OPEN COMPOSABLE COMPATIBILITY LAB

Technical Brief

OpenFlex® Data24 with KIOXIA CM7 Series for Vector Database Storage

Disaggregated NVMe-oF™ via RoCE as a high-performance DiskANN storage backend for Milvus: measured at 1M, 10M, and 100M vector scale on KIOXIA CM7 Series 30 TB NVMe™ SSDs

Modern AI applications, from chatbots that retrieve relevant documents (RAG) to search engines that understand meaning rather than just keywords, rely on vector databases to store and query embeddings: numerical fingerprints that capture the semantic content of text, images, or other data. Each embedding is a long list of floating-point numbers (a vector), and the core operation is always the same: given a new query vector, find the stored vectors most similar to it. This is called Nearest Neighbor search.

At small scale, this is straightforward. At production scale, with hundreds of millions of vectors, it becomes a serious systems problem.

One of the best nearest neighbor search algorithms today is Hierarchical Navigable Small World (HNSW), which builds a layered proximity graph across all vectors, enabling fast, accurate searches. Its weakness is structural: the entire graph must live in host DRAM to function. Do the math on a modest corpus (100 million vectors, each with 768 dimensions, stored as 32-bit floats; and the raw embedding data alone consumes roughly 307 GB. Add the graph structure on top, and you've exceeded what most single-socket servers can hold in memory).

That leaves engineers with two historically unappealing choices: pay more (scale up DRAM, which is expensive) or lose accuracy (compress or prune the index, which degrades recall).

DiskANN, developed by Microsoft Research, offers a third option. It was designed from the ground up to use NVMe SSDs as the primary storage tier, keeping only a small, compressed summary of the index in DRAM for an initial candidate filter, then pulling the relevant graph edges and full-precision vectors from NVMe only for the final ranking step. Because modern NVMe latency (~100µs) is fast enough for this access pattern, DiskANN achieves recall competitive with HNSW - at a fraction of the memory cost.

The architectural insight is simple but important: HNSW is designed for a world where memory is abundant; DiskANN is designed for a world where NVMe is fast.

This technical brief presents VectorDBBench results measuring Milvus 2.6.9 with DiskANN against an in-memory HNSW baseline, using the KIOXIA CM7 Series NVMe SSDs, in a Western Digital OpenFlex Data24 4200 EBOF Storage Platform as a disaggregated NVMe-oF storage target over RoCE v2.

Three dataset scales were evaluated, 1M, 10M, and 100M each with 768-dimensional vectors with remote storage configurations spanning single-path and four-path NVMe-oF access across one and two 200 GbE RoCE NICs. The results quantify both the performance ceiling and the critical importance of multipath configuration at scale.

Challenges

- HNSW index graphs must reside entirely in DRAM, making 100M+ vector corpora impractical on standard single-socket server configurations.
- Locally attached NVMe capacity is constrained by server chassis bay count, coupling vector storage scaling to compute node procurement.
- Single-path NVMe-oF to a remote target can become bandwidth-saturated at large dataset scales, causing catastrophic latency degradation.
- High-recall ANN search requires consistent low-latency storage I/O; any queuing on the storage path directly inflates p95 and p99 query latency.
- Vector database infrastructure must remain operationally manageable as corpus size scales from millions to billions of embeddings.

Highlights

- NVMe-oF DiskANN 4-path exceeds local HNSW Queries Per Second (QPS) by 56.9% at 1M vectors (7,971 vs. 5,081 QPS) with superior recall (0.9953 vs. 0.9799).
- At 10M vectors, 4-path DiskANN delivers 19.4% more QPS than local HNSW (665.9 vs. 557.8) at equivalent recall.
- Multipath is not optional at 100M scale: single-path DiskANN p95 latency is 2,574 ms; four-path significantly reduces it to 24.9 ms, a 103× improvement.
- DiskANN recall consistently exceeds local HNSW recall across all tested configurations and dataset scales.
- RoCE v2 fabric on standard 200 GbE connectivity, requiring no InfiniBand, no proprietary switching, and no specialized host drivers beyond the rdma-core userspace stack.

Test Configuration

Test Scenario

Four configurations were evaluated across three dataset scales (1M, 10M, 100M vectors at 768 dimensions), yielding twelve total measurement sets. All remote configurations targeted the OpenFlex Data24 4200 via NVMe-oF over RoCE v2 (NVMe/RDMA). The HNSW baseline used in-memory storage with Milvus's native graph-based index; DiskANN configurations used disk-resident graph structures managed by the DiskANN index implementation integrated into Milvus.

