KIOXIA

Performance Brief

Accelerating Machine Learning Performance / Decreasing Training Times with KIOXIA CM7 Series SSDs

Introduction

Machine learning (ML) is a subset of artificial intelligence (AI) and has many uses and applications. ML-based systems use a training process to learn from historical data, identify patterns and make logical decisions with little to no human intervention. Compute and storage resources must be able to work fast and efficiently in order to process and train from the vast amounts of data required to create ML models.

The ML process includes **preprocessing** on raw data so that it is cleaned and ready to be fed into the ML algorithms. The model then goes through a **training** cycle, consuming data in batches so it can learn from the entire dataset. Models may go through many iterations on the same dataset so that the algorithms can create an accurate model. The next step is ML **inference**, and the process of running data points into the ML model to calculate an output. It becomes imperative that the storage system responds to multiple queries very quickly and with low latency. If the storage system lags, compute resources may become underutilized and slow down the ML processes.

To keep the many compute cores utilized at all times, it is necessary that the underlying storage system delivers high throughput so that the compute resources can do training on different pieces of data in parallel. It is common for ML workloads to have bursts of throughput, so sustained and burst throughput must both be high.

This performance brief presents an ML comparison of a KIOXIA CM7-R Series PCIe® 5.0 enterprise NVMe™ SSD with a PCIe 4.0 enterprise NVMe SSD from a leading supplier (Vendor A). Both SSDs were deployed in a test server. The comparison featured the DLIO benchmarking tool[1] which created a dataset based on ‘unet3d’ (aka 3D U-Net) - a convolutional neural network developed for medical imaging.

Each SSD was used to hold the created unet3d dataset consisting of 24,000 files and about 3.47 terabytes[2] (TB) of capacity. PyTorch® (a common ML framework) was also used to create a convolutional neural network model that trained from the dataset. The test comparison emulated an NVIDIA® V100 GPU to test, train and check the ML model from the dataset stored on each drive. The emulation enabled this testing without physically having the GPU hardware.

Six tests were conducted and measured the total time to complete training, average accelerator utilization, training sample processing throughput, average drive training input/output (I/O) throughput, maximum drive training I/O throughput and average drive read latency.

The test results show that the KIOXIA CM7-R Series SSD outperformed the Vendor A SSD in all six tests. It also showed that a PCIe 5.0 SSD can help lessen the amount of time taken to train ML/AI models when compared with a PCIe 4.0 SSD.

The results presented include a brief description of each workload test, a graphical depiction of the test results and an analysis. Appendix A covers the various steps of the ML process. Appendix B covers the hardware/software test configuration. Appendix C covers the configuration set-up/test procedures.

Test Results Snapshot

KIOXIA CM7-R Series PCIe® 5.0 SSD deployed in a test server delivered the following ML results when compared with a leading PCIe 4.0 SSD:

SSD Performance Tests

Maximum Drive Training I/O Throughput
(higher is better)
up to 91% Higher

Average Drive Read Latency
(lower is better)
up to 57% Lower

Average Accelerator Utilization
(higher is better)
up to 14% Higher

Training Sample Processing Throughput
(higher is better)
up to 13% Higher

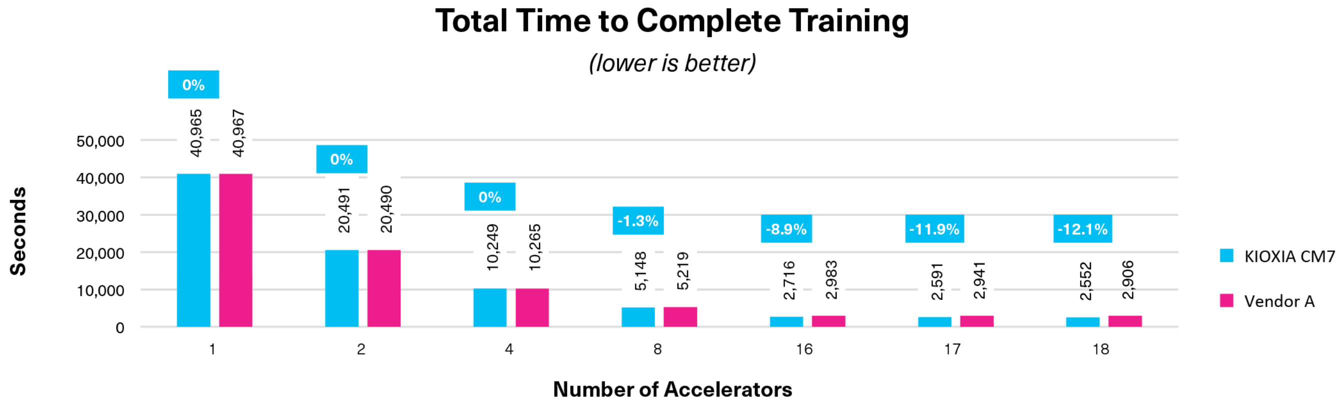
Average Drive Training I/O Throughput
(higher is better)
up to 13% Higher

