Technical Brief

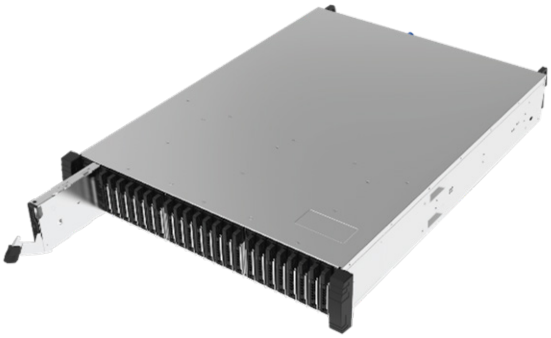
Benchmarking the Ingrasys ES2100 with Western Digital RapidFlex™ A2000 Fabric Bridge using the Deep Learning I/O Benchmark

Executive Summary

This paper presents the benchmarking results achieved, using the Deep Learning I/O Benchmarking (DLIO) tool. The Ingrasys ES2100 chassis is equipped with Western Digital RapidFlex™ A2000 interposers and Kioxia CM7 NVMe™ SSDs.

The system was evaluated using DLIO workloads to assess I/O throughput, consistency, and virtual GPU (vGPU) award. Each RapidFlex interposer provides a 1:1 PCIe® mapping to its corresponding NVMe SSD, ensuring predictable and balanced performance across the entire chassis. This benchmark is not an endorsement of the Ingrasys product by Western Digital and no warranty of the product is expressed or implied by Western Digital.

Ingrasys ES2100 NVMe-oF™ EBOF



Ingrasys redefines next-generation storage with the ES2100, an NVMe-over-Fabrics (NVMe-oF) solution purpose-built for AI and HPC workloads. The 2U storage system supports E3.S and U.2 SSD form factors and features a midplane-less design for enhanced airflow and reliability.

The system is equipped with two hot-swappable, redundant switch modules (SWMs) that not only provide high availability but also allow effortless future system upgrades simply by replacing the SWMs — without impacting the overall architecture.

For additional information on the Ingrasys ES2100 NVMe-oF storage system, see: https://www.ingrasys.com/solutions/15/nvme-of_ebof_solution/

Western Digital RapidFlex A2000 Fabric Bridge Device

Western Digital RapidFlex A2000 is a low-power, high-performance NVMe-oF fabric bridge device that enables more efficient use of NVMe-based storage. NVMe over Fabrics, or NVMe-oF allows external NVMe SSDs to be shared at comparable latency as if they were inside the server. In addition to better efficiency, this disaggregation separates storage from the server refresh cycle. This new version adds initiator capability that allows datacenters to deploy much lower power initiator cards in their servers. This new version also doubles performance over prior generations with an additional 100 GbE port matched to 16 lanes of Gen 4 PCIe.

The Kioxia CM7-V 6.4TB SSD were the NVMe drives that were used in this benchmark.

CM7-V Drive Details

Drive	Kioxia CM7-V
Form Factor	U.3 15mm
Interface	PCIe Gen5, NVMe 2.0
Security	SIE
Power	25W (Active)
Power Idle	5W
Part Number	KCMYXVUG6T40

Deep Learning I/O Benchmark (DLIO)

DLIO is an open-source benchmark developed by NVIDIA. It is specifically designed to model and measure the I/O behavior of deep learning workloads—that is, how efficiently data can be read from or written to storage systems when training or inferring with large models such as ResNet-50 or 3D U-Net.

Typical DLIO metrics include:

- Throughput (MB/s or GiB/s) — effective rate of data feeding into GPUs
- Latency and jitter — consistency of data delivery
- Scalability — performance as the number of GPUs or nodes increases
- CPU utilization and I/O overhead

DLIO measures how effectively the I/O subsystem feeds each active GPU (or vGPU) during data-intensive AI training. The "vGPU" count shown in DLIO results corresponds to the number of logical GPU instances being used by the benchmark—each representing an independent training pipeline consuming data from storage. In turn, this gives a representation of the number of physical GPU that a storage subsystem may cater to. In these benchmarks the NVIDIA® H100 was the vGPU chosen.

