Western Digital.

Storage Benchmarking

MLCommons MLPerf Storage v2.0: Western Digital OpenFlex® Data24 4200 in Focus – Performance, Architecture, and the Necessity for True Comparison

October 2025

Table of Contents

Introduction

The MLPerf Storage benchmark from MLCommons has rapidly evolved into a de facto industry standard for assessing Machine Learning (ML) storage infrastructure performance. The newly released version 2.0 (MLCommons) extends realism in testing, adding workloads like checkpointing for large language models (LLMs) and scaling GPU simulation capabilities, making it a more demanding and relevant benchmark for today's Artificial Intelligence pipelines.

In both standalone testing and in partnership with PEAK:AIO and KIOXIA CM7-V NVMe™ SSDs, Western Digital's OpenFlex Data24 4200 produced strong results on the ResNet-50 and 3D-UNet workloads. PEAK:AIO is tailored to OpenFlex, creating a centrally managed resource that provides RAID options, data services, and cluster-wide volume management. The results confirm the platform's strength in combining high throughput with low latency while showcasing the efficiency of its disaggregated architecture.

Test Procedures – How MLPerf Storage Works

The MLPerf Storage benchmark evaluates infrastructure using synthetic but representative datasets designed to mirror the size and access patterns of real AI workloads. Up to 250 virtual GPU (vGPU) simulators generate concurrent I/O requests that aim to reflect how training jobs consume data in production.

These workloads are executed across multiple client nodes to assess both scaling efficiency and the ability of the storage fabric to sustain network saturation.

The suite evaluates several critical dimensions: sustained throughput in MB/s under load, how effectively vGPU can be driven to full I/O utilization, along with variation in response times (latency). Such factors that are particularly important in synchronous training workloads.

The latest version 2.0 benchmark suite also introduces checkpointing tests, which simulate the multi-terabyte write operations that occur during save points in large-scale LLM training. Additionally, organizations may self-report nameplate power (the maximum amount of electrical power a device or system is rated to draw) and rack unit data, which allows optional normalization of results for density and efficiency, though this remains inconsistently applied across submissions.

The Models in Focus

Residual Network-50

Residual Network-50 (ResNet-50) is a convolutional neural network (CNN) widely used for image classification tasks. Originally introduced by Microsoft Research in 2015 as part of the ResNet family, it pioneered the use of deep residual learning, enabling far deeper networks without vanishing gradient issues. In the MLPerf Storage benchmark, ResNet-50 generates a demanding I/O profile characterized by frequent, smaller reads of image data, with a mix of sequential and random-access patterns and moderate per-batch I/O sizes. This makes it particularly useful for measuring how well a storage system can handle high-frequency metadata lookups and small file reads—scenarios that stress both IOPS and latency as much as raw throughput. With version 2.0, the benchmark has been scaled to represent hundreds of GPUs accessing data simultaneously, creating a realistic model of large-scale AI training environments.

3 Dimension Neural Network

3 Dimension Neural network (3D-Unet) is a three-dimensional CNN designed for volumetric medical image segmentation. First introduced in 2016 by researchers at the University of Freiburg, it extended the popular U-Net architecture to 3D data, quickly becoming a standard for biomedical imaging tasks such as tumor detection and organ delineation. In the MLPerf Storage benchmark, it places heavy demands on storage by generating large, contiguous reads of 3D volume data. This I/O profile emphasizes high sustained bandwidth and sequential streaming, making it an effective test of a system's ability to deliver throughput under concurrent access. Unlike ResNet-50, which stresses small-file IOPS and latency, 3D-UNet is primarily a measure of raw bandwidth and the ability of the storage fabric to maintain performance at scale. This benchmark workload allows the highlighting of weaknesses in bandwidth delivery and parallel file striping, pushing systems to demonstrate consistent throughput under pressure.

OpenFlex Data24 4223 Results Summary

- Unet3D delivered: 101.6 GB/s with 36 vGPU across 3 clients—delivering sustained throughput near media limits for NVMe over RoCE.
- ResNet-50: Supported 186 vGPU across 3 nodes while maintaining over 33 GB/s—showing strong IOPS and metadata performance.
  Results Summary: https://mlcommons.org/benchmarks/storage/

The Challenge of Data Normalization

The MLPerf Storage benchmark's heavy reliance on 'clients' and 'vGPU counts' as normalization metrics can be misleading and difficult for practitioners to interpret.

Counting client servers assumes that more initiators equate to realistic production scaling, but in practice client systems vary widely in allocated CPU, NIC, DRAM and network bandwidth; factors that dramatically influence I/O behavior as well as user cost.

The client architecture used in a given submission have no limitations imposed on quantity and compute capability. This is significant in that the benchmark modelling (i.e. awarded vGPU) is heavily influenced by this capability.

There is no requirement for submissions to disclose associated capital expenditure, licensing fees, or ongoing operational overheads. This lack of transparency obscures not only the true up-front investment but also the longer-term costs of ownership—such as depreciation, maintenance, and support; leaving practitioners without a clear understanding of the real economic impact.

Beyond these normalization concerns, another limitation is the absence of meaningful power data. While nameplate power distribution is required within the submission, MLPerf reports do not provide actual average or peak power consumption. This omission leaves users without visibility into how solutions behave under sustained load versus burst conditions—an increasingly critical factor for data centers facing power and cooling constraints. The result is a benchmark framework that risks oversimplifying real-world trade-offs, favoring configurations tuned for scoring over those optimized for resilience, energy efficiency, operational sustainability, and affordability.

Essentially, results normalization is required. However, in the absence of the mentioned key metrics, meaningful normalization becomes extremely difficult. At present, the only consistent metric is client count, but this is a poor surrogate for understanding true system efficiency. A submission with dozens of lightly configured initiators may achieve high aggregate throughput yet do so with vastly different economics and infrastructure complexity compared to one with fewer, denser nodes.

Normalization itself, risks collapsing into a numbers game, rewarding configurations that scale client servers rather than those that deliver balanced, efficient performance in real deployments. A more complete framework would need to introduce transparent reporting of both economic and operational factors, so that normalization reflects not just performance scaling, but the true efficiency of the solution being benchmarked.