Total Time to Complete Training
(lower is better)
up to 12% Lower

Test Results[3]

Test 1: Total Time to Complete Training

This test measured the total time it took to perform ML training against the unet3d dataset. For ML applications, it is imperative that the underlying storage system responds very quickly to queries, so that compute resources (i.e., CPUs/GPUs) remain active. The results are in seconds. The lower result is better.



Analysis:

The graph shows that as the number of accelerators scale upward, the KIOXIA CM7-R Series PCIe® 5.0 SSD completes model training faster when compared with the Vendor A PCIe 4.0 SSD. With 18 accelerators, the KIOXIA CM7-R Series SSD was able to train the model more than 12% faster, which, when extrapolated over a year, could translate to about 44 fewer days of model training. The chart below depicts the model training days that the KIOXIA CM7 Series SSD can potentially save as accelerators scale upward:

Model Training Days Saved by KIOXIA CM7 Series SSDs:		
18 Accelerators:	12.1% faster	~44 days saved
17 Accelerators	11.9% faster	~43 days saved
16 Accelerators	8.9% faster	~32 days saved
8 Accelerators	1.3% faster	~4.75 days saved

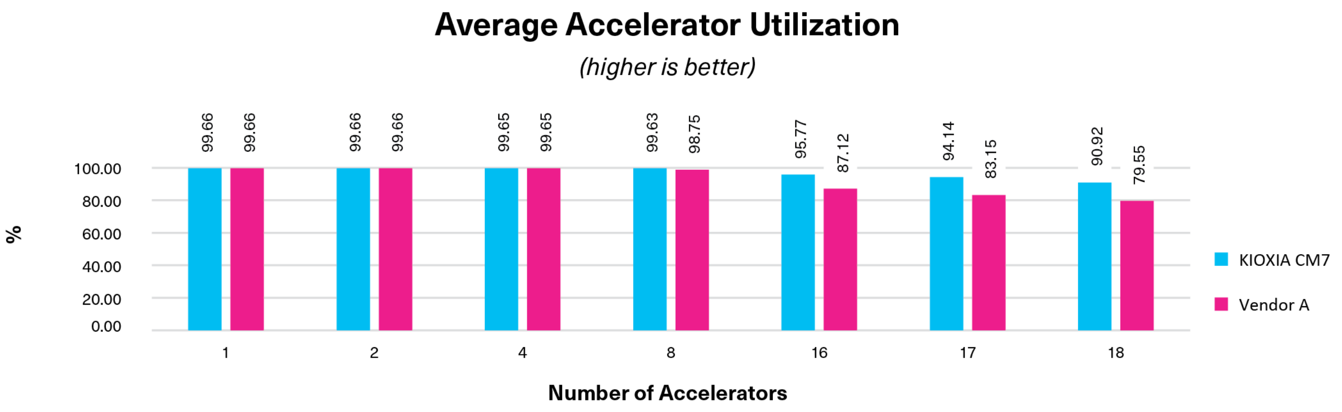
It should be noted that the graph results above are based on 5 epochs of training as set by the DLIO benchmarking tool. As a result, it took the KIOXIA CM7-R Series SSD about 2,552 seconds to complete 5 epochs of training using 18 GPUs. Conversely, it took the Vendor A SSD about 2,906 seconds in the same scenario. Some models can be fully trained using 5 epochs but training with larger numbers of epochs is sometimes necessary to achieve higher training accuracy returned by the models. Given this circumstance, what if 100 epochs of training are required?

