© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

OpenFlex Data24 for AI Infrastructure Storage

Disaggregated NVMe-oF Storage for KV Cache Offloading and Vector Database Workloads

Technical Overview

April 2026

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Topics In This Deck

Two validated AI infrastructure use cases on a single storage platform

- Section 00 : WD OpenFlex Data24 Overview
- Section 01: KV Cache Offloading — Disaggregated NVMe-oF KV cache tier for LLM inference.
- Section 02: Vector Database Storage — DiskANN on NVMe-oF over RoCE v2
- Section 03: Platform Convergence — Common architecture principles & combined takeaways.
- Section 04: OCCL — Open Composable Compatibility Lab.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

00

WD OpenFlex Data24 Overview

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

OpenFlex Data24 – 4100 & 4200

Overview

High Level Specifications

- 2U, 24-Bay, NVMe-oF JBOF
- Broad range of SSD Options
  - Ultrastar DC SN655 (CB II): 3840 / 7680 / 15360 / 30,720 / 61,440 GB
  - SE, ISE and TCG
  - 3rd party SSDs
- High Availability with dual IOMs, Titanium Eff. PSUs, dual-port SSDs
- N+2 fan redundancy
- Twelve 100GbE ports enabled by RapidFlex A2000
- Short chassis depth (< 28”)
- Product Availability: 4200 GA, 4100 expected to sample and GA June 2025

Key Capabilities

- Bandwidth matched performance from network port to SSD
  - PCIe Gen4 from front-end to back-end
- Device Sharing (NTB) to provide connections from any RapidFlex port to any SSD
  - Saves on 100GbE switch ports for direct connect configurations
- Single-port (4100) and dual-port (4200) SSD support
  - Single port optimal for Cloud (CSP) deployments
- Simple unified management through Resource Manager
- Industry-leading 5-year warranty

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Data24 Series Differentiators

High Availability (HA) vs. Standard Availability (non-HA)

4200 Series (High Availability)

- Intended for use with dual-port SSDs
- Redundant path to every SSD (2x2)
- Redundant PSU
- Optimal performance requires active connections to both SSD ports
- Targeted at HPC and other environments requiring redundancy all the way to the SSD

4100 Series (Non-HA)

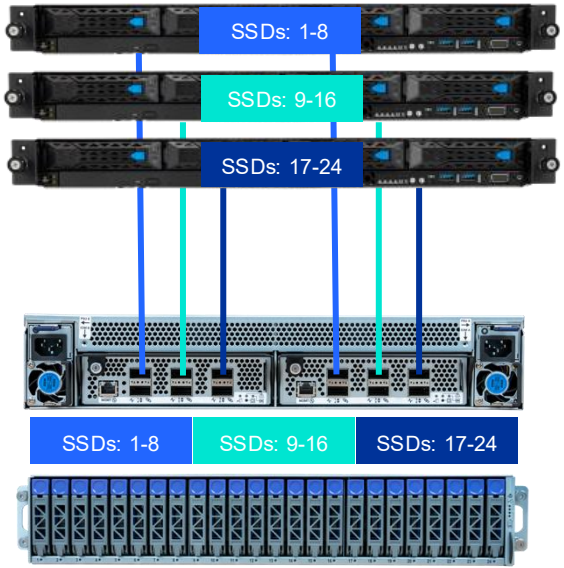
- Intended for use with single-port SSDs
- Single path to every SSD (1x4)
- Redundant PSU
- Performance optimized to a single SSD via a single connection
- Targeted at environments where redundancy is provided by mirroring the storage system (or is not required)

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

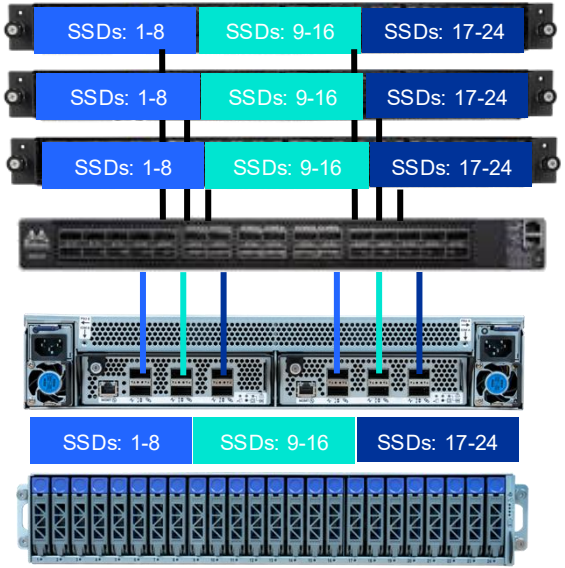
4200 Series Connectivity Options

Device Sharing Enables Any Port to Any SSD Connectivity Without A Switch

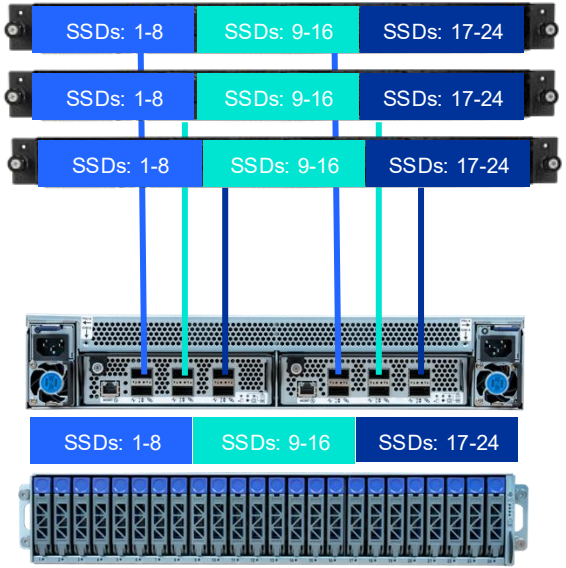
Direct Connect, No Device Sharing



Switched, no Device Sharing



Direct Connect with Device Sharing



© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

01

KV Cache Offloading for LLM Inference

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

What is Key Value Cache

KV Cache (Key–Value Cache) is a memory optimization used in transformer-based large language models to accelerate autoregressive inference. For each token generated, K(key) and V(value) vectors are created and stored in KV Cache within the GPU's High Bandwidth Memory (HBM). Being able to access these previously computed tensors helps avoid redundant computation and dramatically reduces latency. There is a memory tradeoff to take into account. KV cache consumes significant GPU memory (dominating HBM usage) which can cause production LLM inference at scale to run straight into the KV cache memory wall.