Benchmark Runs

ResNet-50

The ResNet-50 model represents a canonical image classification workload used extensively across deep learning benchmarks. It employs a deep convolutional neural network architecture composed of 50 layers with residual connections that mitigate vanishing gradients, allowing efficient training of very deep models. Its data pipeline primarily involves large numbers of small image files (typically JPEG or PNG images resized to 224×224 pixels) , which are read repeatedly during each training epoch.

From an I/O perspective, ResNet-50 places stress on metadata lookup, small sequential reads, and caching efficiency rather than raw throughput. Each vGPU requires a steady stream of modestly sized files at high request rates, meaning that latency and metadata handling can dominate performance, especially in large-scale parallel environments.

The ResNet-50 test is implemented to replicate this real-world pattern using a configurable dataset generator that produces many small samples, each corresponding to an image record. The framework emulates the PyTorch DataLoader behavior by spawning concurrent I/O threads per vGPU, each reading independent mini-batches. During the benchmark run, DLIO measures aggregate throughput, per-vGPU read rates, and tail latencies while simulating the random access and data augmentation stages typical of ResNet-50 training. This profile is valuable for assessing file system scalability, caching effectiveness, and small-I/O performance of disaggregated storage platforms.

3D U-Net

The 3D U-Net model is used in medical image segmentation and is markedly different from ResNet-50 in its data access pattern. Instead of small 2D images, it processes large volumetric datasets such as MRI or CT scans represented as 3D tensors. These files are often hundreds of megabytes in size and are accessed in contiguous blocks to form volumetric patches for training. The model architecture itself consists of encoder–decoder stages with skip connections, emphasizing high-resolution reconstruction and spatial coherence.

I/O demands for 3D U-Net are dominated by large sequential reads and high aggregate bandwidth. Each vGPU consumes multi-megabyte data chunks, often with fewer metadata lookups but heavier sustained data streaming requirements. This makes the workload an ideal stress test for bandwidth scalability and sustained read throughput across RDMA fabrics or NVMe-oF targets.

The 3D U-Net test is implemented by generating synthetic volumetric datasets and configuring larger sample and batch sizes to emulate the streaming behavior of real medical data. Each vGPU reads contiguous blocks concurrently, allowing DLIO to measure how effectively the storage infrastructure delivers high data volumes without bottlenecking GPU training. The results highlight the system's ability to maintain consistent throughput at scale, revealing whether the performance is limited by network transport, queue depth saturation, or SSD parallelism.

Testbed Configuration

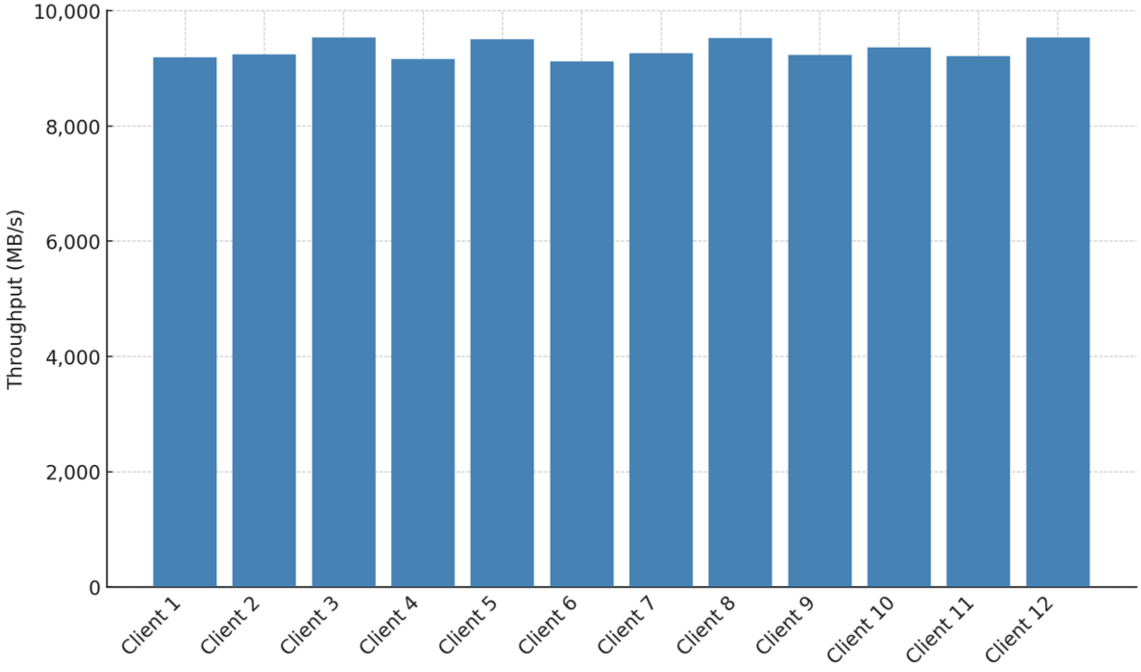
The Ingrasys ES2100 EBOF chassis was populated with 24 Kioxia CM7-V PCIe Gen4x4 NVMe SSDs. Each drive is network attached via a Western Digital RapidFlex A2000 interposer that provide a direct, low-latency PCIe path to the host over Ethernet. This setup eliminates bottlenecks found in traditional PCIe switch fan-out designs.

