SOLIDIGM + GIGABYTE™

Optimizing AI Training with Solidigm™ SSDs and Giga Computing Servers

Customer Challenge

As AI applications grow increasingly complex, organizations face mounting challenges in maximizing GPU utilization and maintaining efficient data throughput for AI training clusters. While storage systems play a crucial role in rapidly delivering training data to keep pace with high-performance GPUs, overall system architecture—including networking, memory hierarchies, and data preprocessing pipelines—also significantly impacts training efficiency. Addressing these challenges holistically is essential to minimizing bottlenecks, optimizing hardware utilization, and reducing operational costs.

For businesses aiming to scale their AI performance, partnerships help overcome the challenge to fully leverage their hardware investments and maximize training performance. Having focused options from Giga Computing along with Solidigm SSDs, customers can find new and improved ways to ensure they reach their goals with AI training.

GIGABYTE Server Solution

Giga Computing collaborated with Solidigm to address these challenges by integrating the high-performance Solidigm D7-PS1010 and the high-capacity D5-P5336. By utilizing the SSDs, they demonstrated significant improvements in GPU utilization and overall training efficiency.

Using the MLPerf Storage v1.0 benchmark — a leading industry standard for evaluating storage performance in AI workloads — Giga Computing and Solidigm tested the solutions across a range of AI models, including Unet3D, ResNet50, and CosmoFlow. Specifically, the Solidigm D7-PS1010 PCIe 5.0 SSD was compared to Micron's 9550 PCIe 5.0 SSD offering, and the Solidigm D5-P5336 PCIe 4.0 SSD was compared to Micron's 6500 ION PCIe 4.0 offering.

Benchmarks Utilized

MLPerf is a standardized benchmark designed to evaluate how efficiently storage systems can supply AI/ML pipeline training data efficiently to the GPU. It is developed and maintained by MLCommons, a collaborative organization that includes academia, industry leaders and AI researchers.

Within the ML Perf suite, we utilized the MLPerf storage benchmark, which is designed to measure how effectively storage solutions can supply data to keep accelerators, such as GPUs, efficiently utilized during ML model training. The benchmark includes simulations of modern accelerators and focuses on three primary workloads:

Unet3D: A model designed for volumetric segmentation,
commonly used in medical imaging to delineate structures within 3D scans.

ResNet50: A deep residual network comprising 50 layers, widely employed for image classification tasks.

CosmoFlow: A 3D convolutional neural network applied to cosmological data,
aiming to predict physical parameters of the universe from simulation data.

These workloads encompass a range of data sizes and access patterns, providing a robust framework for assessing storage performance in diverse ML applications. By simulating the "think time" of accelerators, the benchmark can accurately model storage demands without necessitating actual GPUs, making it accessible for evaluating various storage solutions.

AI Models	Usage	AU (Accelerator Utilization)
Unet3D	Medical - Image Segementation	>90%
ResNet50	Vision - Image Classification	>90%
CosmoFlow	Scientific - Cosmology parameter prediction	>70%

