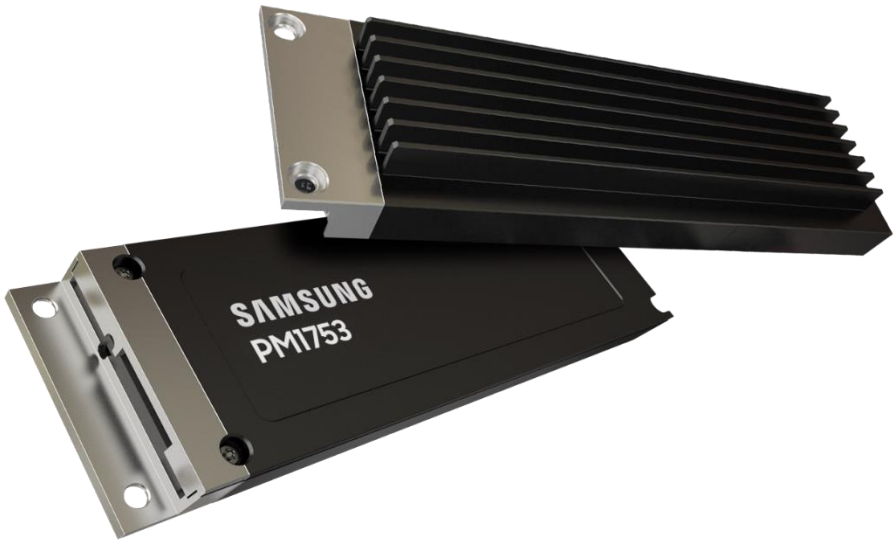
MLPerf Storage benchmark results for Samsung NVMe SSDs

Get highly accelerated performance
with the Samsung Enterprise NVMe SSD optimized for AI application data



Introduction

MLPerf Storage Benchmark Suite for NVMe SSDs

Traditionally, SSD performance in data centers has been evaluated based on four key characteristics: sequential read/write and random read/write. However, as AI applications diversify, GPU-based systems are interacting with SSDs in increasingly varied ways to support operations such as training and inference. In this context, tools like the MLPerf™ Storage benchmark offer valuable metrics that reflect how efficiently a storage system can deliver training data during model training.

A wide range of data center SSDs can be selected to enhance the efficiency of AI platforms. However, the requirements for these SSDs require more complex consideration beyond the traditional four key performance metrics typically used for evaluation. To better predict and optimize storage behavior in AI workloads, the MLPerf Storage benchmark can be used to evaluate deep learning workloads on systems equipped with NVIDIA® A100 and H100 GPU accelerators.

Workloads from MLPerf Storage

The comparison presented here utilizes three representative workloads provided by MLPerf Storage. Each workload simulates dataset access patterns observed during training in real-world deep learning environments across computer vision and scientific computing domains. The benchmark testing can run even in systems without GPUs – the level of compute performance of the benchmark can be configured by specifying the type and number of accelerators for each workload, which in turn affects the intensity of I/O requests to the SSD under test.

- **3D U-Net: Medical image segmentation**
  The 3D U-Net workload is designed for 3D image segmentation tasks. A representative use case of this model is medical image segmentation. This workload performs large-chunk read operations based on the PyTorch framework and uses a training dataset in NPZ format.

- **ResNet50: Image classification**
  The ResNet50 workload is designed for image classification tasks using neural networks. It performs read operations with varying chunk sizes based on the TensorFlow framework and uses a training dataset in TFRecord format.

- **CosmoFlow: Cosmology parameter prediction**
  The CosmoFlow workload is designed for predicting cosmological parameters from 3D simulation data. It is characterized by large sample file sizes, while each individual sample remains relatively small. This workload performs mixed-size chunk read operations based on the TensorFlow framework and uses a training dataset in TFRecord format.

Model	Sample file size (MB)	# of samples per file	Framework / Data loader	Format	H100 computation time (s)
3D U-Net v1.0	142	1	PyTorch	NPZ	0.323
ResNet50 v1.0	137	1251	TensorFlow	TFRecord	0.224
CosmoFlow v1.0	2.8	1	TensorFlow	TFRecord	0.0035

When monitoring SSD behavior during sample file processing, it becomes apparent that all workloads exhibit small sequential I/O patterns that repeatedly access random addresses. From the SSD's perspective, these patterns are neither purely random nor fully sequential. As a result, SSD performance can vary significantly during workload execution.