The same test was run using the same setup multiple times for additional iterations of training for 5 epochs, and the results remained consistent throughout without any depreciation in performance. Accordingly, the results for 100 epochs of training can be approximated by running this same test 20 times consecutively, and thus, the time it would take to complete 100 epochs of training can be estimated to be 20 times the result obtained for 5 epochs of training reported previously.

If 100 epochs of training are performed with 18 accelerators, it would take the KIOXIA CM7-R Series SSD about 14 hours[4] to complete, and the Vendor A SSD about 16 hours[5] to complete. Given 8,760 hours in a year, the KIOXIA CM7-R Series SSD could train over 625 models with 100 epochs of training each year (8,760 hours divided by 14 hours). Conversely, the Vendor A SSD could only train 547 models with 100 epochs of training each year (8,760 hours divided by 16 hours). The potential 12% efficiency gained by using the KIOXIA CM7-R Series SSD over the Vendor A SSD adds up to significant time savings as model training needs are scaled up. The KIOXIA CM7-R Series SSD is able to support training over more epochs faster for an individual model, and has the ability to support training multiple models faster, and get them deployed into production, when compared with the Vendor A SSD.

Test 2: Average Accelerator Utilization

This test measured the percentage of time that a compute resource was used to perform training on the ML workload. In order to decrease the amount of time it takes to train complex models on vastly large datasets, accelerator utilization must be high to ensure that efficient training is conducted. Low accelerator utilization indicates that the underlying storage system is unable to respond fast enough to keep the compute resources busy. The results are in percentage of use. The higher result is better.

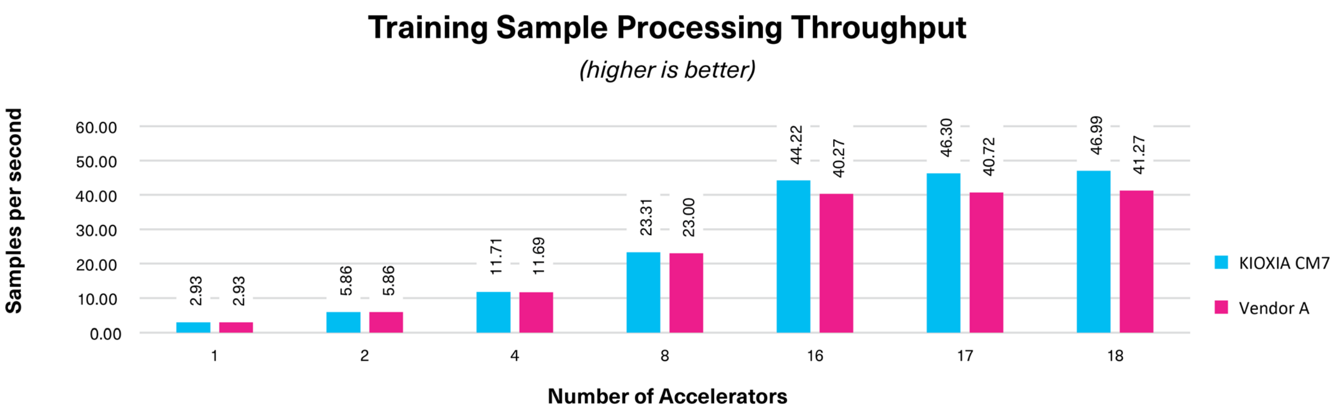


Analysis:

The graph shows that as the number of accelerators increased, the KIOXIA CM7 Series SSD was still able to keep accelerators efficiently utilized and service storage requests faster even under heavy accelerator load when compared with the Vendor A SSD.

Test 3: Training Sample Processing Throughput

This test measured the number of samples or individual units of data from the overall unet3d dataset that was processed per second. High sample processing throughput indicates that the compute resources are working efficiently. The more efficiently the compute resources are working, the faster they can read samples, have the training completed on them, and update parameters to the ML workload. The results are in samples per second. The higher result is better.