Serving long-context, multi-turn conversations with a hundred-billion-parameter class models can leave only a thin margin of GPU VRAM for active KV state before concurrency hits a hard ceiling. Operators must choose between buying more GPU nodes, truncating context, or finding a lower-cost overflow tier that does not destroy user-perceived latency.

The Western Digital OpenFlex Data24 addresses that challenge directly.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

The KV Cache Memory Wall

GPU VRAM is the most expensive memory in the data center

- KV cache stores previously computed key/value tensors in GPU HBM. As context lengths and concurrency grow, KV cache dominates VRAM usage.
- After loading a 70B parameter model in FP16 (~140 GiB), a 4x H100 node retains ~180 GB for KV cache. At 32K context, this exhausts within a handful of concurrent conversations.
- Operators must choose: buy more GPU nodes ($200K–$500K+), truncate context, or add local NVMe per server (couples storage to compute).
- The Data24 provides a fourth option: shared, disaggregated NVMe-oF KV cache that preserves performance while decoupling storage from compute.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

The 4:1 Architecture: One Drive, Four GPUs

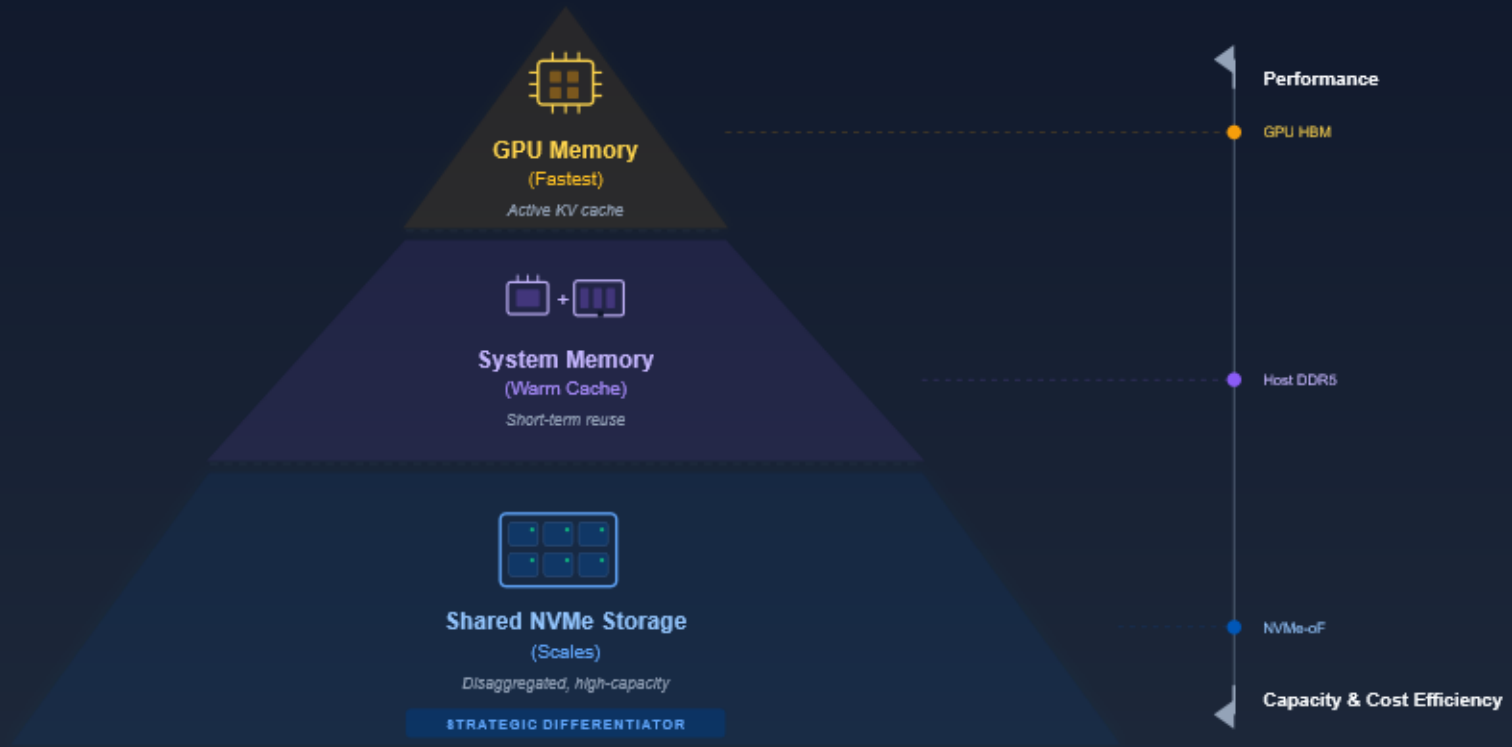
Scale-out at validated concurrency of 16 simultaneous multi-turn conversations per 4-GPU node

Data24 NVMe Drives	GPU Inference Nodes	Total H100-class GPUs	Max Concurrent Sessions
1	1	4	16
6	6	24	96
12	12	48	192
24 (full enclosure)	24	96	384

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Tiered KV Cache Architecture

Hierarchical memory tiers for scalable LLM inference

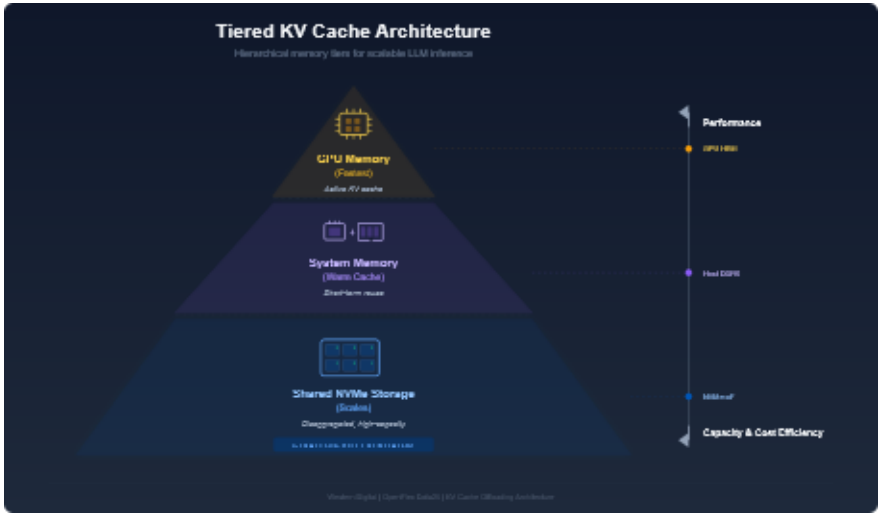


© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Tiered Cache Architecture

LMCache interposes a tiered hierarchy between GPU VRAM and persistent storage

- Tier 0 — GPU VRAM: 4x H100 80 GB HBM3/HBM3e.
- Tier 1 — Host DRAM: 30 GB warm cache per node (TP=4 configuration).
- Tier 2 — NVMe namespace on Data24: 100 GB budget, ext4 file system, NVMe/RDMA transport.
- Network: 4x 200 GbE Ethernet links from server to switch, 12x 100 GbE Ethernet to Data24
- KV access dominated by 128 KB (256 token) sequential reads/writes. Fabric carries traffic without congestion.



© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benchmark Configuration

70 Billion Parameter Model | TP=4 | 4x H100 | 32K context | 4-turn conversations

- Inference: vLLM + LMCache, GPU_MEM_UTIL=0.90, max-model-len 32768.
- Workload: Temperature 0.7, 512 completion tokens/turn, 4 turns/conversation.
- Concurrency: 4 and 8 simultaneous multi-turn conversations.
- Scenarios: Baseline (pure in-VRAM), Local SSD, NVMe-oF (Data24). Same LMCache config for SSD/NVMe-oF.
- Isolation: Fresh vLLM processes and cleared cache directories per scenario.
- Baseline: In the scenario matrix, the Baseline runs with no LMCache, no prefix caching, and no disk or host DRAM tiers active. All KV state stays resident in GPU VRAM across all four H100s (TP=4). The LMCACHE_CONFIG_FILE is explicitly unset in the baseline script, so there's no tiered storage path at all.

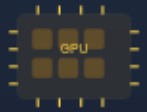
© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Test Scenarios at a Glance

Identical compute, identical workload — only the storage tier changes

In-GPU Memory

BASELINE / CONTROL



KV Cache In HBM

No external storage

All KV cache retained in GPU memory

=

Server-Attached NVMe

LOCAL STORAGE

KV cache spills to server-local SSD

=

Shared NVMe (Data24)

DISAGGREGATED

KV cache offloaded to shared storage over fabric

Same model, same workload, same software — only the storage tier changes.

- ✓ Production-scale model
- ✓ Multi-turn inference
- ✓ Concurrency tested

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Key Concepts for TPS and TTFT

Definitions

Tokens Per Second (TPS)

Throughput / Capacity

TPS measures how fast the model generates output once it starts responding. TPS is the baseline measurement of LLM performance.

- Expressed as tokens per second
- Rough intuition: 30 TPS maps approximately 360 words per minute (fast, fluid generation)
- Determines how many users and how much output a fixed set of GPUs can sustain

Impact on user experience

- Higher TPS = responses stream smoothly without stalling
- Low TPS = replies feel slow, choppy, or "thinking too long" mid-answer
- TPS primarily affects scale and efficiency, not initial responsiveness

Time To First Token (TTFT)

Latency / Responsiveness

TTFT measures how long a user waits before the first word appears. TTFT is a baseline measurement of LLM "feel"

- Expressed in seconds
- Includes prompt processing, KV cache access, and scheduling
- What users feel as *"Did it start responding yet?"*

Impact on user experience

- Lower TTFT = system feels instant and interactive
- High TTFT = users perceive lag, even if total response is fast
- TTFT dominates perceived quality in chat, copilots, and assistants

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Performance Results: Concurrency 4

Baseline = pure in-VRAM. Delta columns relative to Baseline.

Scenario	Avg TPS	TPS Δ	Avg TTFT	TTFT Δ	P95 TTFT	P99 TTFT	Fail %
Baseline	34.39	—	0.687s	—	4.581s	6.776s	0%
Local SSD	33.94	-1.3%	0.793s	+15.4%	4.567s	6.787s	0%
NVMe-oF	33.79	-1.7%	0.770s	+12.1%	4.564s	6.784s	0%

TPS (Tokens Per Second) measures steady-state generation throughput during inference—higher TPS means the system can serve more tokens (and users) with the same GPUs.
TTFT (Time To First Token) measures latency to the first generated token after a request—lower TTFT directly improves perceived responsiveness in interactive LLM applications.
Together, TPS reflects capacity, while TTFT reflects user experience.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Performance Results: Concurrency 8

Baseline = pure in-VRAM. Delta columns relative to Baseline.

Scenario	Avg TPS	TPS Δ	Avg TTFT	TTFT Δ	P95 TTFT	P99 TTFT	Fail %
Baseline	30.55	—	0.552s	—	1.910s	6.196s	0%
Local SSD	30.38	-0.6%	0.591s	+7.2%	2.033s	6.380s	0%
NVMe-oF	30.00	-1.8%	0.577s	+4.6%	2.020s	6.364s	0%

TPS (Tokens Per Second) measures steady-state generation throughput during inference—higher TPS means the system can serve more tokens (and users) with the same GPUs.
TTFT (Time To First Token) measures latency to the first generated token after a request—lower TTFT directly improves perceived responsiveness in interactive LLM applications.
Together, TPS reflects capacity, while TTFT reflects user experience.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Key Observations: KV Cache Offloading

- Throughput: NVMe-oF within statistical noise of local NVMe. 0.4-point TPS gap at concurrency 4 is not operationally significant.
- TTFT: NVMe-oF beats local SSD at both concurrency levels. Conc 4: 0.770s vs 0.793s (2.9% advantage). Conc 8: 0.577s vs 0.591s (2.4%). RDMA provides queue-depth isolation insulating prefill from eviction traffic.
- Tail Latency: P95/P99 effectively flat across all storage tiers. Tail driven by GPU scheduling and PagedAttention eviction, not storage.
- Reliability: Zero failures across all configurations. Disaggregated path stable under load.
- This test is not just comparing "local vs. remote storage.” It is designed to compare a full production-representative workload with tiered configurations