The MLPerf Storage benchmark provides multiple performance metrics to help correlate these I/O patterns with the storage system's processing characteristics. In this paper, we analyze the impact of SSDs on each workload from two perspectives: Throughput and Accelerator Utilization (AU).

Throughput

The throughput metric in the MLPerf Storage benchmark is measured in samples per second: the faster the dataset can be accessed, the higher the SSD's read throughput. The metric generally scales linearly with the SSD's read throughput, but does not represent the SSD's absolute peak performance. Benchmark results vary depending on the underlying storage throughput.

The specific throughput values generated by the benchmark are influenced by the following behaviors.

- SSD performance initially increases as accelerators are added, depending on the type and number of accelerators used.
- As the number of accelerators is increased beyond a certain point, the SSD access becomes a bottleneck and performance improvement tends to stall when the PCIe bus becomes saturated.
- Depending on the workload processing pattern, there can be a significant gap between the maximum and minimum throughput (observed samples per second).
- Even under the same test conditions, results between devices may vary depending on the SSD's read processing capabilities (such as block size handling efficiency).

In Figure 1, the throughput of the PM1753 SSD is compared to that of its chief competitor. The values shown below represent average values obtained over five epochs for each workload. Since the unique I/O pattern of each workload remains consistent regardless of the number of accelerators, the maximum performance of each SSD reflects its ability to handle that specific pattern.

The PM1753 SSD demonstrates high performance against its competition under identical load conditions thanks to its efficient data processing capabilities, and excels particularly in handling large datasets.