Performance and Efficiency Gains

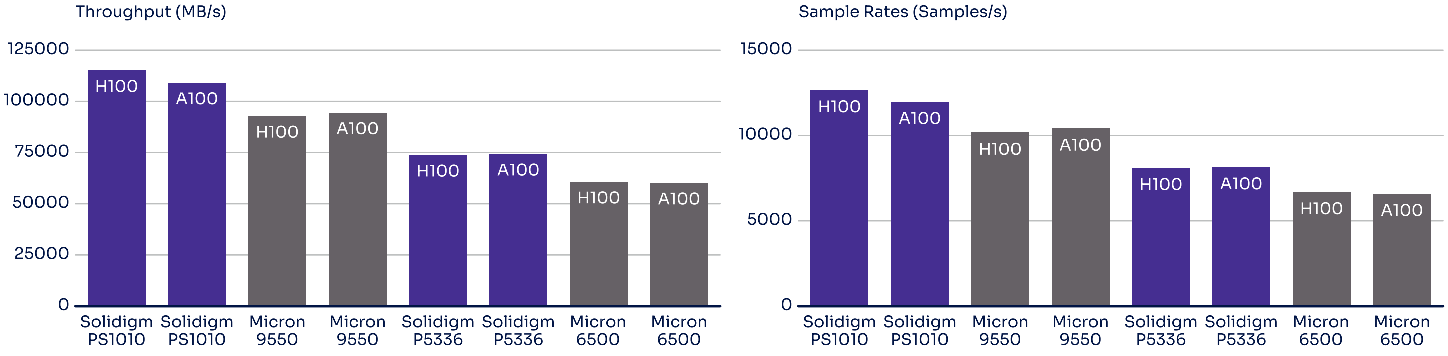
Enhanced GPU Utilization

Across AI models, the introduction of Solidigm SSDs significantly increased the number of GPUs supported during AI training. For example, in the ResNet50 example, the Solidigm D7-PS1010 exhibited an Accelerator Utilization (AU) of 92.68% and supported 70 GPUs on NVIDIA H100 systems, while the Micron 9550 only exhibited an AU of 90.33% and supported 58 GPUs. The Solidigm SSD ensures that expensive GPU resources are used to their fullest potential, reducing idle time and accelerating model training.

Superior Throughput and Read Performance

The combination of GIGABYTE systems and Solidigm SSDs delivered outstanding throughput and read speeds. For instance, the Solidigm D7-PS1010 achieved throughput of 115,805 samples/s and 12,663 MB/s on ResNet50 with NVIDIA H100 GPUs, simulating power to 70 GPUs. This outperformed the Micron 9550 drive by nearly 24%. These results translate into faster data processing model training.

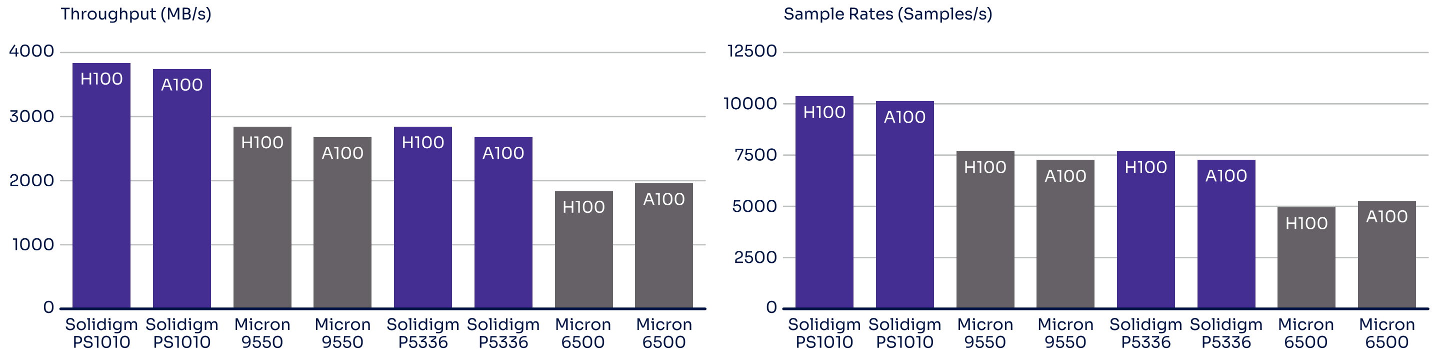
ResNet50



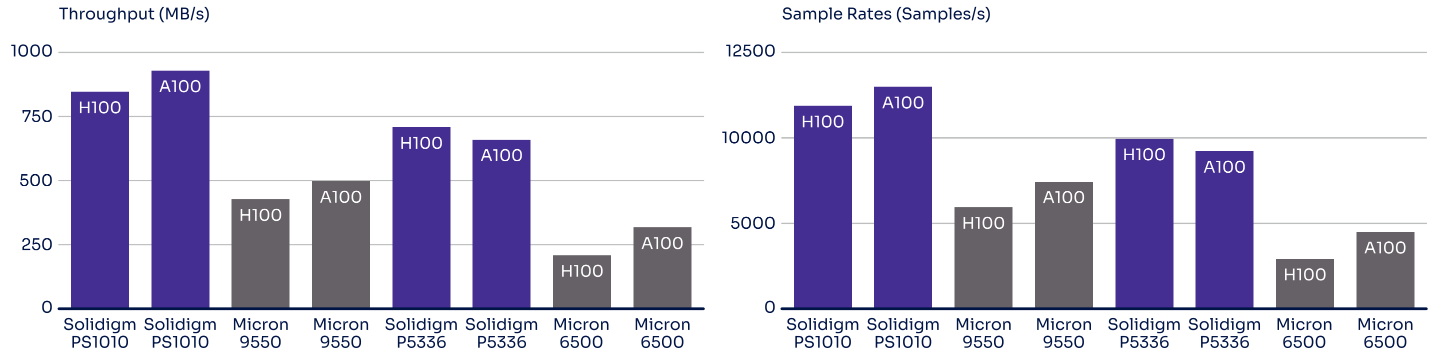
Efficiency Gains Across Multiple AI Models