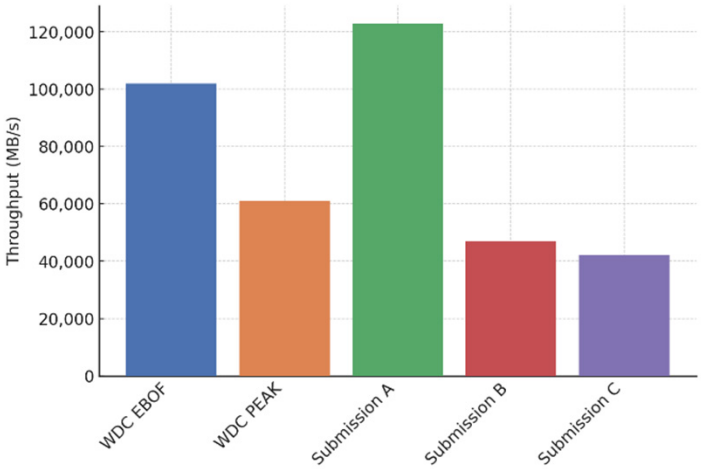
Results Detail

The remaining sections of this document look in more detail at the Western Digital submissions against comparative version 2.0 (Fabric-Attached Block) submissions. Those comparative organizations will remain anonymous out of professional courtesy, with the analysis focused solely on architectural and performance characteristics rather than vendor identity.

3D-UNet Model

Reported Total MB/s throughput per submission

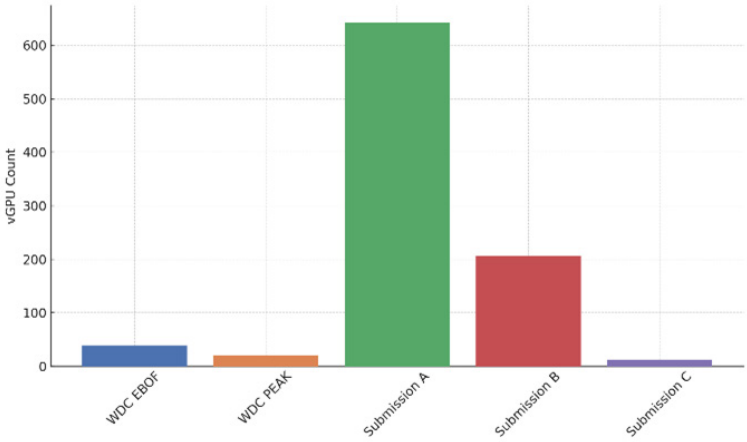
The chart illustrates throughput performance across five submissions. It is important to note that these results are not normalized. Submission A leads decisively, delivering just over 120,000 MB/s, the highest result in the group. The WDC EBOF follows with more than 100,000 MB/s, positioning it as the second-strongest performer and placing it well ahead of the remaining entries. WDC PEAK records around 60,000 MB/s, while Submission B and Submission C deliver approximately 47,000 MB/s and 42,000 MB/s respectively. While these figures highlight peak throughput levels, the absence of normalization means the results should be viewed as indicative rather than strictly comparative, since factors such as client configuration or scale can influence raw performance.



3D-UNet Reported Total vGPU Count per submission

This chart highlights vGPU counts across the same set of submissions, but as with the throughput results, these figures are not normalized. Submission A dominates with over 600 vGPU, which far exceeds the scale of any other entry and is the primary driver behind its high raw throughput in the earlier chart. Submission B shows a more moderate level with around 200 vGPU, while WDC EBOF, WDC PEAK, and Submission C all remain below 50 vGPU.

Taken together with the throughput chart, this demonstrates that raw performance comparisons are heavily influenced by the number of client resources, rather than efficiency alone. There is more to investigate. The WDC EBOF, for example, delivered over 100,000 MB/s throughput with fewer than 50 vGPU, which points to strong per-vGPU efficiency. By contrast, Submission A's leading throughput is tied directly to its very large vGPU count, raising questions about submission scale strategy and costing.

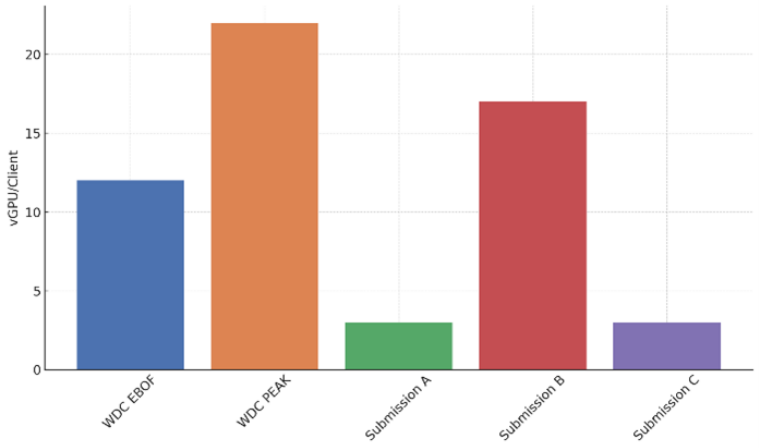


3D-UNet vGPU normalized per client

This chart provides a more meaningful perspective by normalizing vGPU counts per client, reducing the distortion seen in the raw vGPU totals. Unlike the previous slide, where Submission A appeared dominant due to its sheer scale of over 600 vGPU, here it drops to the lower end with fewer than 5 vGPU per client. This indicates that its throughput advantage is tied more to the sheer number of clients deployed rather than the efficiency of each node.

By contrast, the WDC PEAK and Submission B results stand out strongly, supporting over 22 and 17 vGPU per client respectively. This suggests they deliver higher density and potentially better resource utilization per server. The WDC EBOF performs at a solid mid-tier level with around 12 vGPU per client. Submission C, like Submission A, shows comparatively low per-client efficiency.

Normalization highlights that raw throughput alone does not tell the full story. Submissions that achieve high vGPU-per-client ratios demonstrate architectural efficiency and reduced infrastructure sprawl—factors that matter in real-world deployments when considering cost, power, and manageability.

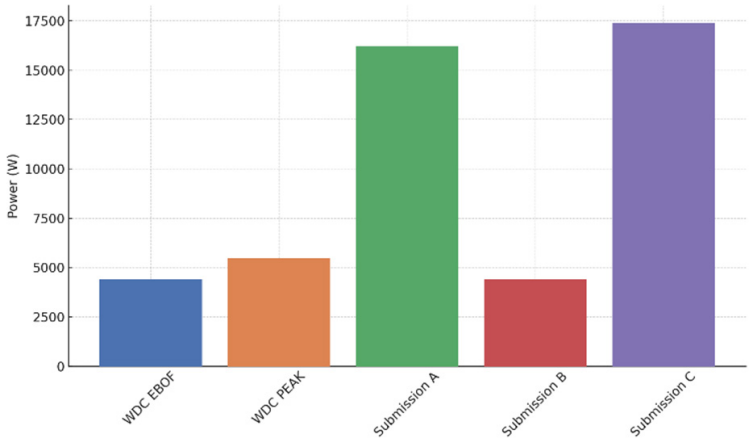


3D-UNet Reported Nameplate power consumption per submission

This chart illustrates reported nameplate power consumption for each submission; but since the values are not normalized, the picture is skewed by system scale rather than efficiency. Submissions A and C stand out with extremely high reported power requirements of ~16,000 W and ~17,500 W respectively, reflecting the large infrastructures deployed in those tests. In contrast, WDC EBOF, WDC PEAK, and Submission B report much lower consumption in the 4,000–6,000 W range, supporting their smaller hardware footprints.

