall power consumption related to AI storage, freeing up more power for the GPUs to maximize their training performance.

Data Storage for Training

There has been an industry perception in optimizing AI and ML workloads that DRAM and SSDs serve different purposes within computing architectures.

Due to its fast read and write speeds, DRAM is typically used for data that needs to be accessed quickly and frequently, such as the working datasets during model training, where parameters are constantly updated, or for holding the weights of a neural network during inferencing. However, some AI models may already exceed locally available GPU DRAM capacity, creating the need to store the data to the next fastest storage, like NVMe™ SSDs.

[2] Understanding Data Storage and Ingestion for Large-Scale Deep Recommendation Model Training, arXiv platform, Mark Zhao et. al., April 22, 2022 (accessed October, 2024)

Also, the cost of DRAM is quite often a prohibitive factor to adding more in current GPU servers. This puts significant pressure on the storage tier to safely store and quickly deliver the training data to keep GPUs fully saturated, and **maximise training efficiency**.

The training stage is one of the most challenging phases for the storage. The training storage read bandwidth and latency requirements vary greatly depending on the model compute boundness – either compute-bound or I/O-bound – and input size. To illustrate this, the SNIA presentation demonstrates a significant contrast in bandwidth requirements (**Figure 4**). For instance, training a **ResNet-50** image model with eight GPUs demands a read bandwidth of **6.1 GB/s**, while training a **3D U-Net** image model on the same number of GPUs requires a substantially higher read bandwidth of **41.6 GB/s**.

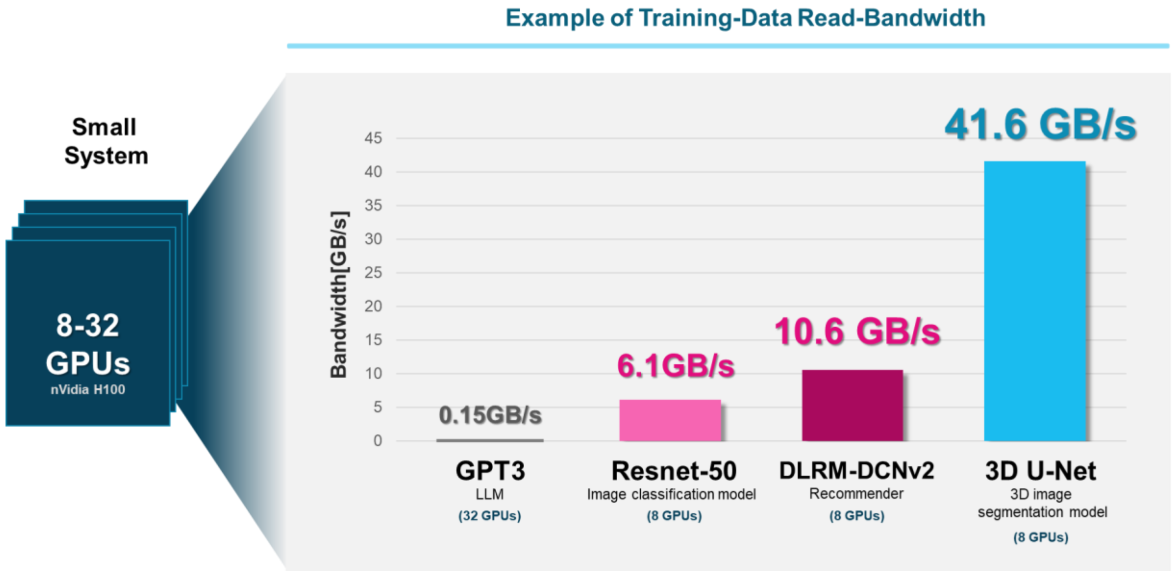


***Figure 4**: Training data storage read bandwidth requirements vary greatly, depending on model compute boundness and input size (Source: SNIA)*

The read workload patterns during training can potentially challenge the data storage system. Nevertheless, a well-engineered AI data storage solution must reliably deliver the necessary read performance with required latency to ensure optimal GPU utilization.

> **KIOXIA CM7** and **KIOXIA CD8P** series, both **PCIe® 5.0** capable, offer very good and stable read performance with low and predictable latencies, which helps improve the training efficiency.

AI Checkpointing

Training modern AI models is resource-intensive, both in terms of time and high infrastructure costs. LLMs have seen widespread adoption, serving a vast user base. However, their underlying complexity is immense. These models typically consist of billions, if not hundreds of billions, of parameters, necessitating substantial computational resources, often involving thousands or even tens of thousands of GPUs to complete training within a reasonable timeframe.

Given the infrastructure complexity combined with an extended training duration – sometimes taking up to several months – **training failures** have become increasingly common. These failures are often attributed to the intricate infrastructure and long training cycles required for large-scale models.

In practice, the success rate for LLM training tasks can sometimes be as low as **56.6 %** for the top 5 % of the most resource-intensive training tasks, as presented in a paper published by Alibaba Group and Nanjing University[3], **Figure 5**. Or in other words, the failure rate can be as high as **43.4 %** for the same group of training tasks.

In the same study, the authors concluded that with a configuration of 128 GPUs, failure rates ranged from one to seven incidents per week. These failures came from various sources, like communication, network issues, software bugs, and hardware malfunctions.