Whether training Unet3D for medical imaging or Cosmoflow for cosmological simulations, Solidigm SSDs consistently outperformed competing solutions. The Solidigm D5-P5336 SSD, for example, showed a 22% improvement in throughput over the Micron ION 6500 in ResNet50 workloads.

Unet3D



CosmoFlow



Future-Proofing with Gen5 SSDs

The current results comparing the Solidigm D5-P5336 PCIe 4.0 SSD to the Micron 6500 ION PCIe 4.0 SSD are impressive for capacity focused solutions in network attached storage. This shows that Solidigm continues to use their deep knowledge of AI to optimize products, enabling the value of solutions across the portfolio within a growing family of products. The recently introduced Solidigm D7-PS1010 PCIe 5.0 SSD family reduces latency when needed and improves data flow in direct attached storage needs, setting the stage for even more efficient AI training clusters.

Real-World Impact

As a result of this collaboration, Giga Computing can now lead with Solidigm SSDs for AI-focused engagements. The demonstrated performance gains have already attracted interest from leading organizations such as BeeGFS, Mangoboost, and NCHC.

More Information

For more details on this collaboration and to explore how Solidigm and Giga Computing can enhance your AI infrastructure, visit:

Solidigm: solidigm.com
Giga Computing: gigacomputing.com
MLPerf Benchmark Details: mlcommons.org

Discover how Solidigm and Giga Computing are redefining AI training performance through cutting-edge storage and system integration. Upgrade your AI infrastructure today and unlock new levels of efficiency and scalability.

SOLIDIGM™ + GIGABYTE™

About Solidigm

Solidigm is a leading global provider of innovative NAND flash memory solutions. Solidigm technology unlocks data's unlimited potential for customers, enabling them to fuel human advancement. Originating from the sale of Intel's NAND and SSD business, Solidigm became a standalone U.S. subsidiary of semiconductor leader SK hynix in December 2021. Headquartered in Rancho Cordova, California, Solidigm is powered by the inventiveness of team members in 13 locations around the world. For more information, please visit solidigm.com and follow us on Twitter and on LinkedIn. "Solidigm" is a trademark of SK hynix NAND Product Solutions Corp. (d/b/a Solidigm).

About Giga Computing

Giga Computing Technology is an industry innovator and leader in the enterprise computing market. Having spun off from GIGABYTE, we maintain hardware expertise in manufacturing and product design, while operating as a standalone business that can drive more investment into core competencies. We offer a complete product portfolio that addresses all workloads from the data center to edge including traditional and emerging workloads in HPC and AI to data analytics, 5G/edge, cloud computing, and more. Our longstanding partnerships with key technology leaders ensure that our new products will be the most advanced and launch with new partner platforms. Our systems embody performance, security, scalability, and sustainability. To find out more, visit https://www.gigacomputing.com/en/ and join our newsletter.

Appendix

Measure following MLPerf Storage benchmark required metrics.

Throughput (MB/s): The rate at which dataset is loaded from storage.

Dataset size (GB): Total size of the dataset used.

Accelerator Utilization (%): Accelerator Utilization (AU) is defined as the percentage of time taken by the simulated accelerators, relative to the total benchmark running time. Higher is better.

AU (percentage) = (total_compute_time/total_benchmark_running_time) * 100

Server system and tuning used in the Micron results are different than the Solidigm + Giga Computing setup. The Micron results can be found at: Optimizing AI Systems With Micron's NVMe SSDs

Server	GIGABYTE R163-Z35
CPU	AMD 9555P - Truin - 64 core x 1
Memory	M321R8GA0BB0-CQKMG - Samsung - 64 GB DDR5 x 12
OS Disk	Intel 1.6 TB x 2, U.2
Test Target SSD 1	Solidigm D5-P5336
Test Target SSD 2	Solidigm D7-PS1010

SOLIDIGM™ + GIGABYTE™