Each client system used in testing was a Dell® PowerEdge R750 configured as follows:

- CPUs: Dual Intel® Xeon® Gold 6354 processors
- Memory: 512 GiB DDR4 ECC (3200 MT/s)
- Networking: 2× NVIDIA ConnectX®-6 200 GbE NICs
- Fabric Switch: NVIDIA Spectrum®-3 SN4700

Results Overview: ResNet-50 Benchmarking

Observations on client throughput



DLIO Client Throughput

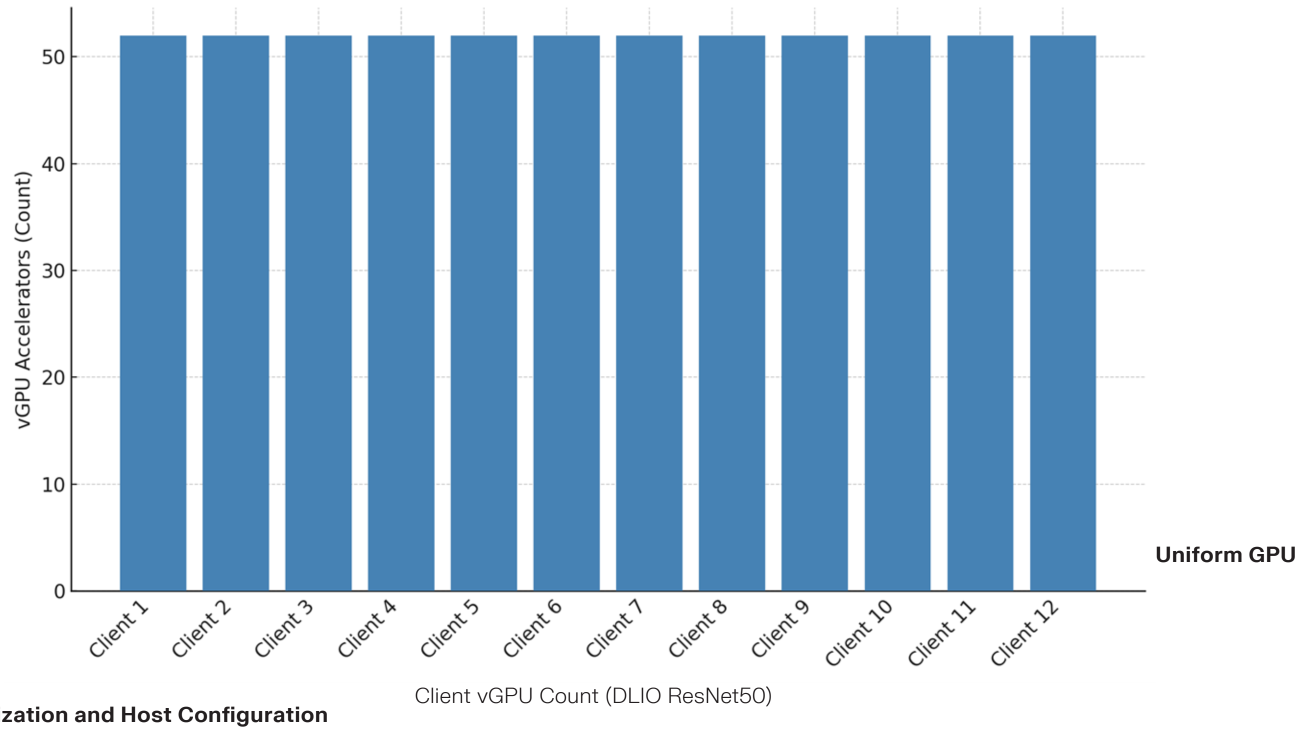
Performance Overview

Across the 12 client hosts, throughput values span ~9.1 GB/s to ~9.5 GB/s, showing only minor variation (≈4% deviation). This level of consistency reflects a highly stable storage fabric and balanced RDMA pathing, confirming that the Ingrasys ES2100 chassis delivered uniform I/O distribution to each host. The aggregate performance at roughly 112 GB/s total is in line with what can be expected considering the ResNet50 I/O characteristics. It should be noted that in preparatory Flexible I/O (fio) benchmarks, the ES2100 demonstrated 167.6 GB/s in 128k sequential read bandwidth and 161 GB/s in 128k sequential writes whilst using the Kioxia CM7 NVMe drives.

ResNet50 Workload Characteristics

ResNet50, as implemented in DLIO, exhibits high read concurrency with sustained sequential access patterns. Each GPU or host thread issues streaming reads on moderately sized data batches (typically 256 KB–1 MB blocks). The consistent throughput observed here implies the ES2100's aggregate bandwidth scaling handled concurrent deep-read pipelines effectively, with minimal interference from metadata lookups or file system locking.

Observations on vGPU Count



Virtualization and Host Configuration

Each of the twelve hosts demonstrated 52 vGPU accelerators, reflecting a perfectly balanced compute topology. This uniformity ensures that DLIO's ResNet50 workload (which scales linearly with GPU count) experienced identical computational demand and I/O generation per client. With 52 vGPUs per node, the test environment represented an aggregate of 624 concurrent GPU data streams, each demanding consistent low-latency access to the dataset. The ES2100 effectively sustained hundreds of thousands of small parallel reads per second, translating into linear data feed performance across all nodes.