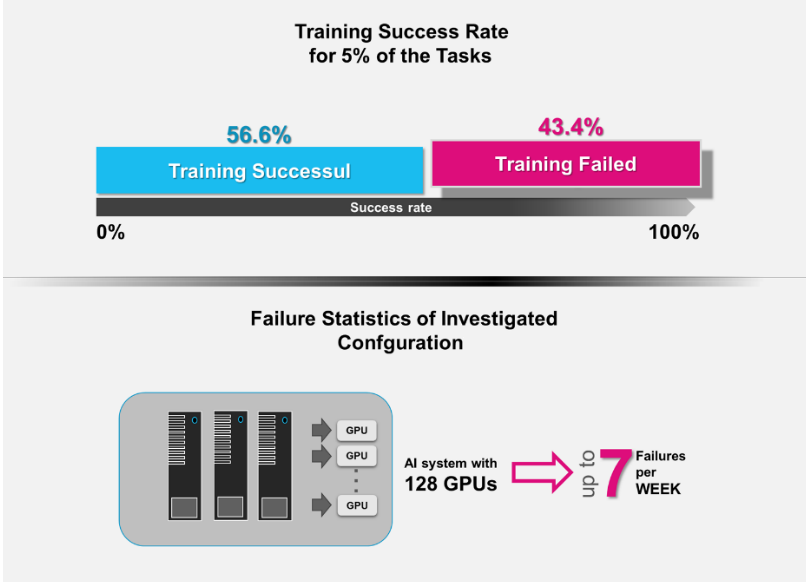


***Figure 5**: Training failure rates for the top 5% of resource-intensive tasks – for a given setup with 128 GPUs, the failure rates are up to seven per week (source: Alibaba Group).*

[3] Unicron: Economizing Self-Healing LLM Training at Scale, Alibaba Group, Nanjing University, Tao He, Xue Li, Zhibin Wang, Kun Qian , Jingbo Xu , Wenyuan Yu , Jingren Zhou, December 30, 2023 (accessed October, 2024)

KIOXIA data center and enterprise drives are designed and produced as high-reliability SSDs with a **MTTF** (Mean Time to Failure) of **2.5 MPOH** (Million Operating Hours), which contributes to overall storage reliability and positively affects failure statistics.

While failures can be minimized in AI systems, they cannot be entirely eliminated due to the increasing complexity of GPU architectures and the intricate topology of supporting networks. However, **a recovery process** must inevitably follow any failure, and this is where **checkpointing** and **data storage** become indispensable.

AI checkpoints are periodic snapshots of the model's internal state, typically created during the training phase. These checkpoints capture the complete model parameters, including weights, optimizer states, and other important metadata, ensuring that all necessary components are preserved for a successful recovery to the most recent stable configuration prior to failure. To understand the important role data storage plays here, let's first look at the **overhead** that AI checkpointing can have on the running training process.
A paper published on the Stanford University website investigates the impact of AI checkpointing and recovery on the training process[4]. **Figure 6** states that: '*Checkpoint-related overheads in full recovery can consume an average of **12 %** of the total training time. And, for the worst 5% of training jobs, training time slowdown can be up to **43 %**.*

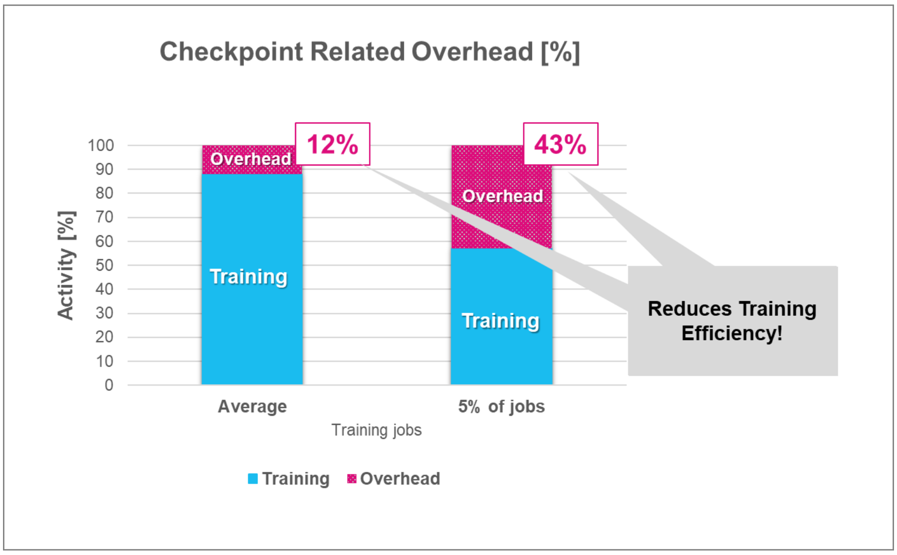


***Figure 6**: Checkpoint-related overheads are responsible for a non-negligible amount of training time (Source: Stanford.edu).*

[4] CPR: Understanding and Improving Failure Tolerant Training for Deep Learning Recommendation with Partial Recovery, stanford.edu, Kiwan Maeng, Shivam Bharuka, Isabel Gao, Mark C. Jeffrey, Vikram Saraph, Bor-Yiing Su, Caroline Trippel, Jiyan Yang, Mike Rabbat, Brandon Lucia, and Carole-Jean Wu, November 20, 2020 (Accessed October, 2024)

This fact, along with the failure statistics presented in the paper published by Alibaba Group and Nanjing University[5], suggests that the impact of AI checkpointing on the overall training efficiency might be significant and cannot be neglected.
To understand the data storage requirements imposed by the AI checkpointing it is important to first examine its dependencies. At the recent Compute, Memory, and Storage Summit organized by the SNIA one presentation proposed that the checkpoint aggregate bandwidth requirements depend on the actual **model size** and **maximum allowed time**[6].

When checkpointing is performed every two hours, the required **write bandwidth** to save a model with 530 billion parameters, which is **7,420 GB** of checkpointing data, within a **72s-time limit**, is **103.1 GB/s**. This results in **1 %** of training time being lost, as the training is generally paused during each AI checkpointing event, **Figure 7**.