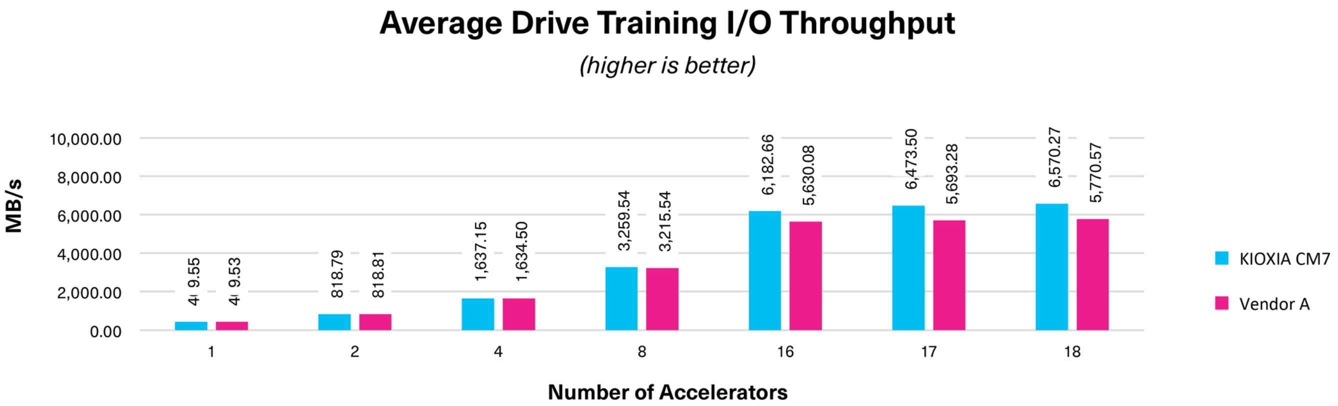


Analysis:

Higher sample processing throughputs lessen the total amount of time it takes to train a model from a training set. For this test, the KIOXIA CM7 Series SSD was able to respond faster when compared with the Vendor A SSD, enabling more samples to be processed in a shorter amount of time.

Test 4: Average Drive Training I/O Throughput

This test measured the average total throughput achieved from the underlying storage device while the ML workload was running. The results are in megabytes per second (MB/s). The higher result is better.

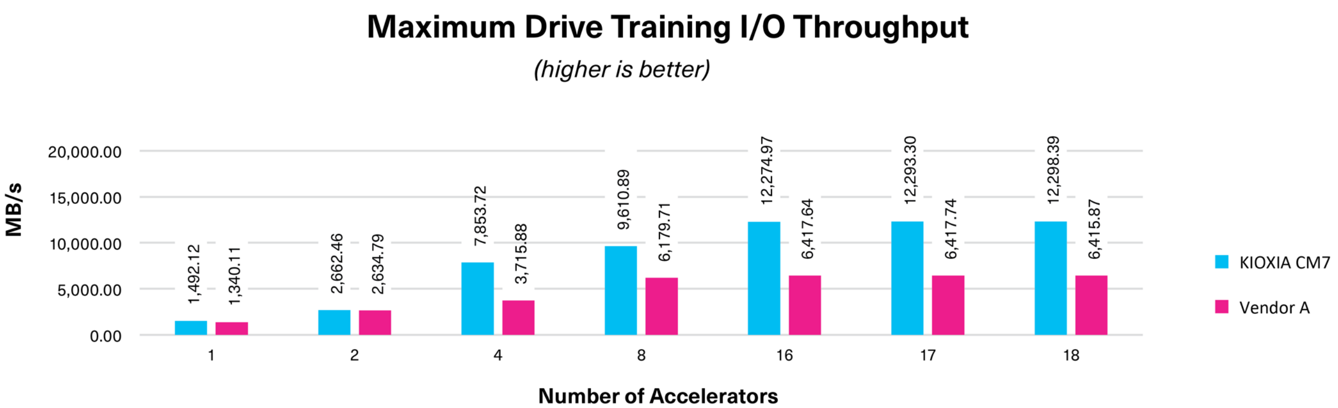


Analysis:

Maintaining higher I/O throughput over the course of the training time enables numerous data samples to be transferred into multiple accelerator's working memory concurrently. The graph shows that the KIOXIA CM7 Series SSD was able to handle more sample requests simultaneously from multiple requestors when compared with the Vendor A SSD.

Test 5: Maximum Drive Training I/O Throughput

This test measured the maximum throughput achieved from the underlying storage device while the ML workload was running. Many ML workloads perform in bursts as GPUs need the next batch of training data within an epoch. As such, storage devices that maintain high maximum storage throughput can quickly move multiple pieces of data within a batch to an emulated GPU. The results are in MB/s. The higher result is better.