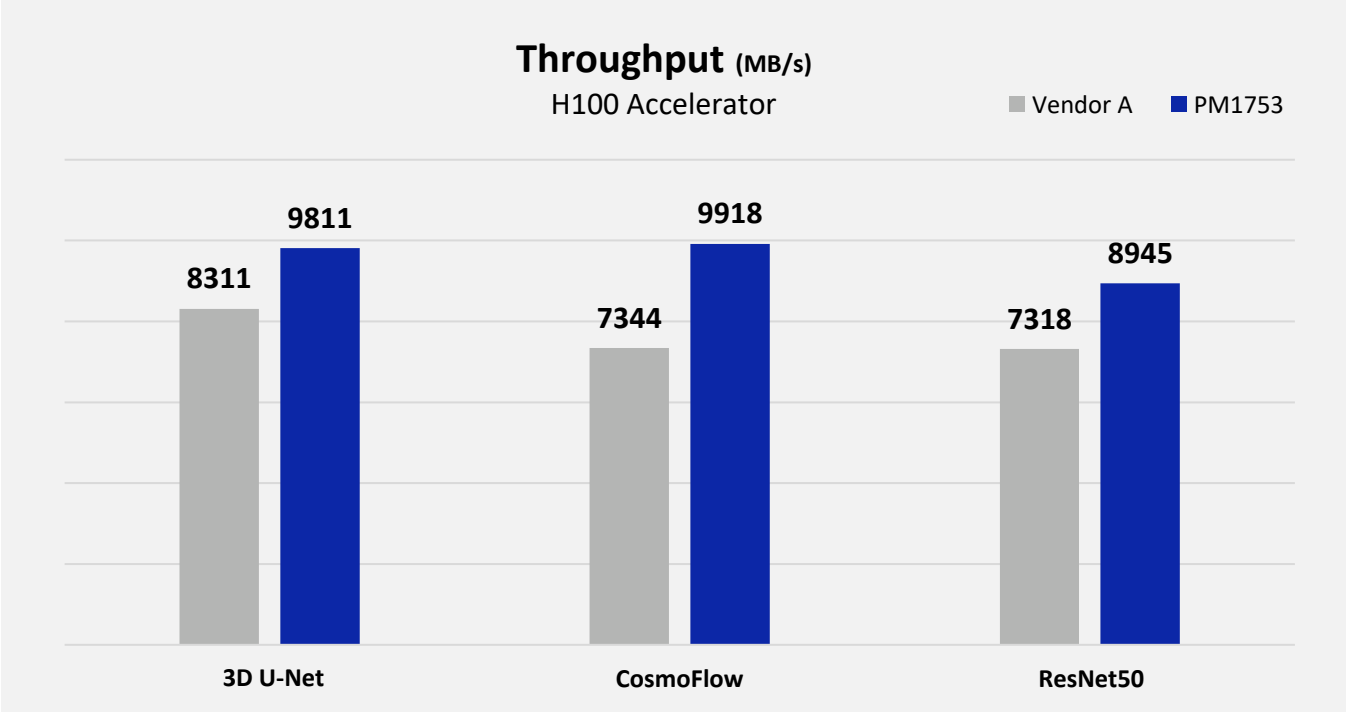


Figure 1: Maximum SSD throughput by workload

1. The experimental results were obtained without modifying the read thread count parameter from the original workload settings published on GitHub. Internal experiments confirmed that increasing the read thread count led to higher throughput.

Accelerator Utilization (AU)

Accelerator Utilization (AU) indicates how well the GPUs are being kept busy, and is calculated as the ratio of total compute time to total benchmark runtime. An ideal compute time is determined based on the batch size, total dataset size, and the number of simulated accelerators, and used for the benchmark. The higher the calculated AU, the better the overall system efficiency. The specific AU values generated by the benchmark are influenced by the following behaviors.

- Each workload has a defined AU minimum threshold, which increases as the number of GPUs increases.
- As noted previously, increasing the number of GPUs leads to higher SSD throughput until bus saturation occurs.
- High AU corresponds to high throughput only up to a point. Once the GPUs are fed sufficient data to keep them occupied, further increases in throughput do not significantly increase AU.
- Consequently, even with the same number of accelerators, AU values differ among SSDs depending on how they process data access. The maximum number of accelerators for which a given SSD can satisfy the AU minimum threshold is heavily influenced by SSD logic design.

The relationship between workload performance and the number of accelerators is well-captured by the AU metric. By analyzing both AU and throughput, it is possible to assess the maximum performance an SSD can deliver relative to the number and type of accelerators used.

In Figure 2, the AU of the PM1753 SSD is compared to that of its chief competitor. The information shown represents average values measured over five epochs for each workload. The results illustrate how performance varies with different accelerator configurations.