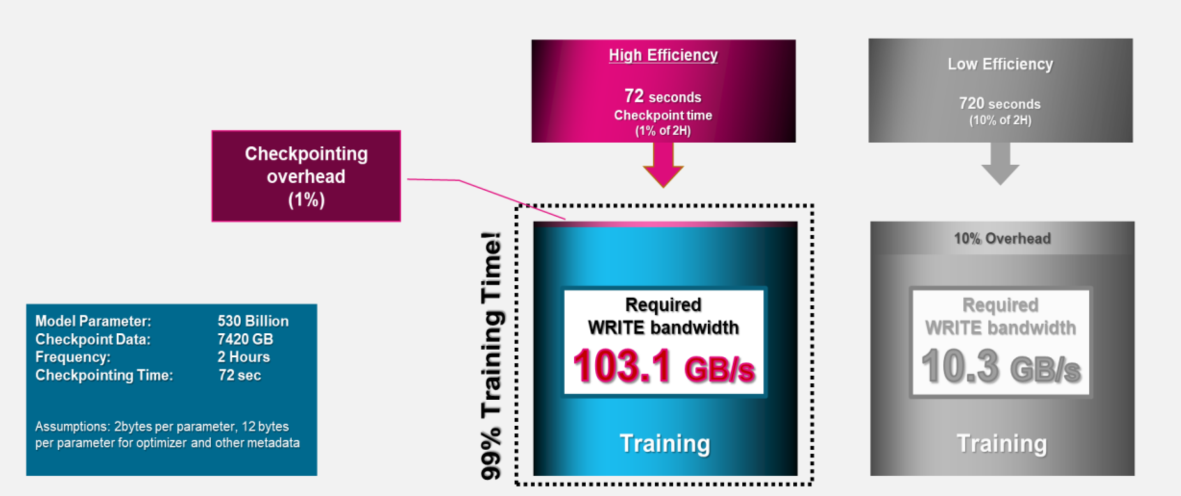


***Figure 7**: Checkpoint-related bandwidth requirement for given model and target overhead of 1% (Source: SNIA)*

Alternatively, if the write bandwidth were only **10.3 GB/s**, the overall portion of checkpointing would be as high as **10 %.** This time would then be directly taken away from the time available for training, causing it to pause for the same portion of the total available time. Therefore, the importance of write bandwidth provided by AI data storage, which is required for efficient AI checkpointing should not be underestimated.

[5] Unicron: Economizing Self-Healing LLM Training at Scale, Alibaba Group, Nanjing University, Tao He, Xue Li, Zhibin Wang, Kun Qian , Jingbo Xu , Wenyuan Yu , Jingren Zhou, December 30, 2023 (accessed October, 2024)

[6] Storage Requirements for AI: Training and Checkpointing, SNIA Data, Networking & Storage Forum (DNSF), John Cardente, Technical Staff, Dell Storage CTO Group, May 21, 2024 (Accessed October, 2024)

To illustrate with a simple example, the write bandwidth required for efficient checkpointing (**103.1 GB/s**) is equal to the aggregated bandwidth of at least 15x **KIOXIA CM7** or 19x **KIOXIA CD8P** SSDs (both PCIe® 5.0 @ 15.3 TB)[7], **Figure 8**.

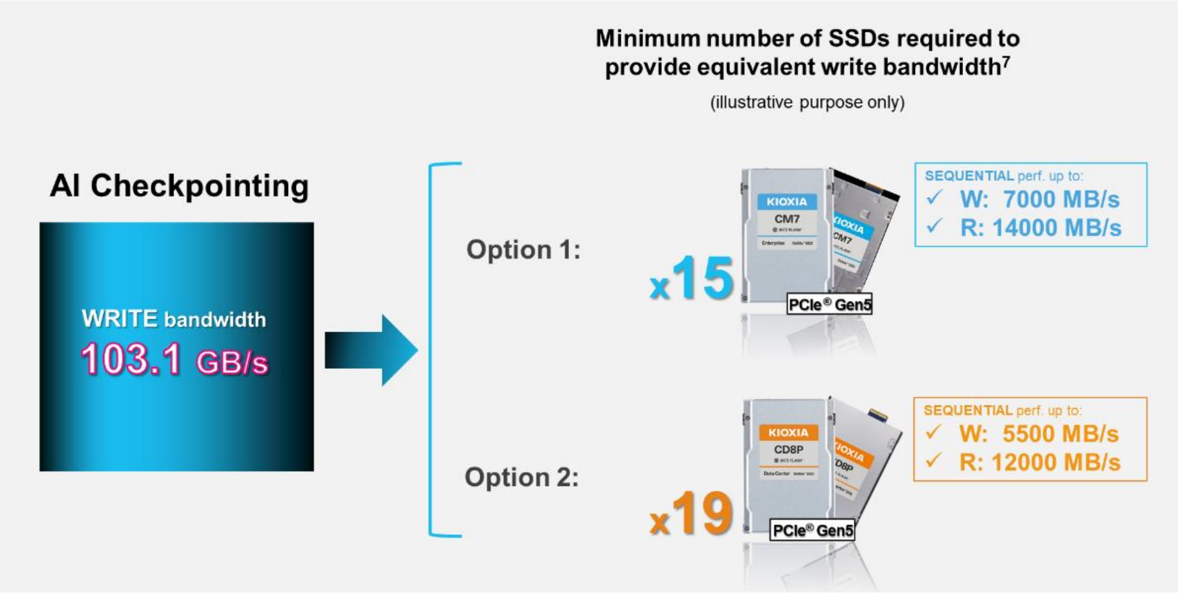


Figure 8*: For illustrative purposes only: For a model with 530 billion parameters the write bandwidth is 103.1 GB/s, which would require a minimum of 15 KIOXIA CM7 series SSDs or 19 KIOXIA CD8P series SSDs.*

It is, therefore, essential that in addition to the data storage requirements imposed by AI training itself, all requirements imposed by checkpointing and recovery are taken into account when planning and designing AI data storage system.

> **The KIOXIA CM7** and **KIOXIA CD8P** SSD series provide very high write bandwidths, which can help reduce the time required for AI checkpointing and by that increase the actual portion of time available for the AI training itself.

Training Data Protection

It is well understood that training data is becoming one of the most valuable assets. Companies that have enough training data want to safeguard it to train the network and provide AI-based products. To defend against data leaks, hardware encryption is more efficient and effective than software encryption.

[7] For illustrative purposes only. Shows that the aggregated sequential write performance of multiple drives meets the required bandwidth. It does NOT indicate, however, that this number of drives is sufficient by any means, as other requirements must be considered.

SED drives protect the data at rest by encrypting it inside the SSD. In case of attempts to steal the SSD, data training data cannot be decrypted onto another system – the drive cannot be unlocked without a password to decrypt the data.