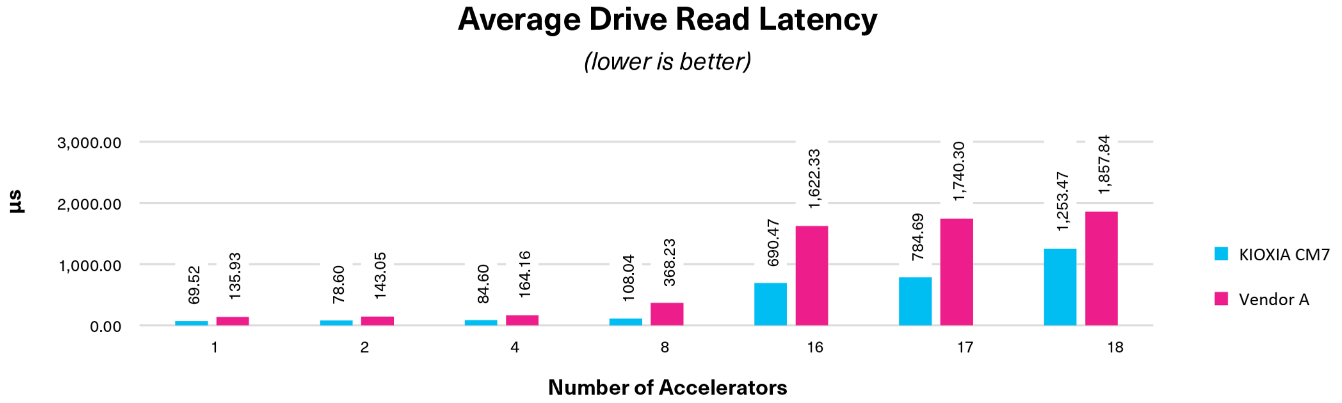


Analysis:

The graph shows that the maximum throughput achieved from the underlying storage device while the ML workload was running was almost twice as fast when 16, 17 or 18 accelerators were accessing data from the drive. In these circumstances, the KIOXIA CM7 Series SSD was able to provide 91% more I/O throughput when compared with the Vendor A SSD.

Test 6: Average Drive Read Latency

This test measured the time it took to perform a drive read operation. It included the average time it took for the workload generator to not only issue the read operation, but also the time it took to complete the operation and receive a 'successfully completed' acknowledgement. These results affect whether compute resources will be constantly in use and training on new samples. If the drive has low latency, compute resources can obtain a given sample faster and start training on it. This enables higher compute resource utilization, which in turn leads to the ML model to be updated faster. The results are in microseconds (µs). The lower result is better.



Analysis:

Fast response times enable end users to build out large scale multi-node systems for ML/AI workloads, where multiple compute resources, such as GPUs, can be better utilized when working in parallel. The graph shows that the KIOXIA CM7 Series SSD had faster response times when compared with the Vendor A SSD.

Additional Observations

For ML/AI applications, GPU compute time is a primary metric that can be changed in order to obtain significant performance gains. For example, switching ML/AI workloads from using a CPU to a GPU can provide a massive performance increase. Switching the GPU to a higher end SKU can also provide big performance increases. However, as GPUs get faster, there is still a concern regarding the speed that relevant data can be moved to these fast GPUs to train the ML models, which must be addressed by the underlying storage.

The ability for storage devices to deliver low read latencies and high average and maximum throughputs enable GPUs to be kept busy, which in turn can increase sample processing throughput training. The higher that this throughput increases, the less time it can take for the GPUs to train on large datasets. Saving even 12% on the total time to complete training, as depicted by the test results, can be a huge cost and time saver for a modern data center. Low read latencies and high throughputs also enables ML models to be updated more regularly with immense amounts of data, which in turn can help the ML model to be more accurate and versatile. The faster that GPUs can train on an ML model, the quicker these GPUs get updated, and more accurate models can be pushed to production on the inference side.

Upgrading to better storage can improve GPU utilization. The fact that a single KIOXIA CM7 Series SSD can handle 18 emulated GPUs at the same time is, in and of itself, an impressive feat. In modern data centers, IT personnel will ask how many GPUs a drive can handle in order to help them in forecasting their data center scaling requirements.