Viewed alongside the earlier throughput and vGPU charts, this raises some considerations. At this stage (without further analysis) Submission A's strong throughput correlates with massive resource and power deployment rather than intrinsic efficiency, while Submission C demonstrates both low throughput and the highest reported power, implying poor efficiency overall. WDC EBOF and WDC PEAK show relatively modest power consumption while still achieving competitive performance, suggesting stronger performance-per-watt profiles once normalized.

Ultimately, nameplate power metrics are useful for sizing and provisioning but does not capture actual average or peak draw. Still, these raw figures reinforce the need for normalized metrics—such as throughput-per-watt or vGPU-per-watt—to meaningfully compare efficiency across architectures.

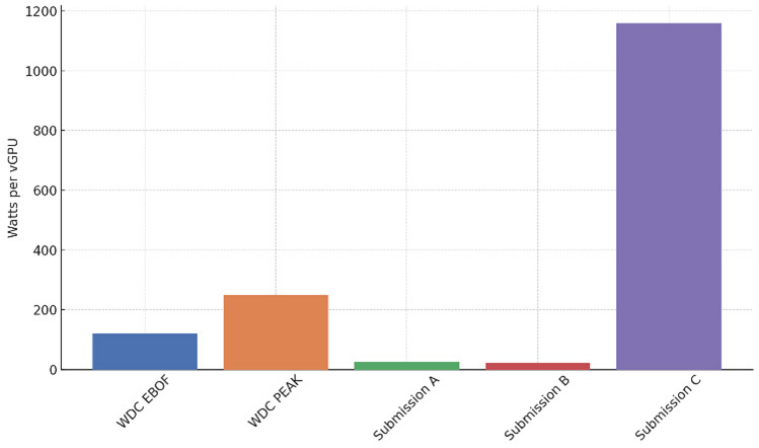


3D-UNet Normalized Efficiency (vGPU per Watt) per Submission

This normalized view demonstrates efficiency—showing how many vGPU each system supports per watt consumed. Submission A, which previously appeared power-hungry, now stands out as one of the most efficient designs, delivering roughly 0.04 vGPU per watt thanks to its large GPU count. Submission B also performs well at about*0.03 vGPU per watt, underscoring balanced efficiency despite more modest overall throughput.

In contrast, Submission C lags significantly, posting only about 0.001 vGPU per watt, by far the least efficient of the group. WDC EBOF lands in a favorable mid-tier position, offering around 0.008 vGPU per watt, which reflects a solid mix of throughput and efficiency. WDC PEAK comes in at about 0.004 vGPU per watt, reflecting its denser architecture and the impact of running a PEAK server node alongside a client.

Overall, this chart highlights how normalization changes the story: systems with high raw power figures (like Submission A) may actually be very efficient per resource, while others with smaller deployments can struggle when evaluated at the vGPU level. WDC's EBOF strikes an appealing balance of efficiency and practicality, while PEAK illustrates the trade-offs of SDS -enabled flexibility.



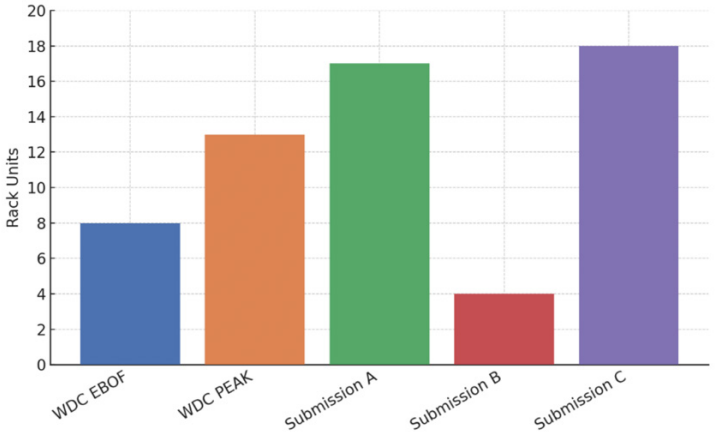
3D-UNet Reported Rack Units Consumed per Submission

This chart shows reported Rack Units (RU) consumed per submission, again in unnormalized form. Submission C stands out as the largest footprint at 18 RU, closely followed by Submission A at 17.

WDC PEAK sits in the mid-range at 13 RU, while WDC EBOF comes in lower at 8 RU. Submission B reports the smallest footprint, requiring only 4.

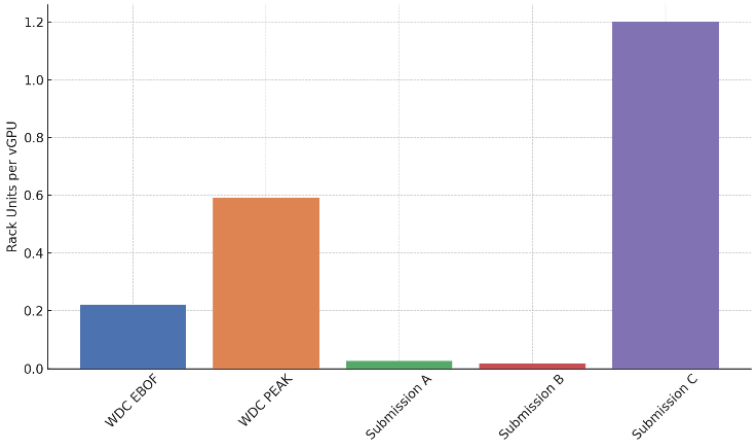
While these raw values are useful for understanding the physical size of each configuration, they do not by themselves reveal the efficiency of the solution beyond RU requirements. Larger deployments, such as Submissions A and C, achieve scale through sheer physical density but at the expense of space, power, and cost. By contrast, the WDC EBOF and Submission B demonstrate more compact footprints, which could translate into better rack efficiency once normalized against throughput or vGPU counts.