KIOXIA self-encrypting drive (SED) SSDs comply with the **TCG (Trusted Computing Group) standard**, as outlined in **Figure 9**. These perform encryption on the fly using dedicated hardware acceleration and do not introduce any visibly increased latency or performance reduction.

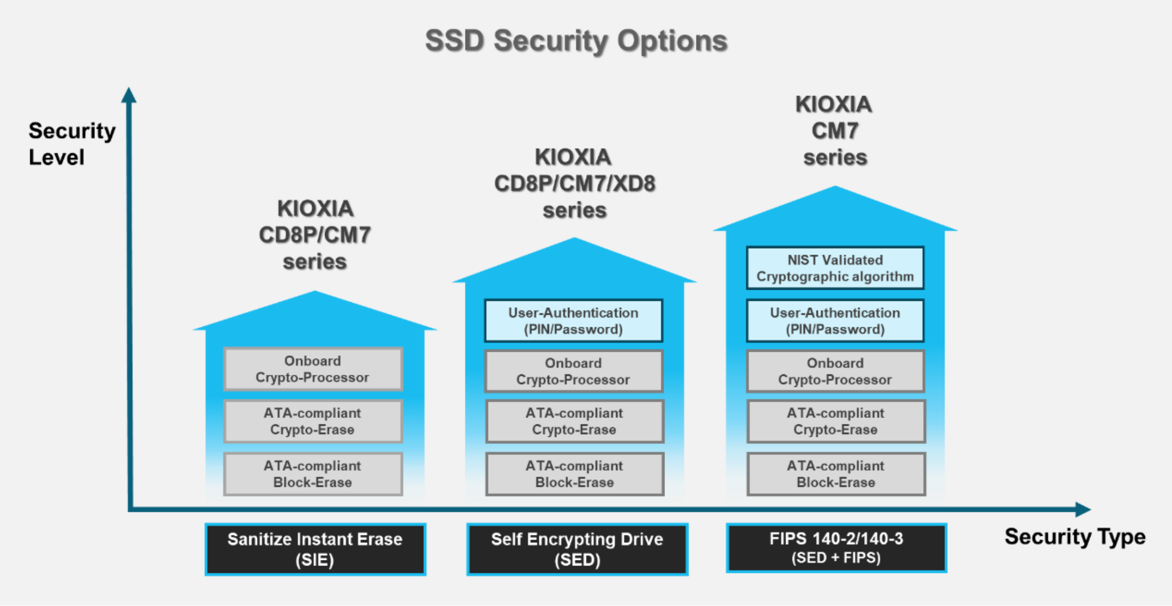

SSD Security Options
Security Level
KIOXIA CD8P/CM7 series
KIOXIA CD8P/CM7/XD8 series
KIOXIA CM7 series
NIST Validated Cryptographic algorithm
User-Authentication (PIN/Password)
User-Authentication (PIN/Password)
Onboard Crypto-Processor
Onboard Crypto-Processor
Onboard Crypto-Processor
ATA-compliant Crypto-Erase
ATA-compliant Crypto-Erase
ATA-compliant Crypto-Erase
ATA-compliant Block-Erase
ATA-compliant Block-Erase
ATA-compliant Block-Erase
Sanitize Instant Erase (SIE)
Self Encrypting Drive (SED)
FIPS 140-2/140-3 (SED + FIPS)
Security Type

***Figure 9**: **KIOXIA** SSDs are available with various security and encryption options: sanitize instant erase (SIE), self-encrypting drive (SED), and SED with Federal Information Processing Standard (FIPS) 140-2/140-3.*

Data Storage for Inferencing

AI inferencing is a process that generally takes place after the training is successfully completed when an AI model is used to make predictions on the new (unseen) input data. The common perception today is that trained models are running in GPU memory and that the storage system plays no role during this stage.

In practice, however, AI storage plays an important role in managing of inference loads and, in some cases, the LLM RAG (Retrieve Augment Generate) approach is used.

In a typical data center, inferencing loads change constantly. So, the ability to quickly adapt to these sudden changes is essential in order to avoid bottlenecks in AI applications or AI services in general. In case of increasing inferencing load, the system should be able to promptly increase the number of inferencing instances, which means it should be able to quickly load AI models - i.e., tens of gigabytes in size - from the storage to GPU Memory. At this point, both latency and the bandwidth of the AI data storage must be able to keep up with the application requirements in order to minimize the fluctuations in the service performance and avoid underutilisation of GPU instances.

In the case of Large Language Model Inferencing, the RAG (Retrieve Augment Generate) approach enables LLMs to retrieve the information from an external knowledge base and use it to generate more accurate responses with up-to-date information.

During RAG, the process of vectorizing proprietary data creates embeddings for RAG solutions that multiply the data's size. Memory-based algorithms like Hierarchical Navigable Small World (HNSW) limit the amount of data these systems can properly utilize. This is mainly due to the fact that the increase in vector database size results in a significant DRAM footprint increase, which makes scaling out quite difficult and very expensive.

The HNSW algorithm performs approximate nearest neighbour (ANN) searches very well in terms of speed and accuracy, making it a very robust vector search algorithm. However, it requires a lot of DRAM in the case of larger datasets, which makes scaling out challenging. So, building on the foundation of Microsoft DiskANN and vector database work, it is now possible to create even larger vector databases using less memory by partially moving from DRAM to SSDs instead.

For RAG applications, SSD-based DiskANN offers performance comparable to in-memory HNSW, making it a cost-effective and scalable alternative to the expensive DRAM-based solution.

KIOXIA AiSAQ™ (all-in-storage ANNS with product quantization) is a research project to transition from **mostly in-storage**, that is, DiskANN, to **all-in-storage** vector databases[8]. It is built upon DiskANN search algorithms through further optimizations for SSDs, and preliminary results suggest its low DRAM footprint without significant degradations in the performance (***Figure 10***). This approach, combined with high performance, and low-latency SSDs, is paving the way to cost-efficient scaling of RAG-based LLM solutions in future applications.