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

02

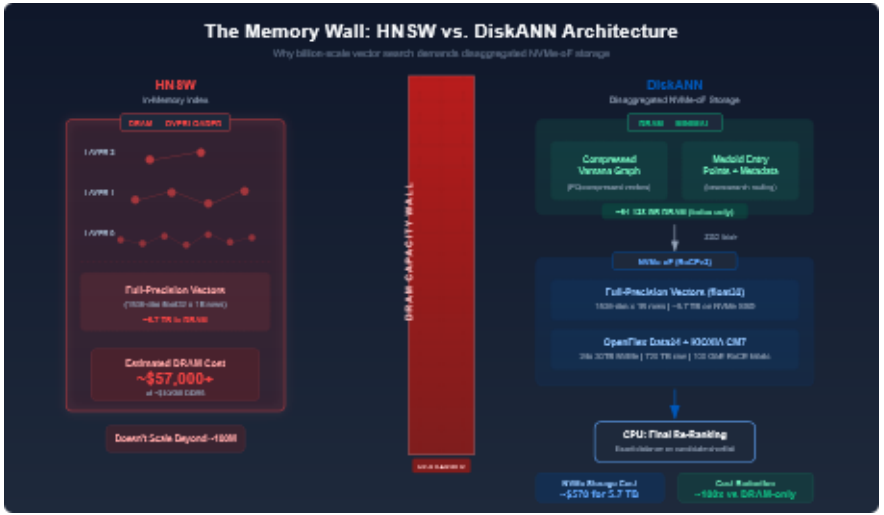
Vector Database Storage with KIOXIA CM7

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

The DRAM Wall for Vector Search

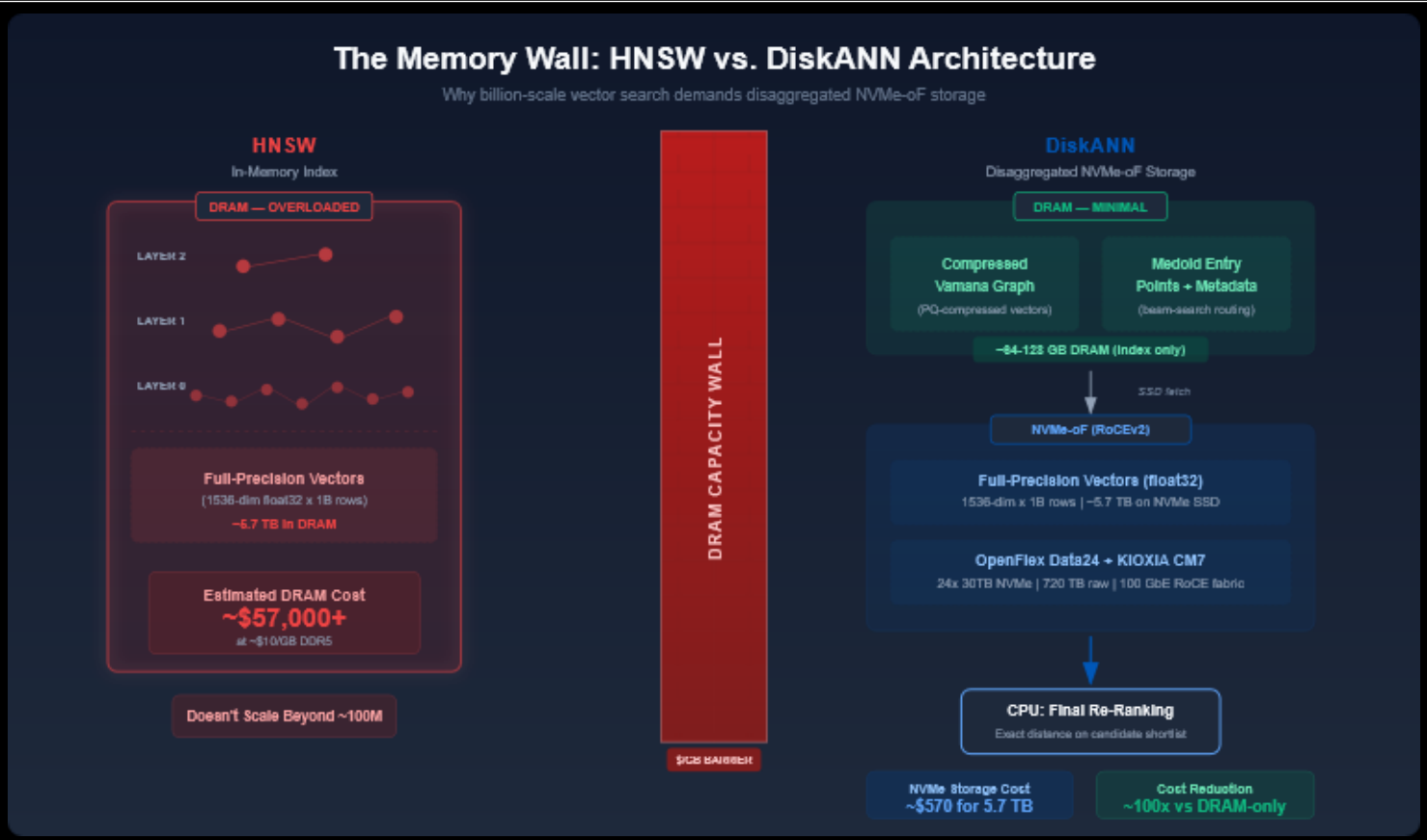
HNSW requires the entire index graph in memory

- 100M vectors x 768 dimensions x 4 bytes = ~307 GB raw embeddings, before HNSW graph overhead.
- Choices: scale DRAM (expensive), quantize/compress (degrades recall), shard across nodes (complexity).
- DiskANN (Microsoft Research): NVMe as primary storage. Small compressed summary in DRAM, full-precision vectors from NVMe for final ranking.
- HNSW is designed for a world where memory is abundant. DiskANN is designed for a world where NVMe is fast.



© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

The DRAM Wall for Vector Search



© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

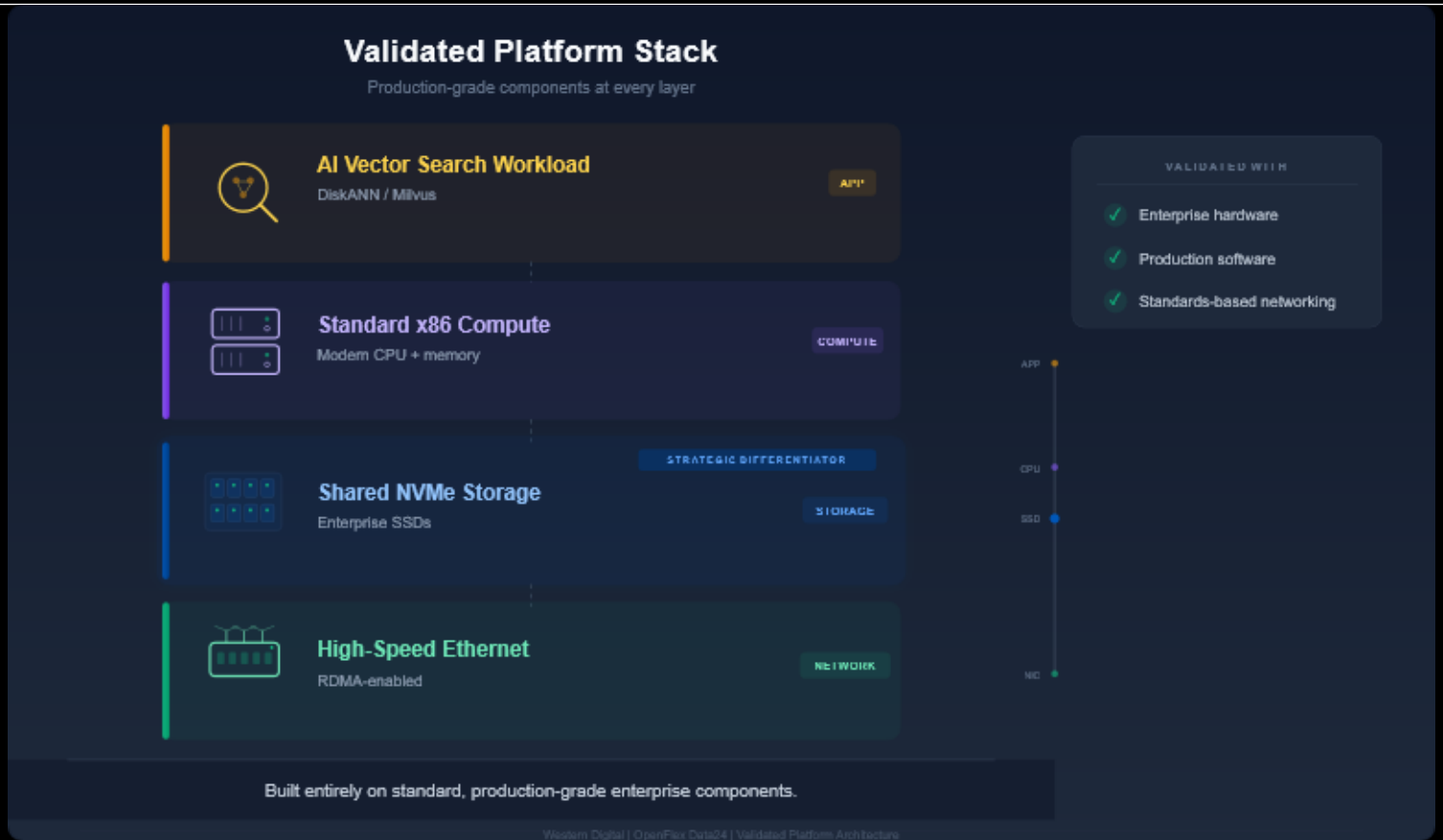
Test Scenarios

Milvus 2.6.9 | DiskANN vs HNSW | Data24 4200 | KIOXIA CM7 30.72 TB | 200 GbE RoCE v2

Configuration	Index	Storage Path	NICs	NVMe Paths
Local HNSW	HNSW	Host DRAM	1	—
Remote DiskANN 1-path 1 NIC	DiskANN	NVMe-oF via RoCE — Data24	1	1
Remote DiskANN 4-path 1 NIC	DiskANN	NVMe-oF via RoCE — Data24	1	4
Remote DiskANN 4-path 2 NIC	DiskANN	NVMe-oF via RoCE — Data24	2	4

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Hardware and Software Stack



© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Hardware and Software Stack

Dell PowerEdge R6615 | AMD EPYC 9454P | 384 GiB DDR5 | Ubuntu 24.04 LTS

- Storage: OpenFlex Data24 4200, PCIe 4.0 backplane, x2 drive lane connectivity.
- SSDs: 3x KIOXIA CM7 (KCMYXRUG30T7) — 30.72 TB, PCIe 5.0 x4, 1 DWPD.
- Fabric: 12x 100 GbE RoCE cables, NVMe/RDMA.
- HNSW: M=30, ef_construction=360, ef_search=100. DiskANN: search_list=100, k=10, Cosine.
- Dataset: 768D float32 vectors at 1M, 10M, 100M corpus sizes.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Vector Database Key Concepts

QPS (Queries Per Second)

Throughput / Scale

QPS measures how many vector search queries the system can complete per second.

- Indicates how much traffic the vector database can handle
- Higher QPS = more users or applications served at once
- Strongly affected by storage latency, parallelism, and multipath I/O

Effect on user experience

- High QPS = fast responses even under load
- Low QPS = slowdowns as concurrency increases
- Determines whether performance holds up at scale

Recall

Accuracy / Result Quality

Recall measures how often the system returns the correct nearest neighbors.

- Expressed as a value between 0 and 1
- Higher recall = results closely match exact search
- Low recall means relevant results are missed

Effect on user experience

- High recall = answers feel relevant and trustworthy
- Low recall = users see wrong or incomplete results
- Critical for RAG, search, and recommendation quality

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Results: 1 Million Vectors

DiskANN 4-path leads at 7,971.8 QPS — 56.9% above local HNSW with higher recall

Configuration	QPS	Recall	p95 Latency	p99 Latency	Load Duration
Local HNSW	5,081.6	0.9799	2.6 ms	2.9 ms	34.7 min
DiskANN 1P/1N	3,713.9	0.9943	4.3 ms	4.6 ms	36.7 min
DiskANN 4P/1N	7,971.8	0.9953	4.3 ms	4.6 ms	36.9 min
DiskANN 4P/2N	6,624.4	0.9953	4.9 ms	5.3 ms	36.9 min

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Results: 10 Million Vectors

DiskANN 4-path delivers 19.4% more QPS than HNSW (665.9 vs 557.8) at equivalent recall

Configuration	QPS	Recall	p95 Latency	p99 Latency	Load Duration
Local HNSW	557.8	0.9835	6.0 ms	6.5 ms	5.84 hr
DiskANN 1P/1N	539.9	0.9962	6.7 ms	7.0 ms	5.96 hr
DiskANN 4P/1N	665.9	0.9960	6.7 ms	7.0 ms	5.96 hr
DiskANN 4P/2N	650.6	0.9955	7.1 ms	7.5 ms	5.99 hr

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Results: 100 Million Vectors

Multipath mandatory. Single-path p95: 2,574 ms. Four-path p95: 24.9 ms. 103x improvement.