As with power consumption, additional normalization could be explored here—for example, metrics such as throughput per rack unit or vGPU per RU would better capture which systems deliver the highest performance density, rather than just highlighting system size.



3D-UNet Normalized vGPU per Rack Unit

This chart highlights efficiency in terms of vGPU density per rack unit. Submissions A and B stand out with the highest densities,—demonstrating extremely compact scaling. WDC EBOF follows at 5 vGPUs per rack unit, offering a balanced mix of space efficiency and system simplicity. WDC PEAK is less dense, with just under 2 vGPUs per rack unit, reflecting the impact of its SDS-based architecture and the use of a separate PEAK node alongside clients. Submission C performs the weakest on this metric, supporting fewer than 1 vGPU per rack unit, indicating a space-intensive configuration.



3D U-Net Model Results Summary

When taken together, the charts reveal how raw throughput figures alone provide an incomplete picture of efficiency and architectural merit. Submission A, for example, delivered the highest throughput of the group at just over 120,000 MB/s, but this result is closely tied to its massive deployment of over 600 vGPU. Once normalized on a per-client or per-GPU basis, its advantage largely disappears, showing that its performance comes primarily from scale rather than per-resource efficiency.

By contrast, the WDC EBOF demonstrated a strong balance of performance and efficiency. Despite operating with fewer than 50 vGPU, it achieved more than 100,000 MB/s, positioning it as one of the most throughput-dense configurations. Its normalized metrics show solid mid-range efficiency in both vGPU-per-client and rack-unit-per-vGPU terms, and its modest nameplate power draw suggests better overall performance-per-watt than the large-scale systems.

WDC PEAK showed a different profile: lower absolute throughput than the WDC EBOF submission, but excellent density, sustaining more than 20 vGPU per client. However, its normalized power and rack metrics placed it behind WDC EBOF in efficiency; suggesting higher per-resource consumption even though it scales vGPU effectively. Submission B fell into a similar category: moderate throughput and relatively small rack footprint in raw terms, but once normalized, it revealed commendable efficiency both in power and rack usage per vGPU.

Submission C performed the weakest across the board. Its throughput was the lowest of all entries, yet its reported power and rack consumption were the highest in absolute terms. When normalized, it fared even worse—consuming more than one rack unit per vGPU and over 1,000 watts per vGPU—making it clearly the least efficient solution.

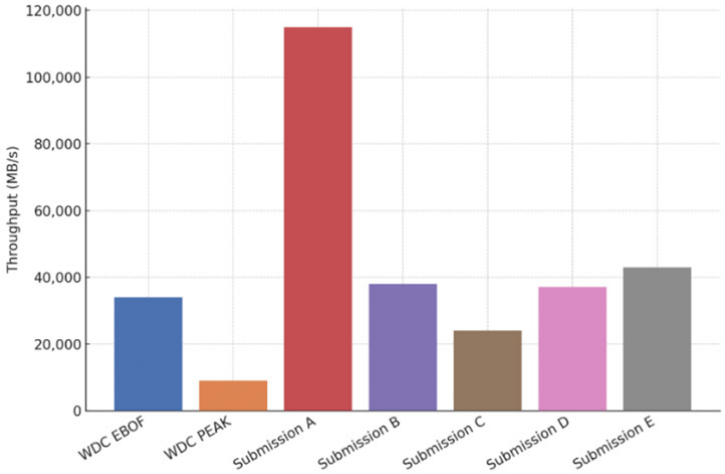
Once normalized, Western Digital submissions consistently land in the upper tier for efficiency and sustainability, avoiding the brute-force scaling used by some competitors. The highest raw performer (Submission A) is shown to rely on brute-force scale, while the WDC EBOF demonstrates strong efficiency with far fewer resources. The WDC PEAK and Submission B deliver competitive density while carrying higher per-resource costs. Submission C illustrates the risks of deploying architectures that fail to balance throughput with power and space efficiency.

Results ResNet-50 Model

ResNet-50 Reported Total MB/s Throughput per submission

The throughput results show a wide distribution across submissions. Submission A clearly dominates, achieving nearly 115,000 MB/s, well above all others and forming its own upper tier. The next group includes Submission E (43,000 MB/s), Submission B (37,000 MB/s), and Submission D (37,000 MB/s), each delivering broadly similar levels of throughput. The WDC EBOF sits just below this mid-tier cluster at approximately 33,000 MB/s, while Submission C trails at 24,000 MB/s. WDC PEAK records the lowest throughput overall, coming in at under 10,000 MB/s.

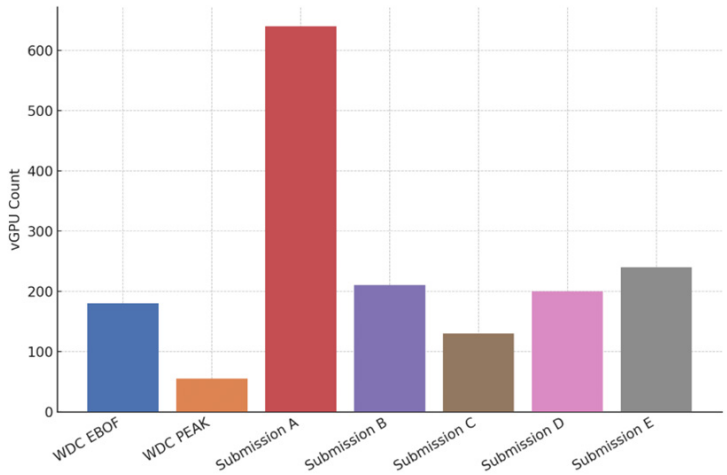
As with the earlier 3D-UNet analysis, configurations with larger client counts or higher vGPU allocations tend to inflate raw throughput figures. For that reason, direct comparisons of absolute throughput alone can be misleading. The more meaningful insights will emerge when these results are normalized against client counts, vGPU density, power consumption, and rack footprint—metrics that better reflect real-world efficiency and sustainability.



ResNet-50 Reported Total MB/s Throughput per submission

This chart reveals the scale disparities that drive much of the raw throughput performance. Submission A dominates with 640 vGPU, dwarfing all other systems and explaining its observed throughput in earlier charts. A second tier of results follow with Submission E (240 vGPU), Submission B (~208 vGPU), and Submission D (200 vGPU). WDC EBOF sits slightly below this range at 186 vGPU, reflecting a leaner but still competitive allocation. Submission C reports a smaller footprint of 128 vGPU, while Wesy PEAK is the most compact entry at around 52 vGPU.

Submission A achieves scale by brute force, while WDC's entries demonstrate more balanced configurations that pair moderate vGPU allocations with efficiency in power and rack utilization.