[8] AiSAQ: All-in-Storage ANNS with Product Quantization for DRAM-free Information Retrieval, Kento Tatsuno, Daisuke Miyashita, Taiga Ikeda, Kiyoshi Ishiyama, Kazunari Sumiyoshi, and Jun Deguchi, **KIOXIA** Corporation Japan, April 9, 2024 (Accessed October, 2024)

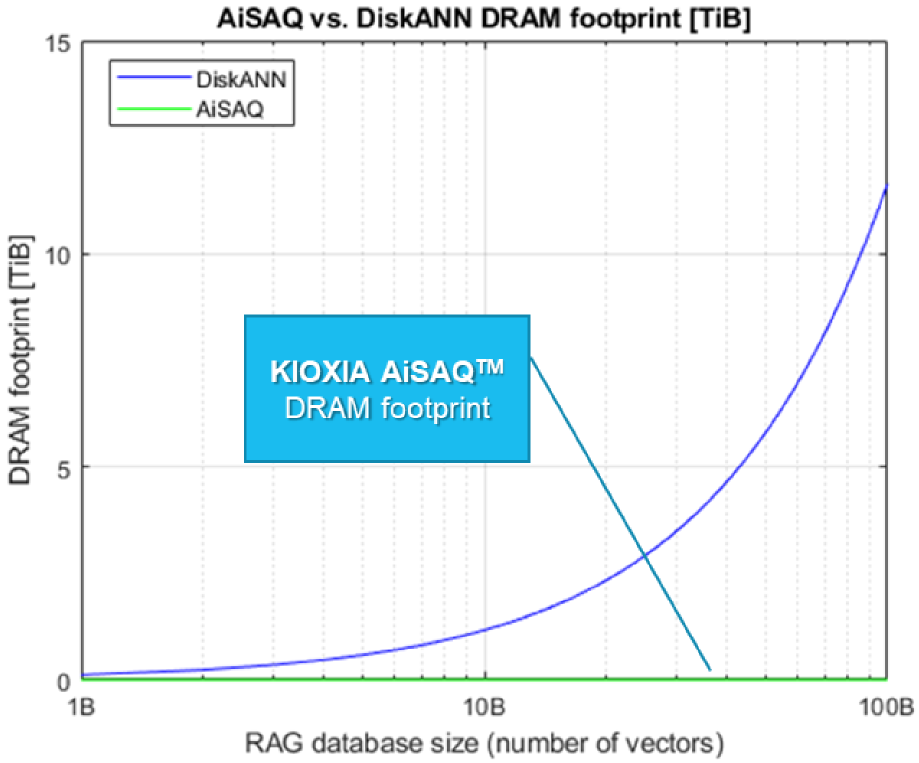


***Figure 10**: KIOXIA AiSAQ^{TM} will enable RAG application at an almost unlimited scale due to its minimum DRAM footprint (Source: KIOXIA).*

KIOXIA SSDs present a very good base for AI systems by providing the necessary bandwidth and latencies required for efficient management of inferencing loads. In the process of LLMs grounding, an external vector database can be scaled out at a relatively low cost, and the data can be retrieved quickly and efficiently using the **KIOXIA** drives and SSD-optimized ANN search methods, like **KIOXIA AiSAQ^{TM}** or DiskANN.

SSDs in AI Systems: Additional Points to Consider

Today's high-capacity SSDs offer significant advantages over previous generation drives in total cost of ownership (TCO) due to the consolidation of power, cooling, and other operational savings, making them particularly suitable for AI-centric storage solutions.

It's also important to recognize that storage accounts for only a small portion of the total costs of an AI system, yet it can significantly influence overall efficiency. Unlike conventional storage systems, decisions regarding SSD selection should be primarily driven by technical requirements and not solely by cost considerations. Let's explore some of those requirements.

SSD Performance and Latency

Storage systems connected to GPU servers must provide high, consistent performance with low and predictable latencies under different workloads to prevent GPU underutilization caused by waiting for data.

As a crucial component of AI storage infrastructure, SSDs must also possess these capabilities to support efficient data flow and enable required system performance. However, SSDs from some manufacturers struggle with situations where there are frequent changes in workloads, jumping from sequential to random or a mixture of both. For some SSDs, this can be observed as a temporary performance drop, or, at certain points, as very long and unexpected latency.

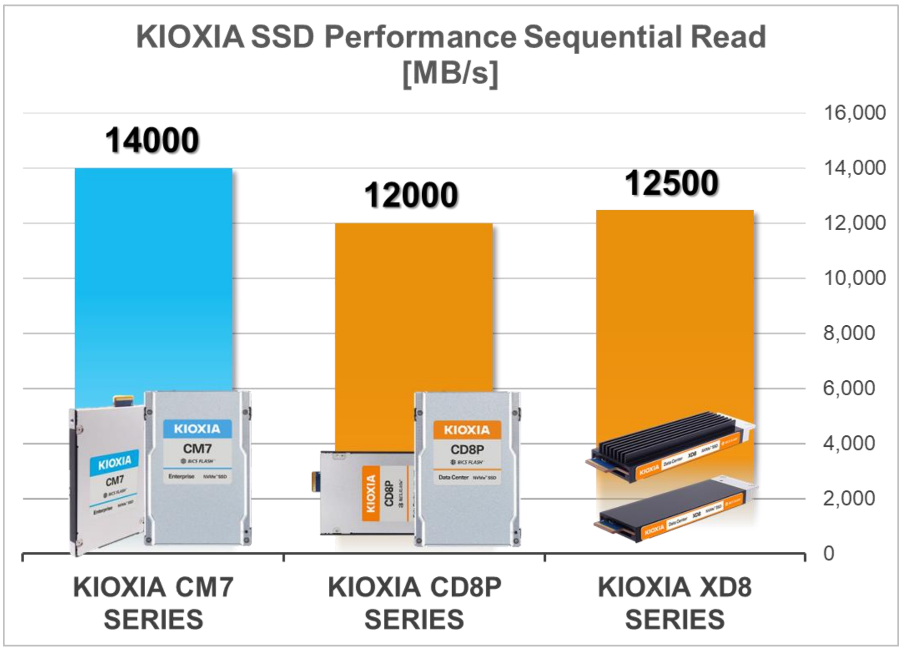


***Figure 11**: The high sequential read performance of KIOXIA SSDs allows for faster iterations and higher training efficiency.*

> **KIOXIA** SSDs offer high performance and low latencies to help data center and enterprise businesses efficiently operate their GPU-based infrastructure (**Figure 11**). A key differentiator is their adaptability to efficiently adapt to changing workloads using **KIOXIA**'s evolved advanced internal algorithms.