Correlation to client AU% (Active Utilization Percentage) behavior

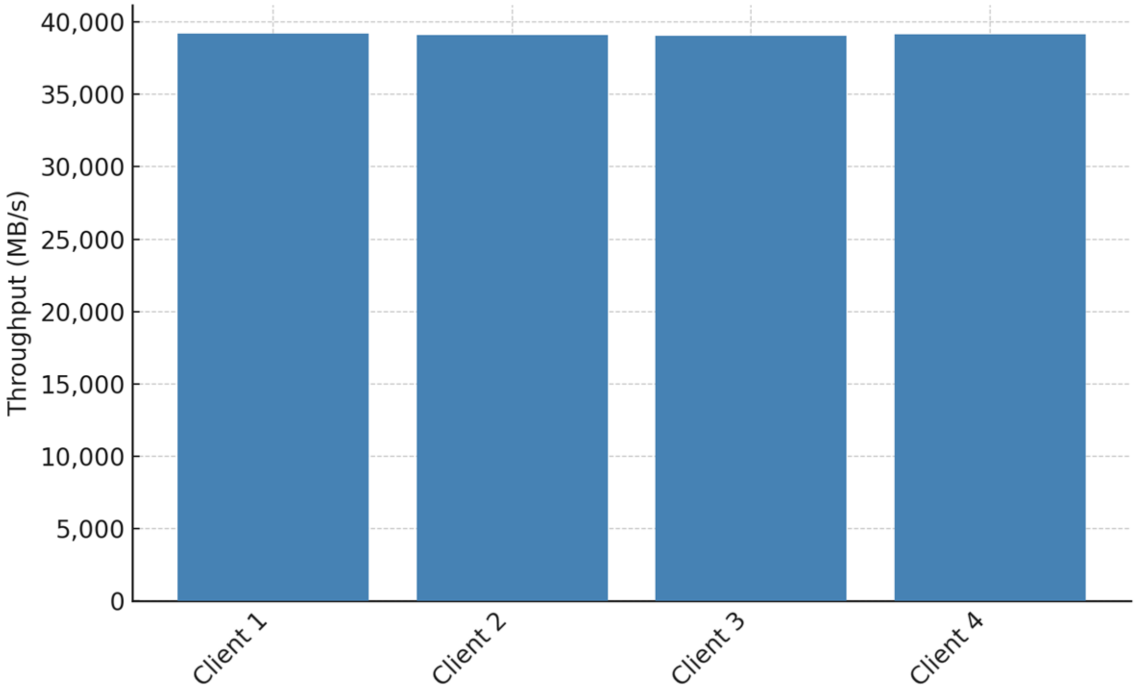
Although not shown in this specific chart, prior AU% results (90–94%) align closely with what's expected for well-saturated GPUs under I/O-balanced conditions. The lack of variation in vGPU count confirms that differences in AU% stemmed from I/O latency or minor RDMA path variations, not from unequal GPU provisioning. The ES2100's internal PCIe topology and RapidFlex™ interposers likely helped maintain consistent queue completions across all initiators.

Observations on vGPU Count: Summary

The vGPU distribution chart underscores the methodical symmetry of the test design—52 vGPUs per host providing a uniform, compute-saturated load to the ES2100. Combined with consistent AU% and throughput behavior, the