Configuration	Index	Storage Path	NICs	NVMe Paths	Description
Local HNSW	HNSW	Host DRAM (in-memory)	1	—	In-memory HNSW graph. Establishes maximum QPS ceiling for memory-resident workloads. Requires full graph in DRAM; impractical beyond ~50M 768D vectors on 384 GB host.
Remote DiskANN 1-path 1 NIC	DiskANN	NVMe-oF via RoCE — Data24 4200	1	1	Single NVMe namespace path to the Data24 over a single 100 GbE RoCE NIC. Baseline for disaggregated DiskANN; exposes single-path bandwidth saturation at 100M scale.
Remote DiskANN 4-path 1 NIC	DiskANN	NVMe-oF via RoCE — Data24 4200	1	4	Four NVMe namespace paths over a single 200 GbE RoCE NIC, enabling multipath I/O. Primary production configuration; eliminates single-path latency collapse at large scale.
Remote DiskANN 4-path 2 NIC	DiskANN	NVMe-oF via RoCE — Data24 4200	2	4	Four NVMe namespace paths are distributed across two 200 GbE RoCE NICs for additional bandwidth headroom. Tests the incremental benefit of dual-NIC fabric attach at each dataset scale.

Hardware and Software Stack

Component	Specification
Client server	Dell® PowerEdge® R6615 (1 socket)
CPU	AMD EPYC™ 9454P — 48 cores, 2.75 GT/s
Host memory	384 GiB DDR5-4800 (12 × 32 GiB DIMMs)
OS	Ubuntu 24.04.3 LTS — kernel 6.8.0-94-generic
Vector database	Milvus 2.6.9 with MinIO (RELEASE.2024-12-18T13-15-44Z) and etcd 3.5.25
Benchmark tool	VectorDBBench 1.0.18
Storage target	Western Digital OpenFlex Data24 4200 — PCIe 4.0 backplane, ×2 drive lane connectivity
NVMe SSDs	3× KIOXIA CM7 Series (KCMYXRUG30T7) — 30.72 TB each, PCIe® 5.0 ×4, 2.5-inch, 1 DWPD
Fabric	12× 100 GbE RoCE cables, NVMe/RDMA (NVMe-oF over RoCE)
Index: HNSW params	M=30, ef_construction=360, ef_search=100
Index: DiskANN params	search_list=100, k=10, metric=Cosine
Dataset	768-dimensional float32 vectors (Performance768D) at 1M, 10M, and 100M corpus sizes

[1] One terabyte (TB) is equal to one trillion bytes. Actual user capacity may be less due to operating environment.

Metric Definitions

Metric	Definition	Why It Matters
QPS (Queries per Second)	Number of ANN search queries completed per second as measured by VectorDBBench during the search phase. Higher is better.	Primary throughput indicator for vector search serving. Directly proportional to concurrent query capacity and cost efficiency per query.
Recall	Fraction of true k-nearest neighbors (k=10, cosine similarity) returned by the ANN index. Measured against a brute-force ground truth computed by VectorDBBench. Range: 0.0–1.0.	Accuracy of the approximate search. A recall of 0.99 means 99% of the correct nearest neighbors were returned. Indexes that trade recall for QPS sacrifice result quality.
p95 Latency	95th-percentile query response time in milliseconds. The worst-case latency experienced by approximately 1 in 20 queries during the search benchmark.	SLA-critical metric for interactive workloads. p95 captures the routine tail of the latency distribution and is commonly used for service level objectives.
p99 Latency	99th-percentile query response time in milliseconds. Represents the worst 1 in 100 queries.	Sensitive to storage I/O queuing, NVMe path saturation, and OS scheduling jitter. Disproportionate p99 values relative to p95 indicate bursty contention in the storage path.
Load Duration	Total wall-clock time required to ingest the full vector corpus, build the index, and commit it to the storage backend. Reported in minutes, hours, or days depending on scale.	Operational planning metric for index (re)build scheduling. DiskANN build times at 100M scale span days; this must be factored into index refresh and capacity planning cycles.
Storage Enclosure	OpenFlex Data24	24x NVMe SSDs, dual 100 GbE, 2U rackmount

Benchmark Results

Local HNSW	Remote DiskANN 1-path 1 NIC	Remote DiskANN 4-path 1 NIC	Remote DiskANN 4-path 2 NIC

Dataset: 768D1M — 1 Million Vectors

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
Remote DiskANN 1-path 1 NIC	3,713.9	0.9943	4.3 ms	4.6 ms	36.7 min
Remote DiskANN 4-path 1 NIC	7,971.8	0.9953	4.3 ms	4.6 ms	36.9 min
Remote DiskANN 4-path 2 NIC	6,624.4	0.9953	4.9 ms	5.3 ms	36.9 min

The table contains information related to DiskANN 4-path 1 NIC leads all configurations at 7,971.8 QPS — 56.9% above local HNSW — while also achieving higher recall (0.9953 vs. 0.9799). The 2-NIC configuration trails 1-NIC 4-path at this scale, suggesting the bottleneck is not fabric bandwidth but index traversal or host CPU at 1M corpus size.