Lastly, latency will get higher when a large number of GPUs are all trying to access a drive at the same time. As GPUs are added during the testing, the latency for both drives begins to increase. However, the key point is when the same number of GPUs are used for each drive in the test comparison, the latency difference between the two is huge, significantly affecting the overall training times of the ML model. The KIOXIA CM7 Series SSD delivered huge latency improvements when compared with the PCIe® 4.0 Vendor A SSD such as 57% lower latency using 16 GPUs, 54% lower latency using 17 GPUs and 32% lower latency using 18 GPUs.

Summary

As presented by the test results, the KIOXIA CM7 Series SSD delivered up to 91% higher I/O throughput and up to 57% lower latency when compared with the Vendor A SSD. For ML/AI workloads, PCIe 5.0 SSDs deliver significant throughput and latency advantages that can help transition users from PCIe 4.0, or in some cases, from PCIe 3.0 products, to PCIe 5.0 solutions. For many other application workloads, customers may be resistant to upgrading because the storage solution they have is acceptable for the workloads they run, such as simple databases, data analytics, etc. However, for the new age of AI, the test comparison shows that upgrading to PCIe 5.0 SSDs can provide significant performance and latency improvements.

A major focus of the test comparison was to shorten the amount of time it takes to perform training on ML models. Though multiple factors contribute to this, the results show that a PCIe 5.0 SSD can lessen the amount of time required to train on ML models when compared with a PCIe 4.0 SSD. As the cost of data center or enterprise class GPUs can be expensive, the ability to conduct ML performance and latency testing by emulating a GPU, and using PCIe 5.0 SSDs and the DLIO benchmarking tool can be a cost effective route to improve system and SSD performance of ML/AI workloads.

KIOXIA CM7 Series SSD Product Info

The latest generation KIOXIA CM7 Series enterprise NVMe™ SSDs support 2.5-inch[6] and Enterprise and Data Center Standard Form Factor (EDSFF) E3.S form factors, and are compliant with the NVMe 2.0 and PCIe 5.0 specifications. These SSDs are available in two configurations: CM7-R Series for read intensive applications (1 DWPD[7], up to 30.72 TB capacities) and CM7-V Series for higher endurance mixed use applications (3 DWPD, up to 12.80 TB capacities). Additional features include a dual-port design for high availability applications, flash die failure protection to maintain reliability in case of a die failure and security options[8].

Additional KIOXIA CM7 Series SSD information is available here.

KIOXIA CM7 Series SSDs[9]
2.5-inch & E3.S form factors

Appendix A

The Various Steps of ML Testing

ML enables machines/systems to perform tasks that normally would require human intervention. It attempts to formulate complex models that allow accurate predictions to solve complex problems. A model becomes more accurate based on the amount of data and variety of data it is given. The more data that is available, the more examples a model can have to train and predict correctly. Additionally, the more diverse a dataset is, the more easily a model can tackle and learn new trends.

Hardware systems must be able to work fast and efficiently in order to process and train on the vast amounts of data that are needed to create an accurate model. This requires compute resources that handle the various steps of the ML process, such as preprocessing, training and inference.

In order to create a model, preprocessing must occur on the raw data so that it is cleaned and ready to be fed into the ML algorithms. This task involves transforming data such as images, sounds or even words/sentences, into a numerical vectorized format.

Once the data has been prepared and cleaned, the compute resources go through the training cycle and consume data in batches, so they can learn on the entire dataset. Models may go through many epochs or iterations on the same dataset so that the algorithms can learn the features they need to create an accurate model. This process can be very lengthy.

Although compute resources are becoming increasingly faster at performing these steps, they cannot start their work until the data is preprocessed and fetched from the underlying storage system. This requires the underlying storage system to be able to respond to multiple queries very quickly and with low latency. If the storage system lags behind, compute resources may become underutilized. In order to keep the many cores utilized at all times, the underlying storage system must deliver high sustained and burst throughput so that the compute resources can do training on different pieces of data in parallel.