results verify that the storage platform delivered predictable, high-efficiency data delivery across a fully parallelized DLIO ResNet50 workload.

Results Overview: 3D U-Net Benchmarking

Observations on client throughput



3D U-Net Client Throughput

Performance Overview

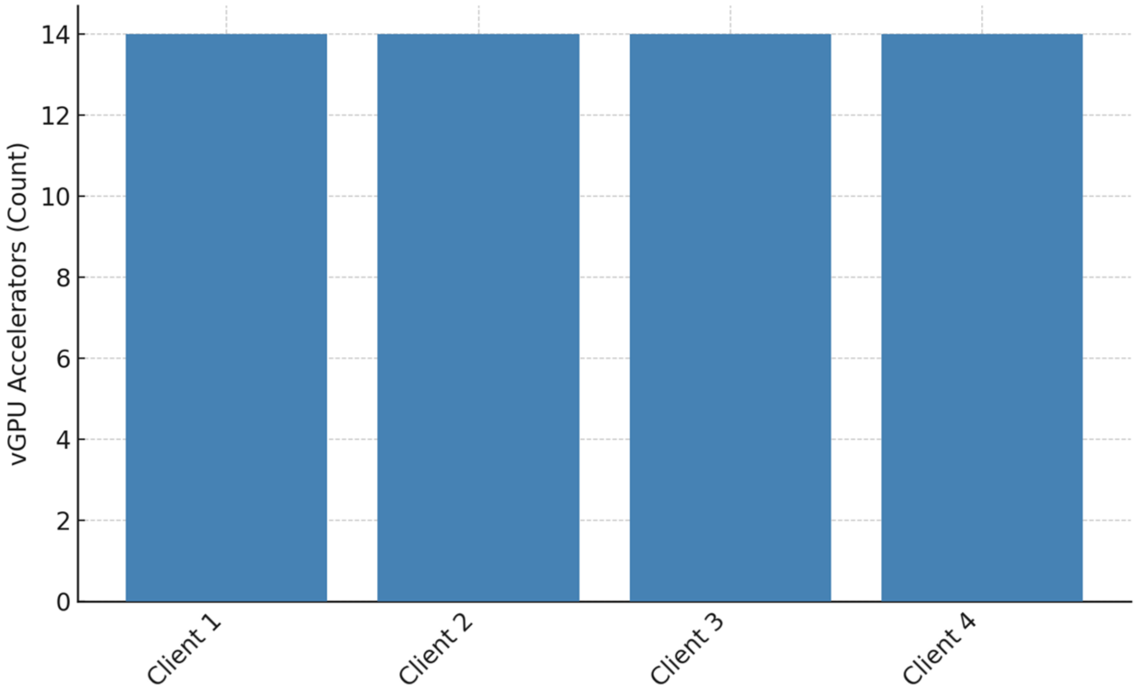
The four clients each achieved 38 GB/s, with values ranging narrowly between 39,070 MB/s and 39,208 MB/s—a deviation of less than 0.4 % across all nodes with a total aggregate chassis bandwidth of 152 GB/s. This uniformity demonstrates exceptional load balance and validates the deterministic I/O behavior of the ES2100 under a distributed DLIO workload. It indicates that both the fabric interconnect and storage subsystem were able to distribute the dataset evenly with negligible queue latency variance.