Besides performance, the **KIOXIA CM7**, **KIOXIA CD8P** and **KIOXIA XD8** series SSDs provide consistently low read/write latencies and very stable I/O operations per second (IOPS), which is very important for handling the access patterns occurring in many AI workloads.

SSD Form Factors

In AI applications where the drives are positioned very close to the GPU (i.e. local NVMe^{TM} storage), server manufacturers can achieve better cooling with the Enterprise and Datacenter Standard Form Factor (EDSFF). Its relatively small height (e.g., E1.S) or thickness (e.g., E3.S 1T) enables it to achieve better airflow to the back of the server where the GPUs are located. If you’re using the thin version of form factor E3.S 1T, which looks like a conventional 2.5-inch drive with half the height (7.5 mm), you could potentially fit in many more drives within a server enclosure.

It is worth mentioning that some server platform makers still struggle to achieve good signal integrity with PCIe® 5.0 with the legacy 2.5-inch form factors. With EDSFF, the contacts on the PCB provide a much higher signal quality. As the industry moves to PCIe® 6.0, which will potentially double the frequency of the signals, the EDSFF signal integrity makes it future-proof – the industry currently sees no bright future for 2.5-inch in the long run (**Figure 12**).

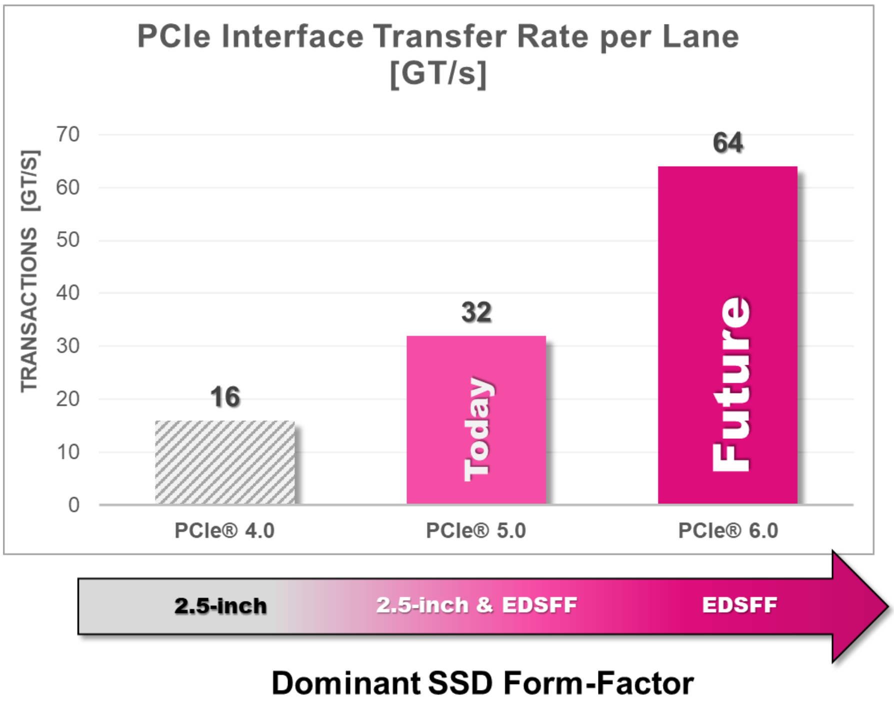
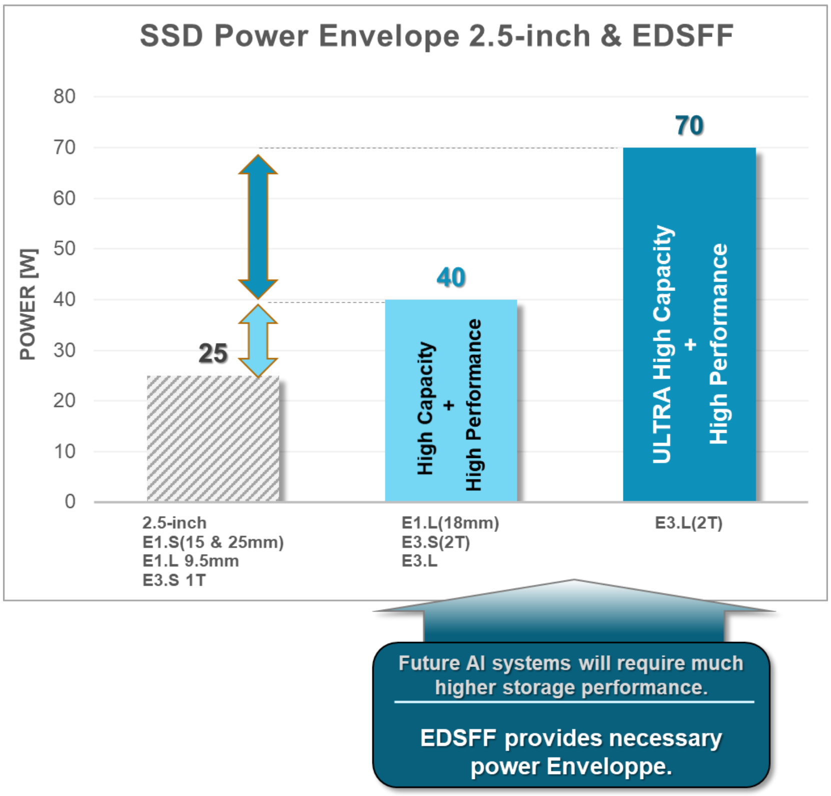


***Figure 12**: As the PCIe® interface rate per lane increases, the dominant SSD form factor migrates from 2.5-inch to EDSFF.*

KIOXIA, in its current portfolio, includes 2.5-inch and EDSFF form factors - more precisely, E1.S (9.5/15 mm) and E3.S 1T.