Appendix B

Configuration Set-up

Server Information	
No. of Servers	1
CPU Model	AMD EPYC™ 9534
No. of Sockets	1
No. of Cores	64
Frequency (in gigahertz)	2.45 GHz
Memory Type	DDR5
Total Memory	768 GB[2]
Memory Frequency	DDR5-4800

SSD Information		
SSD Model	KIOXIA CM7-R Series	Vendor A
Form Factor	2.5-inch (U.2)	2.5-inch (U.2)
Interface	PCIe® 5.0 x4	PCIe 4.0 x4
Interface Speed (in gigatransfers per second)	128 GT/s	64 GT/s
No. of SSDs	1	1
SSD Capacity	3.84 TB[2]	3.84 TB
Drive Writes per Day	1	1

Operating System Information	
Operating System (OS)	Ubuntu®
OS Version	22.04.3 LTS
Kernel	5.15.0-89-generic

File System Information	
File System	XFS®
Mount Options	noatime, nodiratime

Test Software Information	
Test Software	DLIO
Test Software Version	GITHUB® GIT Commit ID: c1b4abc
ML Framework	PyTorch®
Dataset	unet3d
Dataset Size	3.47 TB
No. of Files	24,000

Appendix C

Configuration Set-up/Test Procedures

Configuration Set-up

One test server was set-up with an Ubuntu® 22.04.3 LTS operating system.

One 3.84 TB capacity KIOXIA CM7-R Series SSD was installed into the server.

The drive was set-up with an XFS® file system and mounted with the noatime and nodiratime flags.

The DLIO benchmarking tool was downloaded and set-up on the drive.

A unet3d dataset consisting of 24,000 files was created.

Test Procedures

The machine learning training was then started using the PyTorch® framework on the unet3d dataset, where an NVIDIA® V100 GPU was emulated and used to test the workload against the underlying storage.

The following metrics were recorded:

- *Total time to complete training (in seconds)*
- *Average accelerator utilization (in percentages)*
- *Training sample processing throughput (in samples per second)*
- *Average drive training I/O throughput (in MB/s)*
- *Maximum drive training I/O throughput (in MB/s)*
- *Average drive read latency (in microseconds)*

The configuration set-up and test procedures were repeated for Vendor A.

NOTES:

[1] DLIO is a data centric benchmark for scientific deep learning applications as published by H. Devarajan, H. Zheng, A. Kougkas, X. -H. Sun and V. Vishwanath, “DLIO: A Data-Centric Benchmark for Scientific Deep Learning Applications,” 2021 IEEE/ ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Melbourne, Australia, 2021, pp. 81-91, doi: 10.1109/CCGrid51090.2021.00018, on May 10-13, 2021.

[2] 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.

[3] Read and write speed may vary depending on the host device, read and write conditions and file size.

[4] The 14 hour result is based on this formula: 2,552 seconds * (100 epochs / 5 epochs) / 60 seconds / 60 minutes.

[5] The 16 hour result is based on this formula: 2,906 seconds * (100 epochs / 5 epochs) / 60 seconds / 60 minutes.

[6] 2.5-inch indicates the form factor of the SSD and not the drive’s physical size.

[7] DWPD: Drive Write(s) Per Day (DWPD). 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.

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

[9] The product images shown are a representation of the design model and not an accurate product depiction.

TRADEMARKS:

AMD EPYC and combinations thereof are trademarks of Advanced Micro Devices, Inc. GITHUB® is a registered trademark of GitHub, Inc. in the United States and other countries. NVIDIA is a registered trademark of NVIDIA Corporation. NVMe is a registered or unregistered trademark of NVM Express, Inc. in the United States and other countries. PCIe is a registered trademark of PCI-SIG. PyTorch is a registered trademark of The Linux Foundation. Ubuntu is a registered trademark of Canonical Ltd. XFS is a registered trademark of Silicon Graphics International Corporation or its subsidiaries in the United States and/or other countries. All other company names, product names and service names may be trademarks or registered trademarks of their respective companies.

DISCLAIMERS:

KIOXIA America, Inc. may make changes to specifications and product descriptions at any time. The information presented in this performance brief is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of KIOXIA America, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and KIOXIA America, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may render inaccuracies for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. KIOXIA America, Inc. assumes no obligation to update or otherwise correct or revise this information.

KIOXIA America, Inc. makes no representations or warranties with respect to the contents herein and assumes no responsibility for any inaccuracies, errors or omissions that may appear in this information.

KIOXIA America, Inc. specifically disclaims any implied warranties of merchantability or fitness for any particular purpose. In no event will KIOXIA America, Inc. be liable to any person for any direct, indirect, special or other consequential damages arising from the use of any information contained herein, even if KIOXIA America, Inc. are advised of the possibility of such damages.