Configuration	QPS	Recall	p95 Latency	p99 Latency	Load Duration
Local HNSW	77.9	0.9900	40.6 ms	73.1 ms	57.7 hr
DiskANN 1P/1N	33.3	0.9923	2,574 ms	2,653 ms	58.9 hr
DiskANN 4P/1N	67.2	0.9914	24.9 ms	26.5 ms	59.0 hr
DiskANN 4P/2N	67.2	0.9918	25.7 ms	27.0 ms	59.2 hr

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Key Observations: Vector Database

- QPS: Disaggregated DiskANN is not a compromise. At 1M and 10M, NVMe-oF DiskANN out-queries HNSW with higher recall.
- Recall: DiskANN exceeds HNSW at every scale. 1M: 0.9953 vs 0.9799. 10M: 0.9960 vs 0.9835.
- Multipath at 100M: Single-path = 2,574 ms p95 (unusable). Four-path = 24.9 ms. 103x reduction.
- Dual NIC: Statistically equivalent to single NIC 4-path. One 200 GbE NIC sufficient for tested scales.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

03

Platform Convergence
One Enclosure, Two AI Workloads

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Common Architecture Principles

The same 2U platform addresses two distinct AI infrastructure bottlenecks

Dimension	KV Cache Use Case	Vector DB Use Case
Storage Platform	OpenFlex Data24 4000 Series	OpenFlex Data24 4200
Transport	NVMe/RDMA over 100 GbE	NVMe-oF over RoCE v2 (200 GbE)
NVMe SSDs	Gen4 NVMe (24 drives)	KIOXIA CM7 30.72 TB PCIe 5.0
Key Metric	<1.8% TPS overhead, TTFT lower than local SSD	+57% QPS at 1M, 103x multipath improvement
Proprietary Dependencies	None	None
Scale Model	Independent GPU + storage scaling	Independent compute + storage scaling

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Key Takeaways

The network hop does not materially degrade performance in either use case

- KV Cache: NVMe-oF throughput within 1.8% of baseline. TTFT lower than local SSD. P99 indistinguishable. Zero failures.
- Vector DB: DiskANN outperforms HNSW by 56.9% (1M) and 19.4% (10M) QPS with higher recall. Four-path multipath mandatory at 100M. 103x p95 reduction.
- Platform: Standard Ethernet + RDMA. No InfiniBand, no proprietary drivers, no licensed software. Independent scaling of compute and storage.
- Operations: Namespace reassignment without hardware moves. Capacity scales by adding enclosures. Compute scales independently.
- Value Prop: Limited memory on existing GPU’s, higher DRAM costs in recent months, limited local NVMe, disaggregated storage flexibility.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

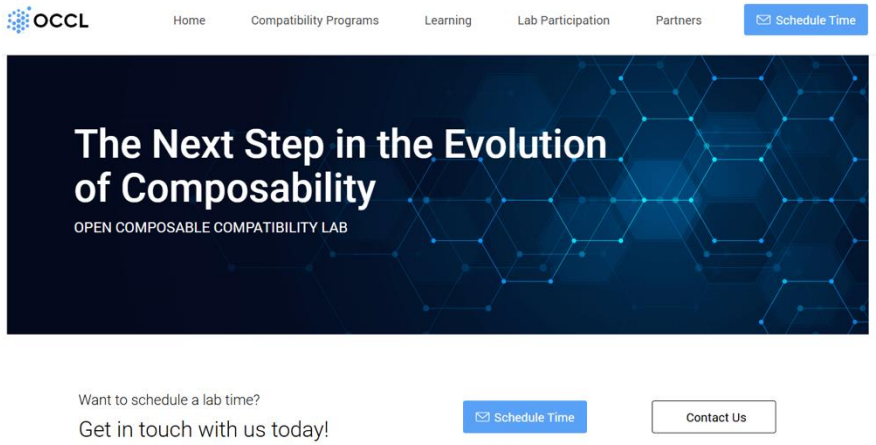
04

Open Composable Compatibility Lab

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

OCCL
History and Purpose

- Traditionally used to demonstrate OC API and Fabric Device Interoperability
- Additional focus on WD Platforms capabilities and solutions:
  - Partner **Collaboration**:
    - Joint Collateral
    - PoC's
  - Thought Leadership etc.



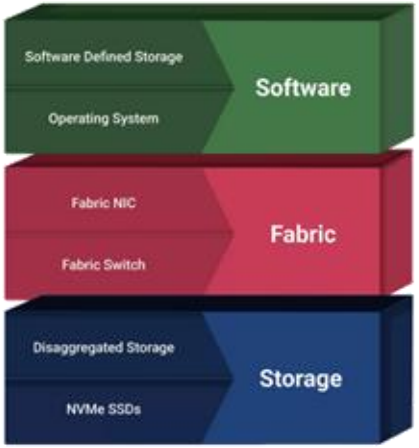
https://www.opencomposable.com/

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Independent Data Storage Infrastructure Provider

Powering the future of data storage through innovation and independence

OCCL Open Composable Compatibility Lab



BeeGFS delivered by thinkparQ · Grald Technology Inc. · MINIO · OSNEXUS · PEAK AIO · RINNOR · ARISTA · BROADCOM · NVIDIA · Western Digital · OpenFlex™ · RapidFlex™ · ingrasys · DapuStor · KIOXIA · PHISON · ScaleFlux · SANDISK

Coming Soon

leil STORAGE · SWISS VAULT · IBM Storage Scale · SOLIDIGM

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benefits

- **Validated Reference Designs**

With validated reference designs, your engineering team gains immediate credibility.

This reduces the time required to fine-tune product specifications or troubleshoot complex interoperability issues, enabling you to meet go-to-market timelines more efficiently.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benefits

- Validated Reference Designs
- **Alignment with Open Standards**

By participating in the OCCL program, partners position their product within an open, composable storage framework that is designed to be vendor-agnostic and future-proof.

This commitment to open standards signals to the market that your SSD is not only innovative but also built to integrate seamlessly into diverse data environments.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benefits

- Validated Reference Designs
- Alignment with Open Standards
- **Enhanced Market Visibility**

Once a project is complete, the collateral will be published on the opencomposable.com website, which is a go-to resource for companies seeking cutting-edge composable storage solutions.

Inclusion on this platform effectively endorses your solutons proven capabilities, raising your profile in a competitive market.

The program also allows you to leverage the credibility of Western Digital’s brand and technical reputation.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benefits

- Validated Reference Designs
- Alignment with Open Standards
- Enhanced Market Visibility
- **Strengthened Customer Confidence**

End-users evaluating storage solutions often look for the assurance that products will interoperate without hidden pitfalls.

By joining a recognized testing program, you demonstrate transparency and reliability.