New Power Envelopes

Despite constant improvements in Flash technology and SSD efficiency, it is inevitable that, as the performance increases, the power consumption also increases. The 2.5-inch SSD allows up to 25 W per slot per drive – for the E1/E3 models, it can be up to 40 W or even 70 W. Though it doesn't need to be used, such a power envelope leaves plenty of room for future performance and capacity increases, which will be necessary to efficiently support current and especially future AI ecosystems in the years to come (**Figure 13**). Even though higher power maybe required for devices with increased capacity, the actual TB/W will be reduced due to gains made in NAND flash density and power efficiency.



Future AI systems will require much higher storage performance.

EDSFF provides necessary power Enveloppe.

Figure 13*: Comparison between the 2.5-inch and EDSFF SSD power envelopes – only EDSFF SDDs will provide the necessary power envelope for future AI systems requiring higher storage performance.*

NVMe^{TM} Drive Integration

Modern GPU-based systems integrate locally attached NVMe^{TM} drives alongside DRAM. These NVMe^{TM} drives work in tandem with DRAM, serving as a high-speed layer for caching or data staging. This collaboration is crucial, especially when large datasets exceed the capacity of DRAM. In such cases, locally attached NVMe^{TM} storage acts as an extended cache, keeping the next chunk or batch of data readily available for the GPU, thereby optimizing the overall data flow and training efficiency.

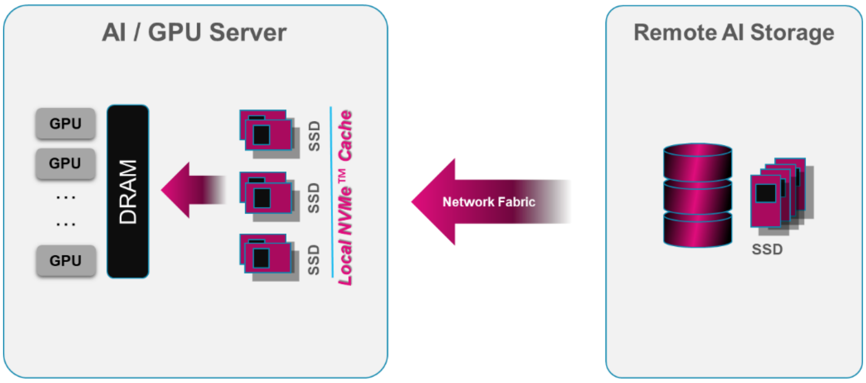


***Figure 14**: Modern GPU-based systems integrate local NVMe™ drives with DRAM, reducing the need to traverse the network fabric.*

The advantage of incorporating local NVMe™ storage alongside remote storage lies in its low-latency and performance, as it reduces the need to frequently traverse a network fabric when delivering data to the GPU.

To seamlessly integrate local and remote SSD storage, NVIDIA has developed **GPUDirect Storage (GDS)**, which allows for highly efficient data transfers from local NVMe™ drives or remote NVMe™ (i.e., NVMe-oF™) storage directly into GPU memory, resulting in higher I/O performance and reduced latency. This approach eliminates the so-called 'bounce buffer' from a data path between the GPU and SSD (**Figure 15**). According to NVIDIA, it can achieve 2-8x higher bandwidth and up to 3.8x lower end-to-end latency[9] has been demonstrated as well.

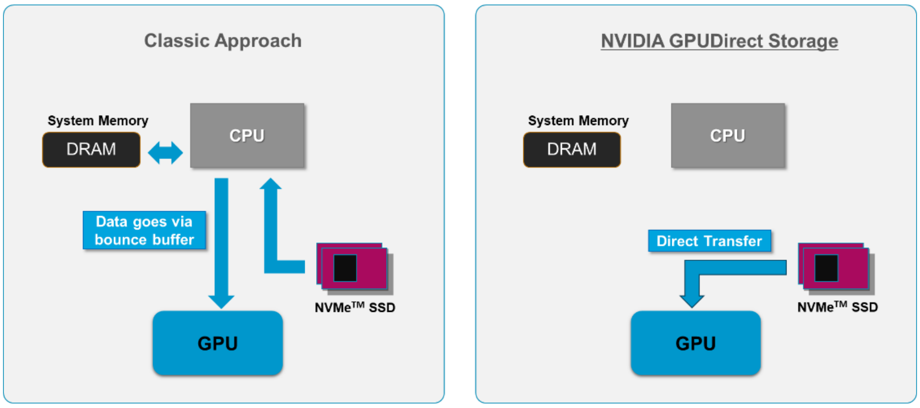


***Figure 15**: NVIDIA's GPUDirect Storage eliminates the 'bounce buffer' from the data path between the GPU and SSD.*

[9] GPUDirect Storage®: A Direct Path Between Storage and GPU Memory, NVIDIA® Developer blog, Adam Thompson and CJ Newburn, August 06, 2019 9 (accessed October, 2024)

Conclusion

In the last few years, the AI industry has made a big leap forward with rapidly growing GPU capabilities and a series of new ground-breaking AI models, which have grown enormously - both in size and complexity. In order to keep GPUs fully saturated, the surrounding infrastructure, including AI data storage, must be carefully planned. To achieve high efficiency and avoid GPU underutilisation, the storage system must be able to reliably deliver excellent read & write performance under changing workloads simultaneously ensuring sufficiently low latencies.

Advanced SSDs (Solid State Drives) are pivotal in modern AI data centers. **KIOXIA** is one of the first to release a PCIe® 5.0 drive for the data center and enterprise markets – the dual-port high-performance **KIOXIA CM7**, single-port **KIOXIA CD8P**, and single-port **KIOXIA XD8** series – that enable next-generation applications like AI, ML and deep learning. Both the **KIOXIA CM7** and **KIOXIA CD8P** drives are suitable for remote AI storage where a high capacity and PCIe® 5.0 level of performance is required or for local NVMe™ caching in the GPU server. **KIOXIA XD8** SSD, on the other hand, might be more suitable for local caching in the GPU server, as its maximum capacity goes up to 7.68 TB and some GPU-based chassis leverage E1.S form factor instead of 2.5-inch (**Figure 16**).