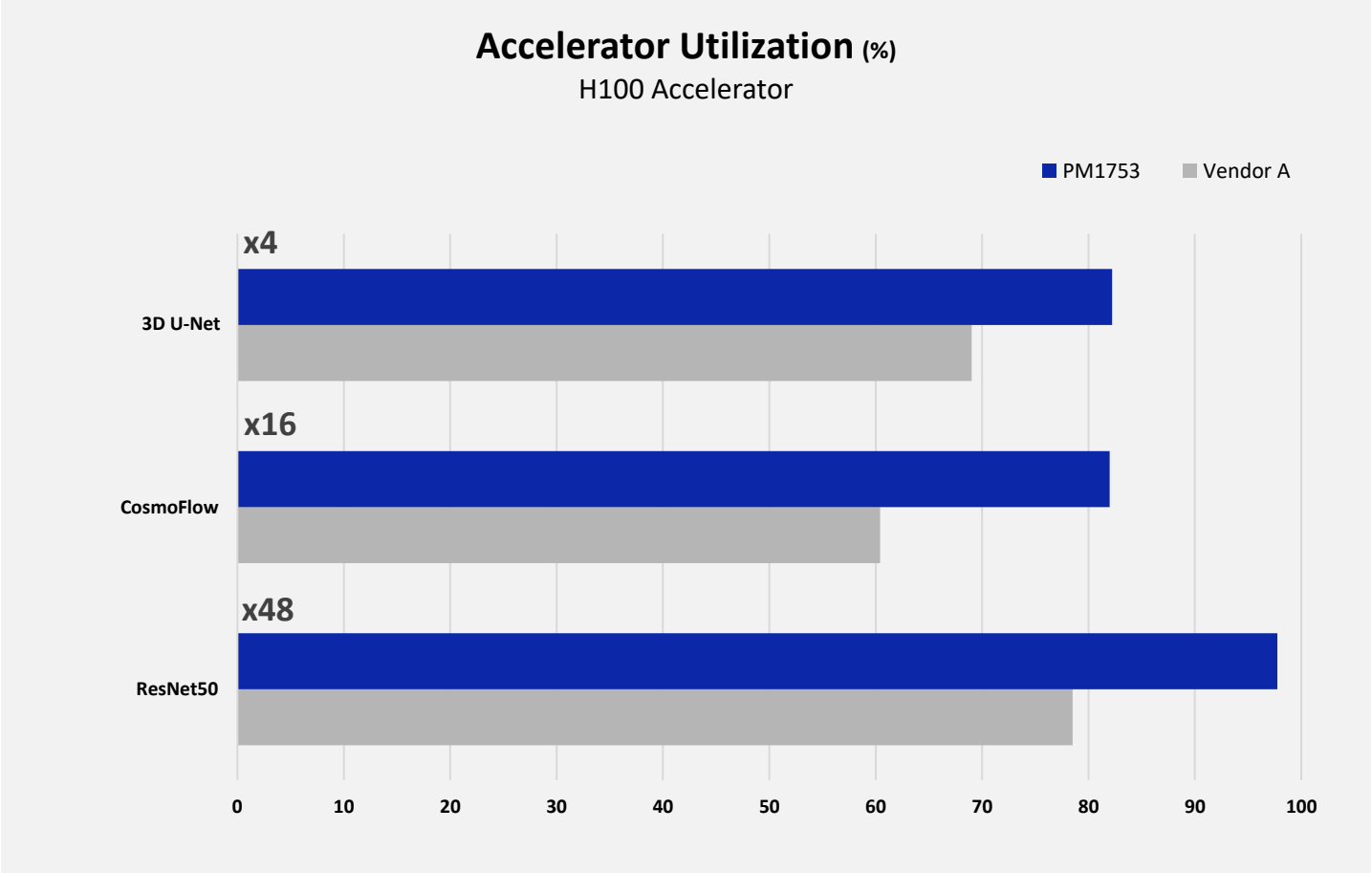


Figure 2: AU per workload when the number of accelerators is fixed

2. The above experimental results were obtained using the default read thread count as defined in the workload parameters published on GitHub.

Identifying Inefficiencies

A key observation in these experiments is that, due to limitations of a given SSD's load-handling capability, there may be a point where increasing the number of accelerators leads to inefficiencies—reflected as longer overall execution times. This inflection point can be identified quantitatively through the AU value.

In a real-world multi-GPU training environment, a higher AU indicates shorter overall training time and reduced GPU idle time, resulting in improved system efficiency.

This concept can be more clearly illustrated by examining the processing rate in samples per second (Figure 3) and total execution time (Figure 4) for each workload on the SSD. A higher sample rate corresponds to a shorter total execution time. Since the computation time per step is fixed by the workload parameters, differences in total execution time are primarily determined by the SSD's I/O processing capabilities.