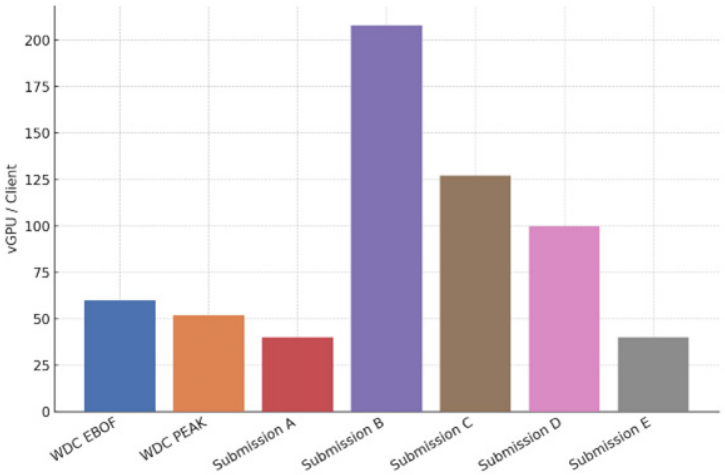


ResNet-50 vGPU normalized per client

This chart highlights efficiency in terms of how many vGPU each client server can support. Submission B is the standout, sustaining 208 vGPU per client, a dense configuration that reflects strong per-node utilization. Submission C also performs well at 128 vGPU per client, with Submission D following at 100 vGPU per client. It is worth noting that the same vendor offered Submissions C and D. The difference between the entries is the client count (1 and 2 respectively). What is observed is a nonlinear increase in vGPU count which raises questions about true performance scalability.

In the mid-range, WDC EBOF (62 vGPU per client) and WDC PEAK (52 vGPU per client) demonstrate balanced efficiency, consistent with their compact system designs. Submission A trails with 40 vGPU per client, while Submission E comes in lowest with 40 vGPU per client, underscoring limited per-node scaling.

The contrast between Submission A's dominance in total vGPU count and its relatively weak per-client density suggests a strategy that prioritizes brute-force scaling rather than per-node efficiency. By comparison, Submissions B, C, and D stand out as the most efficient in client utilization, an approach that could translate into lower infrastructure overhead, simpler scaling, and improved cost-effectiveness in practical deployments.

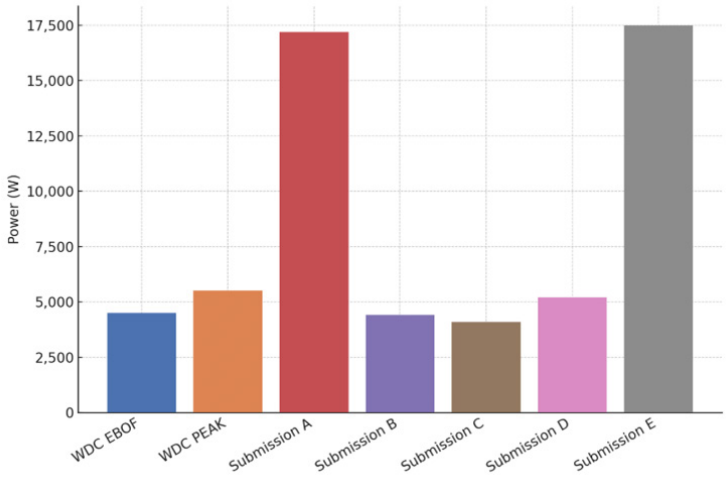


ResNet-50 Nameplate power consumption per submission

Submissions A and E are the most power-hungry, each drawing 17,120 Watts (W) and 17,400 W respectively which help underscore the scale and resource intensity of these deployments. WDC PEAK (5,500 W) and WDC EBOF (4,400 W) sit in the lower range, reflecting their more compact designs. The mid-tier includes Submission B (4,500 W), Submission C (4,200 W, and Submission D (5,200 W), all of which consume considerably less than the largest-scale entries.

When normalized against vGPU counts, the picture changes. WDC EBOF proves among the most efficient systems, delivering solid throughput while maintaining one of the lowest power-per-vGPU figures.

Overall, the data suggests that while Submissions A and E achieve top-line throughput through sheer scale, they do so at a steep power cost. By contrast, the WDC entries demonstrate leaner, more power-conscious designs, striking a balance between performance and sustainability—an increasingly important metric for data centers operating under power and cooling constraints.

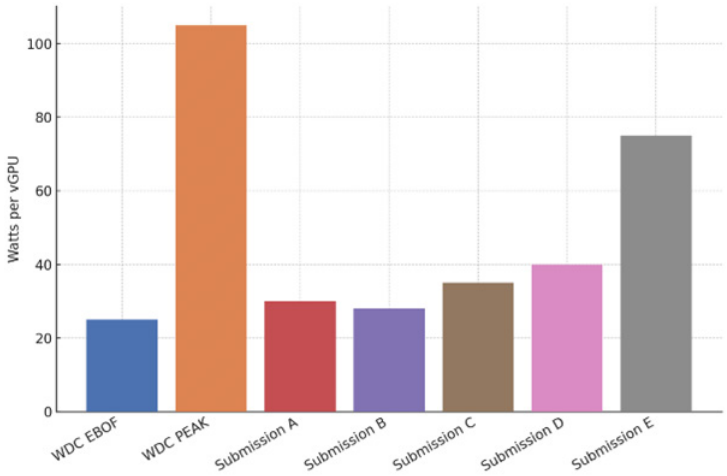


ResNet-50 Normalized Nameplate Power Efficiency (vGPU per Watt)

This chart shows how many vGPU each system can support per watt of power. On this scale, higher values indicate stronger efficiency. WDC EBOF emerges as one of the top performers, delivering the highest vGPU per watt ratio and reinforcing its profile as a balanced, power-conscious architecture. Submissions A, B, C, and D also cluster in the efficient range, each sustaining proportionate GPU allocation with relatively modest power draw.

In contrast, Submission E and WDC PEAK rank lowest on this metric, supporting fewer vGPU per watt than the rest of the field. For PEAK, this result is not purely a reflection of its Software Defined Storage (SDS) layer. The test setup included both a dedicated PEAK server node and a separate client system, which inflates the apparent per-vGPU power cost relative to configurations where these roles are integrated. While SDS may introduce some additional overhead, it does not inherently impose a significant penalty on power efficiency.