Dataset: 768D10M — 10 Million Vectors

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

The table contains information related to DiskANN 4-path 1 NIC delivers 665.9 QPS against local HNSW's 557.8, a 19.4% advantage. Single-path DiskANN (539.9 QPS) falls marginally below the HNSW baseline, confirming that multipath is required for competitive throughput at this scale. Recall for all DiskANN configurations exceeds HNSW (0.9960–0.9962 vs. 0.9835).

Dataset: 768D10M — 100 Million Vectors

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

The table contains information related to Single-path DiskANN at 100M scale suffers 2,574 ms p95 and 2,653 ms p99 latency — a direct consequence of NVMe namespace saturation on a single path at this I/O depth. Red cells indicate pathological values. Four-path configuration resolves this entirely: p95 drops to 24.9 ms, a 103× improvement.

Observations

QPS: DiskANN on NVMe-oF Outperforms In-Memory HNSW at 1M and 10M

The most immediate result is that disaggregated DiskANN is not a compromise: at both 1M and 10M vector scales, a correctly configured NVMe-oF DiskANN deployment out-queries an in-memory HNSW baseline on the same hardware. At 1M, 4-path DiskANN (1 NIC) reaches 7,971.8 QPS against HNSW's 5,081.6 — 56.9% more throughput from a disaggregated storage-backed index. At 10M, the advantage narrows to 19.4% (665.9 vs. 557.8 QPS), reflecting the higher per-query I/O cost as the graph grows. The QPS advantage derives from DiskANN's graph structure, which is optimized specifically for sequential NVMe access patterns, combined with the Data24's ability to sustain those access patterns over multiple concurrent NVMe paths without contention.

Recall: DiskANN Delivers Higher Accuracy Than HNSW Across All Scales

DiskANN recall is consistently higher than HNSW at every tested scale and every remote configuration. At 1M, DiskANN 4-path scores 0.9953 against HNSW's 0.9799, a 1.5-point gap. At 10M, DiskANN achieved 0.9960–0.9962 versus HNSW's 0.9835, a 1.3-point gap. At 100M, parity is essentially restored (0.9914–0.9923 DiskANN vs. 0.9925 HNSW), within noise. For RAG pipelines where retrieval accuracy directly affects generation quality, this is not a secondary metric.

Multipath Is Mandatory at 100M Scale: A 103× Latency Difference

The 100M single-path setup clearly shows the value of multipathing to NVMe-oF targets. At single path the test yields a p95 latency at 2,574 ms and p99 at 2,653 ms which is not viable for production use. For 100M vectors, a single NVMe path can't handle concurrent DiskANN queries efficiently. Using multiple NVMe-oF paths eliminates saturation, lowering p95 to 24.9 ms and p99 to 26.5 ms, making it suitable for production. Four-path deployment is required for this scale.

Dual NIC: Headroom With Minimal Incremental Gain at Current Scale

Adding a second 200 GbE RoCE NIC (4-path across 2 NICs) produces results that are statistically equivalent to 4-path on a single NIC at all three tested scales. At 1M, the 2-NIC configuration trails the 1-NIC 4-path by 17.0% in QPS (6,624 vs. 7,971), likely due to path balancing overhead or NUMA effects at this dataset size. At 10M and 100M, QPS and latency are nearly identical between the two configurations (10M: 650.6 vs. 665.9 QPS; 100M: 67.20 vs. 67.22). The practical implication is that a single 200 GbE RoCE NIC with four-path NVMe-oF is sufficient for the tested scales, and the second NIC provides bandwidth headroom for corpus growth or additional concurrent workloads rather than a latency or throughput improvement at current corpus sizes.

Disaggregated Storage Architecture

The OpenFlex Data24 4200 used in this benchmark houses up to 24 NVMe SSDs in a compact 2U enclosure with a PCIe 4.0 backplane and dual-lane per-drive connectivity. In this test, three KIOXIA CM7 Series drives at 30.72 TB each provided the storage backend, connected via twelve 100 GbE RoCE cables. The Data24 presents each drive as an independent NVMe namespace, allowing multipath NVMe-oF connections from the host server to distribute I/O across multiple physical paths to the same or different physical SSDs.