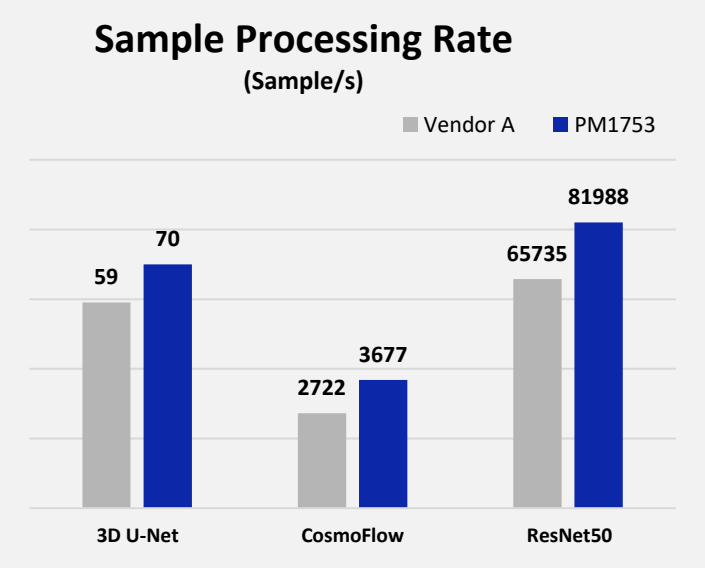


Figure 3: Sample rate by workload

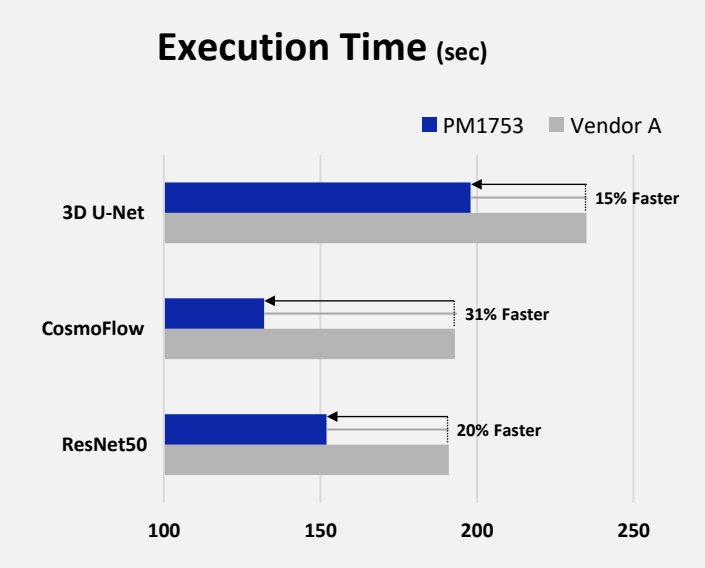


Figure 4: Execution time by workload

Power Consumption

From the perspective of a real GPU environment, execution time affects not only training efficiency but also the power consumption of the H100 accelerators in use. Since each workload assumes a multi-GPU setting, we can estimate energy consumption by multiplying the number of H100 accelerators by the total execution time required to process the same dataset. Therefore, a shorter execution time implies greater energy efficiency and reduced power usage.