Overall, this chart highlights two complementary stories: WDC EBOF demonstrates the advantages of a fabric-attached block approach for maximizing efficiency, while WDC PEAK illustrates how SDS architectures can deliver flexibility and density at the cost of some normalized efficiency. Together, they underscore that architectural design choices—not just raw throughput—play a pivotal role in shaping the real-world efficiency of ML storage infrastructures.

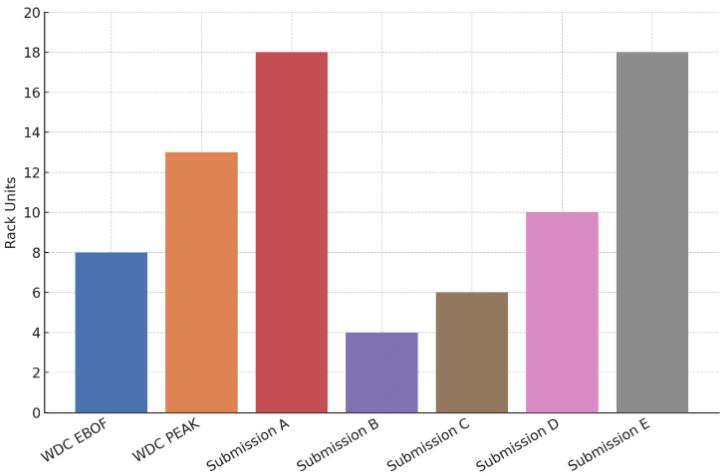


ResNet-50 Rack Units Consumed per Submission

This chart shows the physical footprint of each system in rack units. Submissions A and E occupy the largest footprints at 18 rack units each, indicating substantial infrastructure requirements. WDC PEAK and Submission C fall into the mid-high tier at around 13 and 6 rack units respectively, while Submission D sits at 10 rack units.

On the smaller end, the WDC EBOF requires just 8 rack units, offering a compact deployment relative to most of the field. Submission B is even leaner at 4 rack units, making it the most space-efficient of the group.

While these raw rack figures provide useful context for deployment planning, they do not reveal overall efficiency in isolation. To fully understand deployment efficiency, these values should be considered alongside normalized metrics such as throughput-per-rack-unit and vGPU-per-rack-unit.



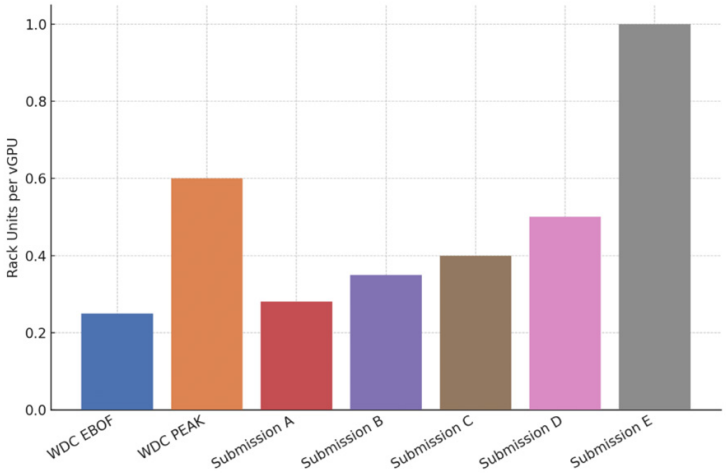
ResNet-50 Normalized Rack Efficiency (vGPU per RU)

This chart normalizes vGPU allocation by rack space, providing a clearer view of how effectively each system uses physical footprint. Submission E is the least efficient, supporting only about 1 vGPU per RU, underscoring its reliance on raw scale rather than density. WDC PEAK also trends toward the lower end at roughly 1.7 vGPU per RU, reflecting its test configuration that included both a PEAK server node and a separate client, which naturally reduces density.

In contrast, most other systems fall into a more efficient band of 2–3 vGPU per RU. Submissions A, B, C, and D demonstrate moderate rack efficiency, with Submission A improving its relative standing when normalized versus its raw rack footprint.

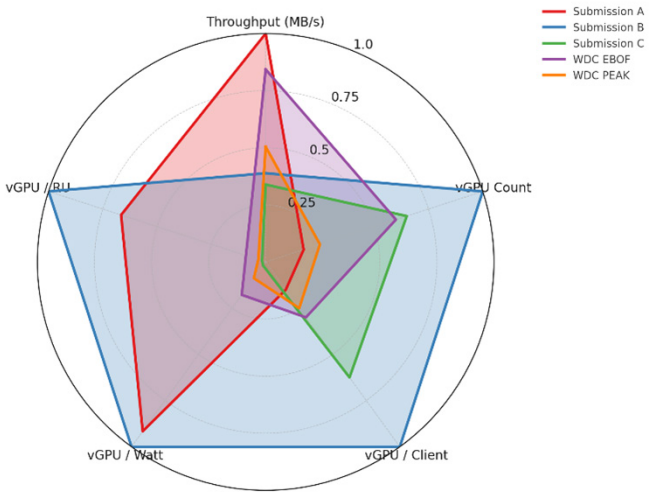
At the top end, WDC EBOF clearly stands out at more than 4 vGPU per RU, positioning it as the most space-efficient solution of the group.

The comparison highlights an important distinction: raw rack counts exaggerate the size of scale-heavy systems, but normalization reveals which architectures make the most effective use of physical space. Here, WDC EBOF demonstrates the strongest balance of density and efficiency, while systems like Submission E show how scale can inflate footprint without proportional returns.

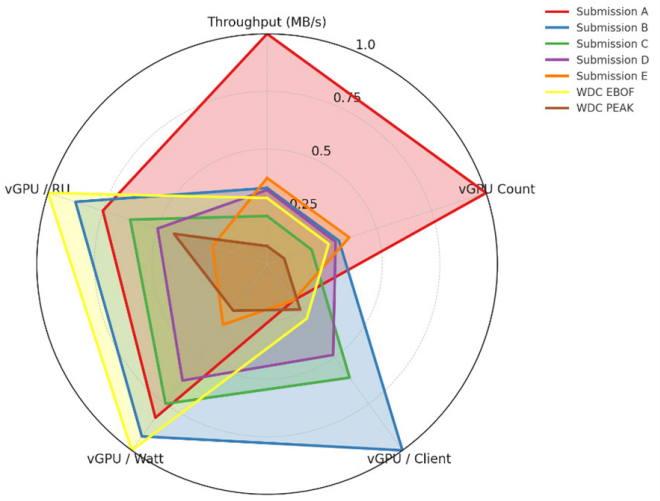


Closing Thoughts

ResNet-50 Radar Chat



3D-Unet Radar Chart



The MLPerf results reinforce both the value and the limitations of the current benchmark framework. Absolute throughput continues to reward scale-heavy deployments, but once normalized, a much more nuanced story emerges around efficiency, density, and sustainability. The above radar charts underscore this point visually. Scale-oriented systems sprawl outward on raw throughput and GPU count, while Western Digital's submissions achieve a more balanced and rounded profile across all five efficiency axes.