This is a powerful differentiator when competing for large enterprise and hyperscale deployments, where rigorous vetting is the norm.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benefits

- Validated Reference Designs
- Alignment with Open Standards
- Enhanced Market Visibility
- Strengthened Customer Confidence
- **Accelerated Time to Market**

Validated solutions can see faster acceptance in the data center ecosystem. Commercially, that means quicker transition from product evaluation to design-win, enabling faster revenue realization. Technical compatibility, proven by OCCL testing, can significantly reduce the negotiation time with potential customers who would otherwise conduct extensive in-house evaluation.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benefits

- Validated Reference Designs
- Alignment with Open Standards
- Enhanced Market Visibility
- Strengthened Customer Confidence
- Accelerated Time to Market
- **Long-Term Partnership Opportunities**

The program promotes an ongoing collaboration between Western Digital and partners.

By working closely with our engineering and product teams, you gain early insights into future hardware and software roadmaps.

This fosters deeper partnerships that can lead to co-marketing opportunities, joint innovation, and alignment on industry standards—driving sustained commercial success.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Benefits

- Validated Reference Designs
- Alignment with Open Standards
- Enhanced Market Visibility
- Strengthened Customer Confidence
- Accelerated Time to Market
- Long-Term Partnership Opportunities
- **OCCL Labs (Global VPN access)**

- The OCCL labs (based out of Colorado Springs) enjoys a healthy offering of compute, storage and network infrastructure which allows for partner and customer PoC’s with guided support and best practice guidance from WD platforms SME’s.

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Partner SSD Test Reports

- **Robust Compatibility Validation (SSD)** Gaining approval from Western Digital's OCCL demonstrates that an SSD has been rigorously evaluated for compatibility and stability within the OpenFlex Data24 4200 / 4100 EBOF.
- It also ensures that the SSD meets stringent requirements for data-intensive, latency-sensitive applications. This alignment helps accelerate customer decision-making by providing clear performance metrics and a proven reference architecture.

Validated Partner Drives List

The following is a comprehensive list of SDDs that have been tested using comprehensive evaluation to ensure compatibility, performance, and reliability performed by the Open Composable Compatibility Lab (OCCL). The assessment covers basic interoperability and specialized workload performance using industry-standard benchmarking tools.

For more information related OCCL, see: https://www.opencomposable.com/

Evaluated using OpenFlex Data24 4100 Series NVMe-oF Storage Platform

Drive	Capacity	Form Factor	Interface	Part Number
Kioxia CM7-V	6.4TB	U.3 15mm	PCIe® Gen5, NVMe™ 2.0	KCMYXVUG6T40

Evaluated using OpenFlex Data24 4200 Series NVMe-oF Storage Platform

Drive	Capacity	Form Factor	Interface	Part Number
DapuStor R5100D	15.36TB	U.2 15mm	PCIe® Gen4, NVMe™ 1.4	DPRD31016TT515T3010
Kioxia CM7-V	6.4TB	U.3 15mm	PCIe Gen5, NVMe 2.0	KCMYXVUG6T40
Kioxia CM7-V	12.8TB	U.3 15mm	PCIe Gen5, NVMe 2.0	KCMYXVUG12T8
Phison X200P	3.84TB	U.2 15mm	PCIe Gen5	XX208H023T84P324T0910
ScaleFlux CSD5310	7.68TB	U.3 15mm	PCIe Gen5, NVMe 1.4	CSDU7UVG76
Ultrastar DC SN655	15.38TB	U.3 15mm	PCIe Gen4, NVMe 1.4	WUS5EA1A1ESP7E1
Ultrastar DC SN655	30.72TB	U.3 15mm	PCIe Gen4, NVMe 1.4	WUS5EC0B1ESP7Y1
Ultrastar DC SN655	61.44TB	U.3 15mm	PCIe Gen4, NVMe 1.4	WUS5EC0C1ESP7Y1

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

Partner SSD Test Reports

Dapustor X2900P 400GB Data24-4200 x1 x24

Third Party Test Results are based on a comprehensive evaluation to ensure compatibility, performance, and reliability performed by the Open Composable Compatibility Lab (OCCL). The assessment covers basic interoperability and specialized workload performance using industry standard benchmarking tools. This test result is not an endorsement of the third-party product by Western Digital and no warranty of the product is expressed or implied by Western Digital or the OCCL. For more information related OCCL, see:

https://www.opencomposable.com/

Drive Details

Drive:	Dapustor X2900P
Form Factor:	U.2 15mm
Interface:	PCIe-4, NVMe 1.4a
Security:	N/A
Power:	14W (Active)
Power Idle:	6W
Part Number:	DPXD3101T0S100T4000

Top Line Performance Results — EXTRAPOLATED

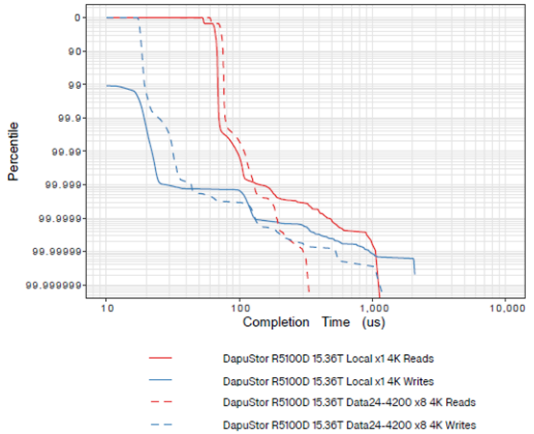
Test Description	Local x1	Data24 x1	Data24 x8	Data24 x24
Read Bandwidth (128KB) MB/s	7,042	7,160	41,488	~124,464
Write Bandwidth (128KB) MB/s	6,794	6,496	36,048	~108,144
Random Read (4KB)K IOPS	1,680	1,710	8,956	~26,867
Random Write (4KB) K IOPS	1,305	1,305	8,621	~25,863
Random Mixed (4KB) K IOPS	1,624	1,618	8,978	~26,935
4K Random Write Latency us	6	16	19	~19
4K Random Read Latency us	19	26	29	~29
4K Random Write 4-9s us	121	23	24	~24
4K Random Read 4-9s us	144	32	34	~34

DapuStor R5100D SSD Test Results

Exceedance Chart

The following charts contain the technical and most accurate information. Exceedance charts are typically used for only single drive, single process, queue depth one analysis as in the following example. The chart is the most fundamental measurement of SSD or disk performance as it shows the minimum latency as well as the expected error rate and latency for the device under test. These results are often referred to as the "number of nines". For example, "4-9s" shows the latency or response time for 9999 of 10000 IOs. The number of IOs grows exponentially with the increase in the number of nines. This chart shows for 6-9s that the best performer for random writes is the small blue dotted line at approximately 90 µs. Random writes are faster than random reads, because random writes are cached in the asynchronous write buffer and are periodically written to the underlying NAND media. Exceedance charts can be run and compared as long as all tests were run on similar systems using the same workload.