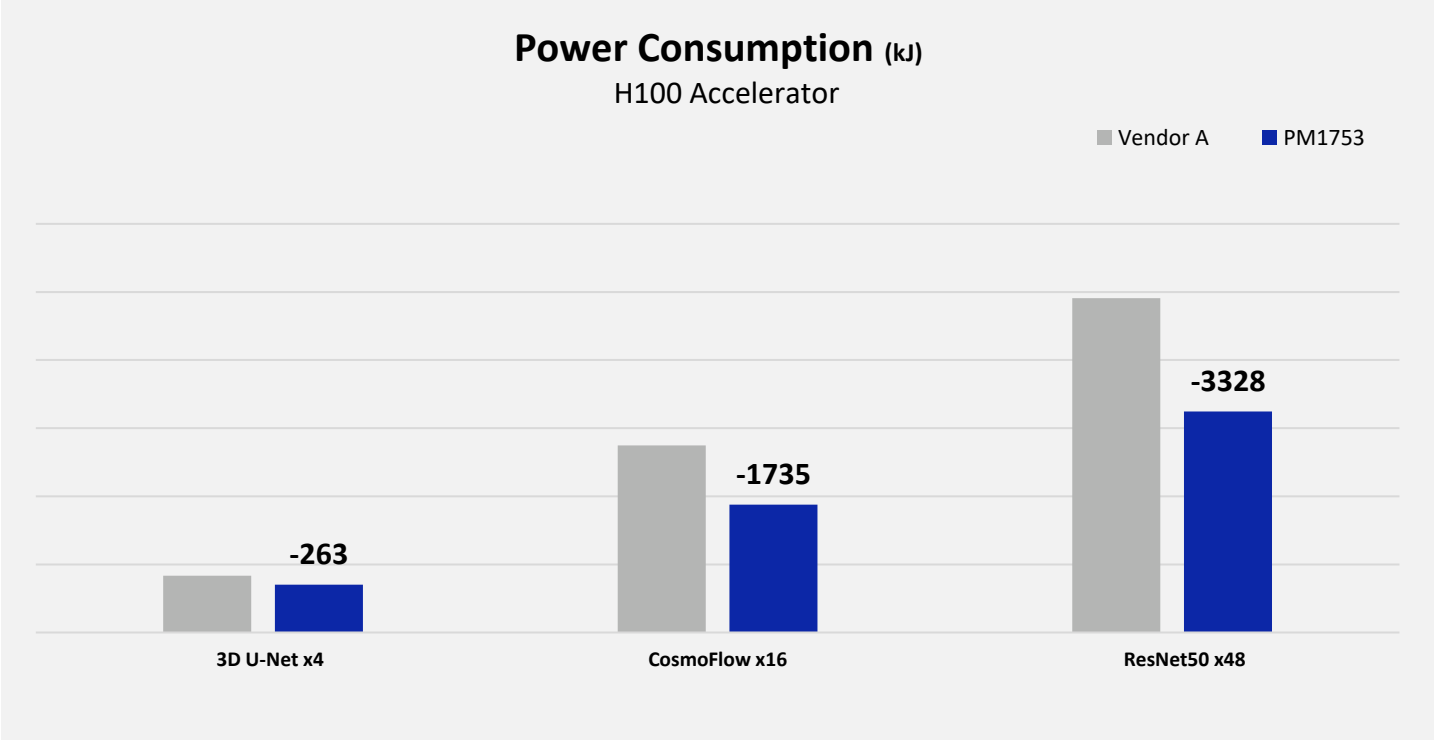


Figure 5: Power Consumption

3. For additional details on H100 power consumption, refer to the relevant article on the Power Electronics News website.
4. Actual power consumption values may vary depending on the GPU's idle entry pattern during training.

Workload Characteristics

All workloads in MLPerf Storage assume a multi-GPU environment by default, with the number of accelerators explicitly specified. Conceptually, each workload accesses a large dataset from multiple GPU hosts. As a result, the order in which each GPU accesses the dataset is not deterministic, leading to mixed sequential I/O operations with varying block sizes being issued to the SSD.

This I/O behavior deviates from the conventional four major I/O patterns, and requires more advanced logic at the SSD level to handle operations efficiently.

Although the I/O pattern in Figure 6 may appear entirely random, a closer examination of each I/O group reveals underlying sequential read patterns within smaller units. Due to the complexity of the I/O access pattern, the device alternates between random and sequential reads.

This behavior arises from the conceptual model of multiple GPUs concurrently reading large dataset files from the host. As the number of accelerators or read threads increases, the complexity of the access pattern can further intensify.

In future AI training environments involving large sample datasets such as images and graphical data across many GPUs, the processing performance of SSDs like the one evaluated above may play a critical role in improving overall training efficiency.