Relative to ResNet50 (which averaged around 9 GB/s per host), each 3D-UNet client sustained over 4× the per-client throughput. This difference reflects the larger I/O footprint and more aggressive data streaming pattern of 3D-UNet, where volumetric datasets (3D tensors) require deeper prefetch queues and higher sustained bandwidth.

The ES2100's ability to scale throughput proportionally across multiple initiators—without observable degradation—underscores its headroom for mixed-model AI workloads that blend convolutional and volumetric training pipelines.

In short, the ES2100 delivers enterprise-class linearity in a compact 2U footprint, proving its suitability as a shared NVMe-oF storage platform for multi-GPU and multi-node AI pipelines where 3D-UNet-like workloads dominate.

Observations on vGPU Count



3D U-Net vGPU Count Chart

Uniform vGPU Allocation Across Clients

Each of the four clients demonstrated 14 vGPU accelerators, resulting in a total of 56 vGPUs being available. This even allocation is essential for DLIO's 3D-UNet benchmark, ensuring that each node exerts an identical computational and I/O demand on the storage subsystem. The flat vGPU distribution seen in the chart confirms orchestration symmetry — no host was oversubscribed or under-allocated.

Throughput Consistency and Linearity

Per-client throughput ranged narrowly between 39,070 – 39,208 MB/s, yielding an aggregate of 156,515 MB/s (152.9 GB/s). This near-perfect alignment (less than 0.4 % variance) indicates that the ES2100 delivered highly deterministic data access, even under concurrent 3D tensor streaming loads. The balanced throughput distribution is a clear indicator that the PCIe topology, RapidFlex™ interposers, and NVMe-oF fabric are working in concert to eliminate arbitration bottlenecks and drive uniform latency across initiators.

The AU percentages (92–93 %) further validate that each vGPU was being fed data at an optimal rate. In deep learning workloads, utilization above 90 % is typically indicative of a storage-fed compute-saturated regime, meaning that the GPUs were waiting minimally for data.

Summary

The benchmarking of the Ingrasys ES2100 NVMe-oF EBOF with Western Digital RapidFlex A2000 interposers and Kioxia CM7-V SSDs using NVIDIA's open-source DLIO benchmark demonstrated exceptional I/O stability and scalability across AI workloads.

ResNet-50 achieved roughly 112 GB/s aggregate throughput with only ~4 % deviation, reflecting consistent low-latency delivery for metadata-intensive small-file access.

The 3D U-Net workload, emphasizing large sequential reads, reached 152–156 GB/s aggregate throughput with less than 0.4 % node-to-node variation and GPU utilization around 93 %. Uniform vGPU allocation across clients confirmed balanced compute and storage distribution. Overall, results validate the ES2100's systems predictable, repeatable, and consistent throughput and latency performance. The chassis demonstrates linear scalability, and suitability for multi-node AI/ML training where predictable throughput and minimal latency are critical.

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com