Latency Exceedance for DapuStor R5100D 15.36TB



© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

OCCL – Thought Leadership example

- **MLCommons**
  - Mission is to accelerate machine learning innovation and accessibility by developing fair, transparent, and standardized benchmarks, datasets, and best practices. Through collaborative industry and academic partnerships, MLCommons fosters reproducible performance measurement, drives efficiency across AI systems, and ensures that ML technology benefits the broadest possible community worldwide.

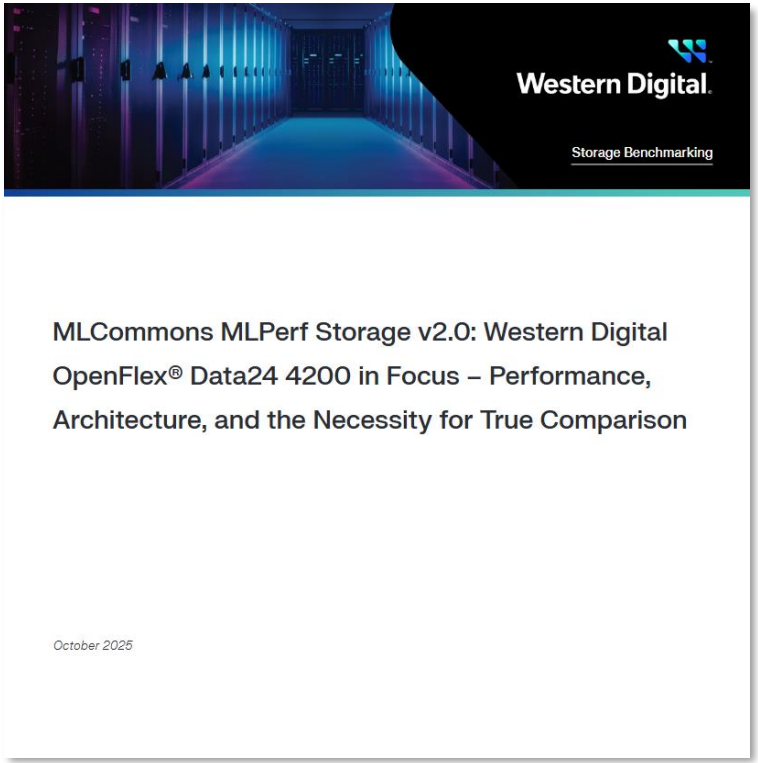
Data24 ResNet50	
Simulated H100 GPU	186
Throughput MB/s	33,301.64

Data24 PEAK ResNet50	
Simulated H100 GPU	52
Throughput MB/s	9,165.78

Data24 3D-Unet	
Simulated H100 GPU	36
Throughput MB/s	101,627.75

Data24 PEAK 3D-Unet	
Simulated H100 GPU	22
Throughput MB/s	61,079.54

https://mlcommons.org/benchmarks/storage/



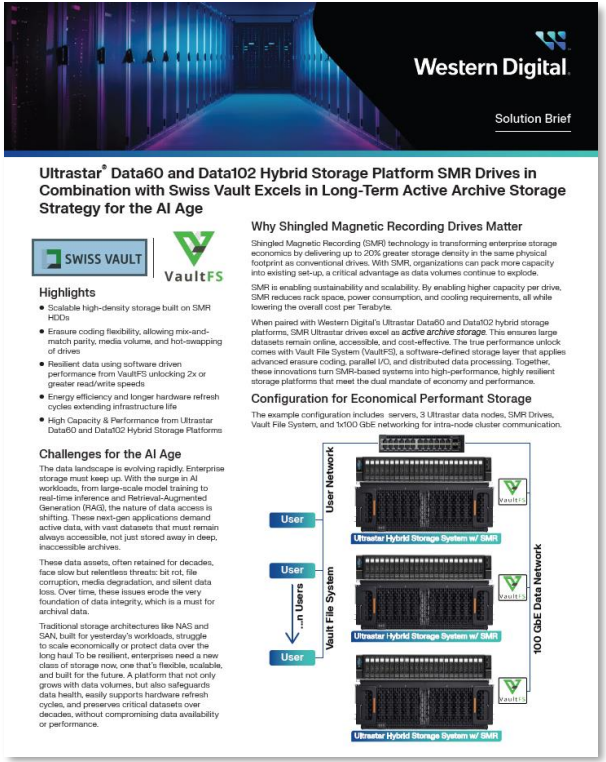
MLPerf - True Comparison

© 2026 WESTERN DIGITAL CORPORATION OR ITS AFFILIATES. ALL RIGHTS RESERVED.

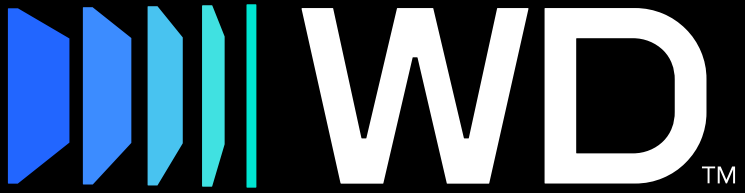
OCCL- Solutions Brief

Summary:

- **Active Archive for the AI Era:** The solution targets long-term, always-online datasets needed for AI training, inference, and RAG, combining cost efficiency with accessibility.
- **High-Density SMR Drives:** Shingled Magnetic Recording (SMR) delivers up to 20% greater storage density and lower cost per terabyte while reducing rack, power, and cooling footprint.
- **VaultFS Software Acceleration:** Swiss Vault's VaultFS unlocks 2× or higher read/write performance through parallel I/O, distributed processing, and customizable erasure coding.
- **Flexible Resilience and Economics:** VaultFS lets operators tune Data+Parity ratios per workload, balancing performance, durability, and capacity to fit business priorities.
- **Scalability and Efficiency:** The architecture scales linearly from a few nodes to hundreds of exabytes, maintaining energy efficiency and extending hardware refresh cycles.
- **Optimized Ultrastar Platforms**: Western Digital's Data60 and Data102 hybrid enclosures provide up to 3.06 PB per 4U, with IsoVibe™ vibration isolation and ArcticFlow™ cooling for performance and reliability.
- **Infrastructure Reuse and Simplified Operations:** Supports existing network and server hardware under one namespace, enabling automated data migration and long-term cost control



VaultFS


WD
TM