The four-path configuration used in testing connects the host to four independent NVMe namespaces on the Data24 over RoCE. DiskANN's index is striped or distributed across these namespaces by the Milvus storage layer, so concurrent graph traversal queries can issue I/O to multiple NVMe namespaces in parallel without serializing through a single path's queue depth. At 100M vectors, this parallelism is what separates a functional deployment (24.9 ms p95) from an unusable one (2,574 ms p95).

Scale-out Dimension	Approach
More vectors per node	Add NVMe drives to the Data24 (up to 24 SSDs). Each KIOXIA CM7 Series at 30.72 TB provides significant headroom for corpus growth.
More concurrent indexes	Allocate additional NVMe namespaces per index. DiskANN indexes for different collections can be isolated to separate namespace groups.
More query throughput	Increase NVMe path count (four-path is the tested production minimum). Add a second 200 GbE RoCE NIC for additional fabric bandwidth.
More index nodes	Deploy additional compute nodes, each connecting to designated namespaces on one or more Data24 enclosures. Storage scales independently of compute.
Higher resiliency	Data24 supports multipath NVMe-oF; path failover can be configured at the NVMe-oF subsystem level without application changes.

Deployment Requirements

Component	Requirement
Storage Enclosure	Western Digital OpenFlex Data24 4200 (up to 24 NVMe SSDs, 12× 100 GbE RoCE uplinks)
NVMe SSDs	KIOXIA CM7 Series or equivalent enterprise PCIe 5.0 NVMe (30 TB+ per drive for large corpora)
Network	200 GbE RoCE fabric — standard data center Ethernet switches with PFC/ECN configuration for RDMA
Host NIC	RoCE-capable NIC (e.g., NVIDIA ConnectX-7 or equivalent); one NIC minimum, two for headroom; RDMA userspace stack (rdma-core)
NVMe paths	Minimum 4 NVMe-oF paths for 10M+ vector workloads. Single-path is not viable at 100M scale.
Host CPU/memory	Single-socket AMD EPYC or equivalent; DiskANN DRAM footprint is a fraction of HNSW — 384 GB sufficient for 100M vectors
OS	Linux® kernel 6.x with nvme-rdma / nvme-fabrics modules; Ubuntu 24.04 LTS validated
Vector database	Milvus 2.6.9 with DiskANN index type; MinIO and etcd as configured by default Milvus deployment
Benchmark framework	VectorDBBench 1.0.18 for reproducible methodology

Conclusion

Three results from this benchmark are noteworthy. First, disaggregated NVMe-oF DiskANN is faster than local in-memory HNSW at 1M and 10M vector scales, not comparable, not close: 56.9% and 19.4% more QPS, respectively, with higher recall. While this can be attributed due to PCIe Architectures and server design it is notable. Any notion that disaggregated storage carries some inherent performance penalty is falsified at these workload types and scales.

Second, at 100M vectors, the choice of path configuration is the single most important deployment decision. Single-path NVMe-oF produces p95 latency of 2,574 ms. This shows that incorrectly configured deployments are unusable for any interactive workload. multi-path configuration on the same fabric and hardware produces 24.9 ms p95. This reiterates the key difference between a functional system and a non-functional one. Every deployment targeting 100M+ vectors must be configured with a minimum of four NVMe-oF paths.

Third, the Data24 4200 with KIOXIA CM7 Series NVMe SSDs provides the storage bandwidth and queue depth necessary to sustain DiskANN's access pattern at all tested scales on standard 200 GbE RoCE infrastructure. No specialized fabric, no InfiniBand, no proprietary drivers. The operational advantage of disaggregation, independent scaling of storage and compute, shared storage pools, simpler server BOM, and lower TCO, comes without a performance cost when the system is correctly configured.

OpenFlex Data24 4000 Series Storage Platform

The OpenFlex Data24 4000 series NVMe-oF storage platform extends the high performance of NVMe flash to shared storage. The 4000 series provide low -latency sharing of NVMe SSDs over a high-performance Ethernet fabric to deliver similar performance to locally attached NVMe SSDs. Western Digital RapidFlex NVMe-oF controllers, allows up to six dual pathed hosts to be attached without a switch.

The OpenFlex Data24 4000 series uses three of Western Digital's RapidFlex A2000 Fabric Bridge Adapters per IOM to provide up to 12 ports of 100GbE which can connect to RDMA and/or RDMA configured host ports.

Kioxia CM7 Series

KIOXIA CM7 Series enterprise NVMe SSDs support EDSFF E3.S and 2.5-inch form factors and are PCIe 5.0 and NVMe 2.0 specification compliant. They are available in read-intensive (1 DWPD4, up to 30.72 TB) and mixed-use (3 DWPD, up to 12.8 TB) endurances. Additional features include a dual-port design for high availability applications, flash die failure protection and security/encryption options.

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