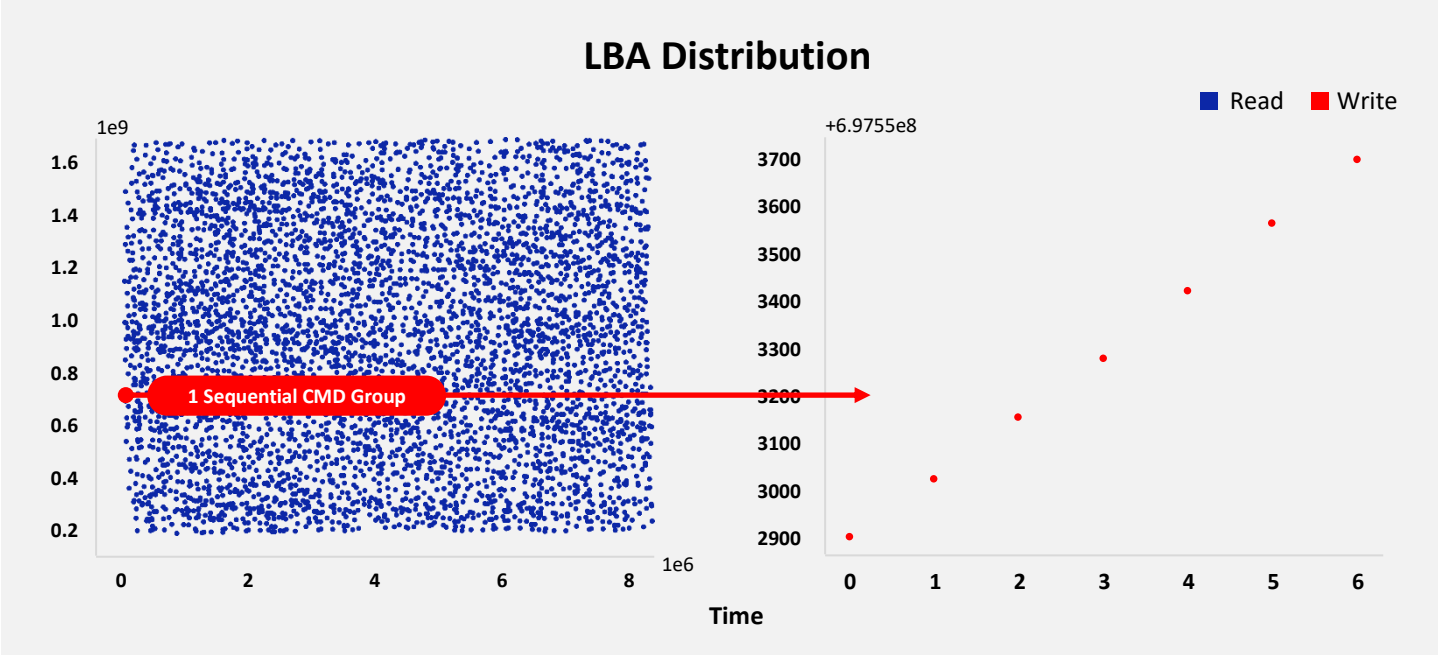


Figure 6: Logical Block Address (LBA) distribution graph of SSD I/O during benchmark execution

5. The block sizes primarily handled by SSDs are typically 128 KB or larger; however, they can vary depending on the file system and workload characteristics.
6. The LBA access distribution differs across the three workloads.

Conclusion

In the future, SSDs for AI platforms will be expected to support a variety of operating modes beyond those that target current standardized performance requirements. These modes will focus on the hybrid sequential-random access patterns essential to high-performance AI storage operations. Additionally, the size of datasets used for AI training such as images and videos is anticipated to grow significantly, especially in multi-GPU environments.

In such scenarios, selecting SSDs capable of processing large volumes of data without introducing system bottlenecks will be essential to maximizing overall training and retrieval efficiency.

The MLPerf Storage benchmark results offer valuable insights into future storage requirements. Our analysis of the MLPerf Storage workloads reveals that SSDs must handle a variety of block sizes and complex access patterns that may not necessarily be optimized in today's SSDs.

Samsung's PM1753 NVMe SSD excels at traditional performance standards while also delivering high throughput and capacity optimized for such demanding environments. As AI architectures continue to evolve rapidly, the ability to proactively address these emerging performance demands will become increasingly important – and Samsung will remain at the forefront of researching these demands and implementing solutions to stay ahead of the needs of AI accelerator toechnology.

Appendix: Implementation Notes

MLPerf Storage dataset volumes are structured differently based on the available host memory[1]. Additionally, the maximum performance for each workload can vary depending on the host system architecture. SSD performance may also be influenced by the type and number of sample files generated per workload, as well as the type of accelerator used within the same environment. Therefore, performance measured in a single environment should not be considered a definitive specification of the storage itself.

In general, the impact of storage becomes more pronounced when evaluated on high-performance servers (e.g., systems with more CPU cores, higher per-core performance, and better memory bandwidth). Even with identical parameters and SSDs, the sampling rate per second can vary depending on the characteristics of the server used for evaluation.[2]

We did not alter any components of the workload. Instead, we focused on maximizing storage I/O and minimizing execution time to obtain optimal test results. Accordingly, we isolated system resources as much as possible and minimized potential sources of interference within the evaluation environment during benchmarking.

PM1753 SSD Test Platform	
Server platform	Gigabyte R283-Z95-AAV1
CPU	AMD EPYC™ 9655 (96-Core Processor)
Memory	256GB
Storage	Samsung PM1753 SSD 7.68TB
OS	CentOS Linux release 8.5.2

7. Depending on the host memory configuration, the number of datasets generated is adjusted to minimize caching effects.
8. A larger number of datasets increases SSD capacity usage and can also affect sampling throughput.
9. All results are unverified MLPerf v1.0 Storage scores. Results not verified by MLCommons Association. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

For more information
For more information about the Samsung Semiconductor products, visit semiconductor.samsung.com

About Samsung Electronics Co., Ltd.
Samsung Electronics Co. Ltd inspires the world and shapes the future with transformative ideas and technologies. The company is redefining the worlds of TVs, smartphones, wearable devices, tablets, digital appliances, network systems, memory, system LSI and LED solution.
For the latest news, please visit the Samsung Newsroom at news.samsung.com

Samsung Electronics Co., Ltd.
1-1, Samsungjeonja-ro, Hwaseong-si, Gyeonggi-do 18448, Korea
www.samsung.com 1995-2021

SAMSUNG