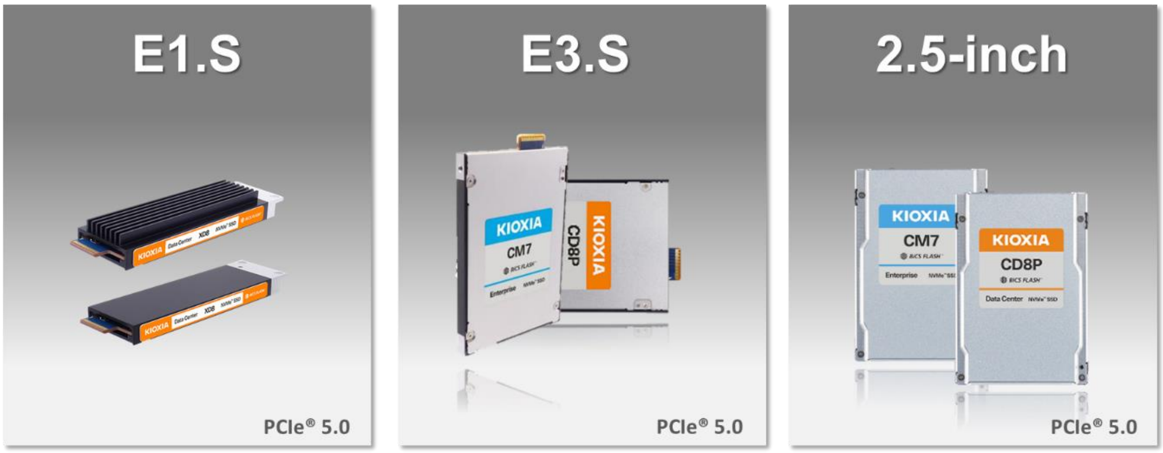


Figure 16*: The PCIe® 5.0 KIOXIA CM7, KIOXIA CD8P series SSDs are offered in 2.5-inch and E3.S form factors, whereas KIOXIA XD8 series SSDs are in the E1.S form factor.*

KIOXIA SSDs are engineered to meet the demands of AI systems across critical stages, like data ingestion, preparation, training, and inference, delivering excellent performance with stable, low-latency operation. Combined with proven reliability, **KIOXIA CM7**, **KIOXIA CD8P** & **KIOXIA XD8** series SSDs provide an excellent base for AI data storage infrastructure, enabling data centers to focus on AI Applications with confidence in the SSDs as a robust and reliable core component.

Sources:

1. Source: “Move Over, Moore’s Law: Make Way for Huang’s Law”, IEEE Spectrum, Tekla S. Perry (Accessed October, 2024), https://spectrum.ieee.org/move-over-moores-law-make-way-for-huangs-law

2. Source: “Understanding Data Storage and Ingestion for Large-Scale Deep Recommendation Model Training”, arXiv platform, Mark Zhao et. al., April 22, 2022 (accessed October, 2024), https://arxiv.org/pdf/2108.09373

3. Source: “Unicron: Economizing Self-Healing LLM Training at Scale”, Alibaba Group, Nanjing University, Tao He, Xue Li, Zhibin Wang, Kun Qian , Jingbo Xu , Wenyuan Yu , Jingren Zhou, December 30, 2023 (accessed October, 2024), https://arxiv.org/pdf/2401.00134

4. Source: “CPR: Understanding and Improving Failure Tolerant Training for Deep Learning Recommendation with Partial Recovery”, stanford.edu, Kiwan Maeng, Shivam Bharuka, Isabel Gao, Mark C. Jeffrey, Vikram Saraph, Bor-Yiing Su, Caroline Trippel, Jiyan Yang, Mike Rabbat, Brandon Lucia, and Carole-Jean Wu, November 20, 2020 (Accessed October, 2024), https://cs.stanford.edu/people/trippel/pubs/cpr-mlsys-21.pdf

5. Source: “Storage Requirements for AI: Training and Checkpointing”, SNIA Data, Networking & Storage Forum (DNSF), John Cardente, Technical Staff, Dell Storage CTO Group, May 21, 2024 (Accessed October, 2024), https://www.snia.org/sites/default/files/SSSI/CMSS24/CMSS24-Cardente-Storage-Requirements-for-AI.pdf

6. Source: “AiSAQ: All-in-Storage ANNS with Product Quantization for DRAM-free Information Retrieval”, Kento Tatsuno, Daisuke Miyashita, Taiga Ikeda, Kiyoshi Ishiyama, Kazunari Sumiyoshi, and Jun Deguchi, **KIOXIA** Corporation Japan, April 9, 2024 (Accessed October, 2024) , https://arxiv.org/pdf/2404.06004

7. Source: GPUDirect Storage®: A Direct Path Between Storage and GPU Memory, NVIDIA® Developer blog, Adam Thompson and CJ Newburn, August 06, 2019 9 (accessed October, 2024), https://developer.nvidia.com/blog/gpudirect-storage/

Trademarks:

Disclaimer:

Definition of capacity: KIOXIA defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 2^{30} = 1,073,741,824 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, such as Microsoft Operating System and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

All company names, product names and service names may be trademarks of their respective companies.

Product illustrations may present design models or previous versions. Actual product may vary.
Images are for illustration purposes only.

Information, including product pricing and specifications, content of services, and contact information is current and believed to be accurate on the date of the announcement, but is subject to change without prior notice. Technical and application information contained here is subject to the most recent applicable KIOXIA product specifications.

About KIOXIA

KIOXIA is a world leader in memory solutions, dedicated to the development, production and sale of flash memory and solid-state drives (SSDs). In April 2017, its predecessor Toshiba Memory was spun off from Toshiba Corporation, the company that invented NAND flash memory in 1987. KIOXIA is committed to uplifting the world with “memory” by offering products, services and systems that create choice for customers and memory-based value for society. KIOXIA's innovative 3D flash memory technology, BiCS FLASH™, is shaping the future of storage in high-density applications, including advanced smartphones, PCs, automotive systems, data centers and generative AI systems. For more information, please visit **KIOXIA** Website**.**