The WDC EBOF stands out in particular, highlighting the strengths of the OpenFlex Data24 4200 architecture. Despite deploying fewer vGPU and clients than some scale-out competitors, it delivered over 100 GB/s in 3D-UNet and supported 186 vGPU in ResNet-50, while maintaining modest power draw and compact rack usage. In the radar charts, this translates into a well-rounded polygon that extends strongly across the vGPU-per-Watt and vGPU-per-RU axes, confirming its role as one of the most efficient and space-conscious designs in the set.

The WDC PEAK submission, though delivering lower raw throughput, validated a different strength: exceptional vGPU density per client. This capability is clearly reflected in the vGPU-per-Client axis of the radar charts, where PEAK maintains a competitive position. This vGPU density per client illustrates how PEAK:AIO can maximize GPU utilization even with limited client hardware, which is a critical advantage in environments where compute slots are scarce or licensing costs are high. At the same time, its relatively higher watts-per-vGPU shows up as a contraction on the power efficiency axis, illustrating the impact of introducing a server / client layer compared to the direct efficiency of fabric-attached block designs such as EBOF.

By contrast, the anonymous competitive submissions often achieve headline throughput figures by deploying very high client numbers or consuming disproportionately large amounts of power and rack space. The radar charts make these trade-offs explicit: polygons spike on throughput and vGPU count, but collapse inward on efficiency axes, showing how raw scale can mask significant resource costs.

In sum, Western Digital's OpenFlex Data24 4200-based entries demonstrate that strong benchmark results do not require brute-force scaling or inflated resource consumption. Instead, they represent a credible, real-world approach to AI storage infrastructure: competitive in raw output, efficient under normalization, and aligned with modern data center priorities of power, space, and cost discipline. The radar visualizations reinforce this story—showing EBOF as the most balanced and sustainable design, with PEAK adding density flexibility—and they highlight the ongoing need for greater architectural and economic transparency in MLPerf reporting to enable more equitable and actionable comparisons. Together, the EBOF and PEAK submissions underline Western Digital's ability to address both efficiency-driven deployments and software-defined flexibility, positioning OpenFlex as one of the most adaptable storage solutions in AI today

Appendix

OpenFlex Data24 4000 Series Overview

The OpenFlex Data24 4200 is a high-performance, NVMe-over-Fabrics (NVMe-oF) storage JBOF designed to extend the full potential of NVMe SSD performance across Ethernet fabrics. Architecturally, it leverages a PCIe Gen4 backplane populated with dual-port NVMe SSDs, fronted by Western Digital RapidFlex NVMe™-oF Controllers, to deliver low-latency, high-bandwidth access to storage resources over the network.

The chassis supports industry-standard dual ported NVMe drives from multiple vendors, giving operators flexibility in capacity, endurance, and performance tiers. As part of Western Digital's "open composable" strategy, the Data24 4200 separates compute and storage into independent resource pools, enabling organizations to dynamically compose and recompose infrastructure based on workload requirements. This disaggregated architecture directly addresses several persistent pain points in AI/ML, HPC, and enterprise data environments:

- Stranded resources — By decoupling storage from compute, unused capacity in one node can be reallocated elsewhere without physical intervention.
- Scaling inefficiency — Traditional direct-attached storage (DAS) forces storage growth to be tied to server growth; the Data24 allows each to scale independently.
- Utilization bottlenecks — NVMe SSD performance can be shared across multiple hosts, improving overall asset utilization.
- Operational complexity — Centralized storage pools reduce the management overhead of per-node storage provisioning.
- Lifecycle mismatch — Compute and storage refresh cycles often differ; disaggregation allows independent upgrade paths, reducing TCO. In essence, the OpenFlex Data24 4200 transforms NVMe storage from a server-bound resource into a composable fabric asset, providing the flexibility, performance, and efficiency demanded by modern, data-intensive workloads.

Benchmark Architecture – Block to Data24 4200

Client Server

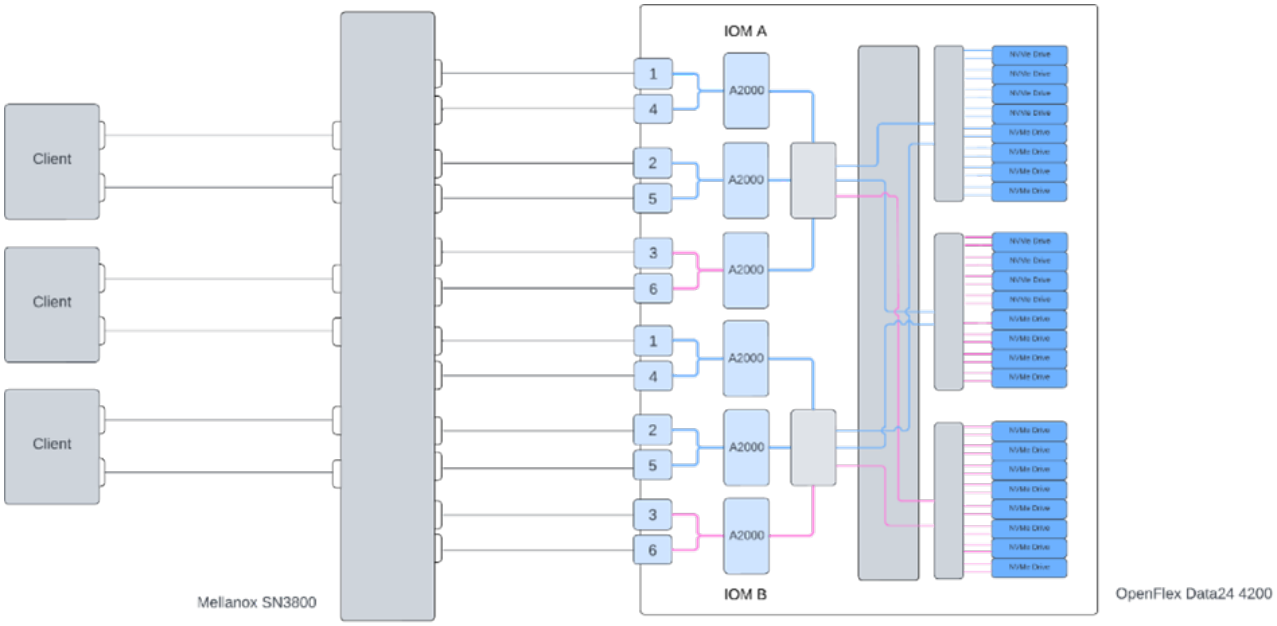
- 3 x SuperMicro H12DSU-iN
  - CPU: 2 x AMD EPYC 7452
  - Memory: 512 GB DDR5
  - Data Network: 1 x NVIDIA ConnectX-6
  - OS: Ubuntu 22.04

Disaggregated Storage

- 1 x Western Digital OpenFlex Data24 4223
- 24 x Kioxia CM7-V NVMe™ SSD – 6.4TB

Network

- Mellanox SN3800 64 Port 100GbE QSFP28 with ONYX OS



Benchmark Architecture – PEAK:AIO to Data24 4200

Client Server

Note: The PEAK:AIO submission ran on a single MLPerf client. This highlights that PEAK:AIO achieved competitive efficiency even at smaller scale

- 1 x SuperMicro 521-GE-TNRT
  - CPU: 2 x Intel Xeon Platinum 8462Y
  - Memory: 1024 GB DDR5
  - Data Network: 2 x NVIDIA ConnectX-7
  - OS: Ubuntu 22.04

PEAK:AIO Server

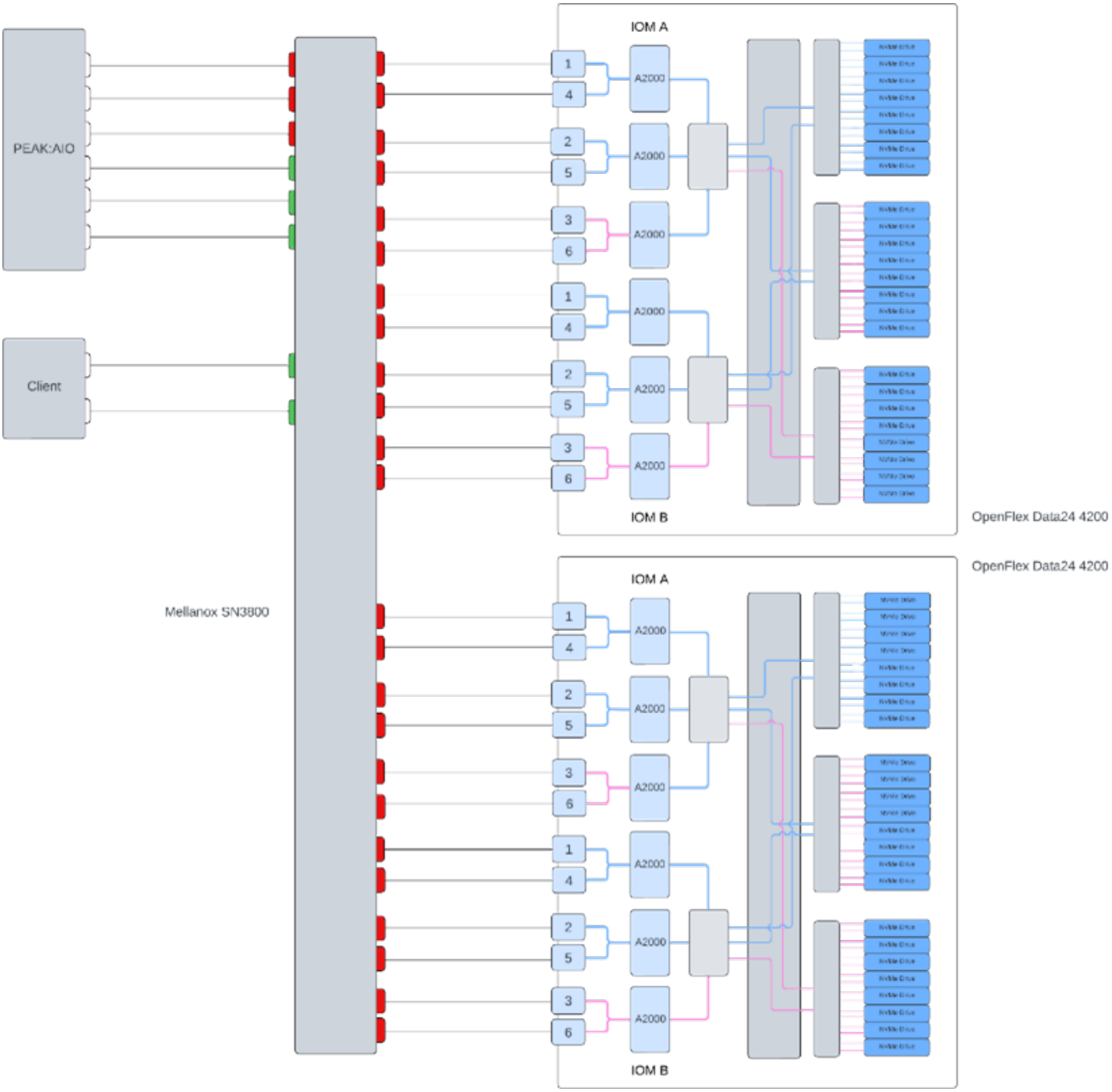
- ASUS WS Motherboard
  - CPU: 1 x Intel W9-3475X
  - Memory: 512GB DDR5
  - Data Network: 3 x NVIDIA ConnectX-7
  - Storage Network: 3 x NVIDIA ConnectX-7
  - OS: PEAK:AIO Linux 24.04

Disaggregated Storage

- 2 x Western Digital OpenFlex Data24 4200
- 48 x SanDisk SN655 15.36TB NVMe™ SSD

Network

- Mellanox SN3800 64 Port 100GbE QSFP28 with ONYX OS



- Green Switch Ports – Data Network
- Red Switch Ports – Storage Network

PEAK:AIO

PEAK:AIO is a high-performance software-defined storage (SDS) platform designed for AI, HPC, and data-intensive workloads. It delivers shared NVMe-oF storage optimized for GPU-centric environments, enabling low-latency access and high throughput across training, inference, and analytics pipelines. By separating storage from compute, PEAK:AIO helps organizations scale efficiently, supporting GPU-direct storage (GDS), RDMA networking, and composable architectures. Its SDS layer provides enterprise-class features—data services, resiliency, and management—while minimizing overhead, allowing integrators to pair it with industry-standard hardware like Western Digital's OpenFlex Data24. The result is a flexible, cost-effective solution that maximizes GPU utilization and accelerates AI workflows.

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

October 2025 / Revision 03