signal65

Local Vector Search at 10-Billion Scale

AI Performance on Dell PowerEdge R770 with Dell PERC H975i

AUTHOR

Brian Martin

AI Data Center Performance | Signal65

IN PARTNERSHIP WITH

DELL Technologies

JUNE 2026

Executive Summary

Enterprise AI has moved beyond model experimentation and into operational deployment. At this stage, success is no longer defined only by model quality, but by whether the infrastructure can deliver data to those models reliably, consistently, and cost-effectively at production scale. For organizations deploying retrieval-augmented generation (RAG), vector search, and real-time inference, the storage layer has a direct impact on user experience, service levels, infrastructure efficiency, and ultimately business value. When storage becomes the bottleneck, AI investments underperform, latency rises, and expensive compute resources sit underutilized.

Key Highlights

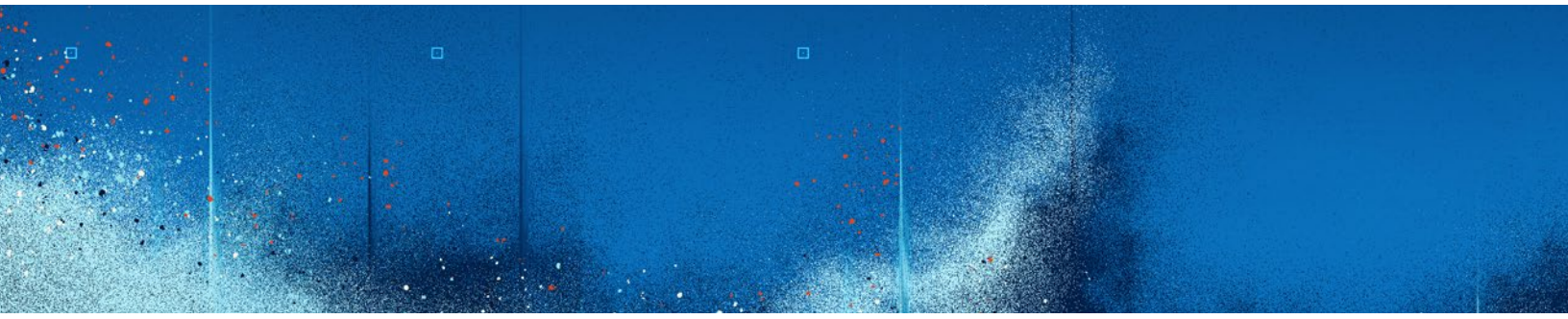
Production Scale
10 billion vectors with 768 dimensions requiring 29TB on-disk local storage

NVMe Performance
Over 50GB/sec and 860 queries per second from storage-based indexes

Resilient Performance
Less than 10% query and 15% bandwidth impact during RAID5 rebuild

In prior work, AI Storage Pipeline Acceleration with Dell PERC13 established that the Dell PERC H975i (PERC13) controller can deliver industry-leading synthetic performance, reaching up to 56 GB/s of sequential throughput and 13 million IOPS. Those results demonstrated the platform's technical ceiling. The more important question for enterprise decision makers, however, is whether that performance holds up in real deployments, where access patterns are irregular, concurrency is high, and data protection cannot be compromised. That is the difference between an impressive benchmark and a platform that can be trusted in production.

This paper answers that question using a 10-billion-vector FAISS index, sized to exceed system memory and validate storage-based queries, on a Dell PowerEdge R770 equipped with dual PERC H975i controllers, 32 NVMe drives in RAID5 configuration, and 512 GiB of DDR5 memory. Under this realistic workload, the platform delivered 51.6 GB/s while serving concurrent queries at scale and maintaining RAID5 protection, demonstrating high-performance AI infrastructure does not have to come at the expense of resilience. With the right storage architecture and proper tuning, organizations can accelerate retrieval-intensive AI workloads, protect critical data, and improve infrastructure efficiency while reducing deployment risk and increasing confidence in production-scale AI.



The Changing Nature of AI Storage Requirements

As AI workloads have evolved, so have the demands they place on storage infrastructure. Inference serving introduces sensitivity to latency because requests are user facing and response time matters. RAG adds another layer of complexity by combining vector database operations with document retrieval in a workload that is both highly concurrent and inherently unpredictable. The practical consequence is that production systems must simultaneously support high throughput, acceptable latency, and strong concurrency while preserving predictable behavior and protecting data.

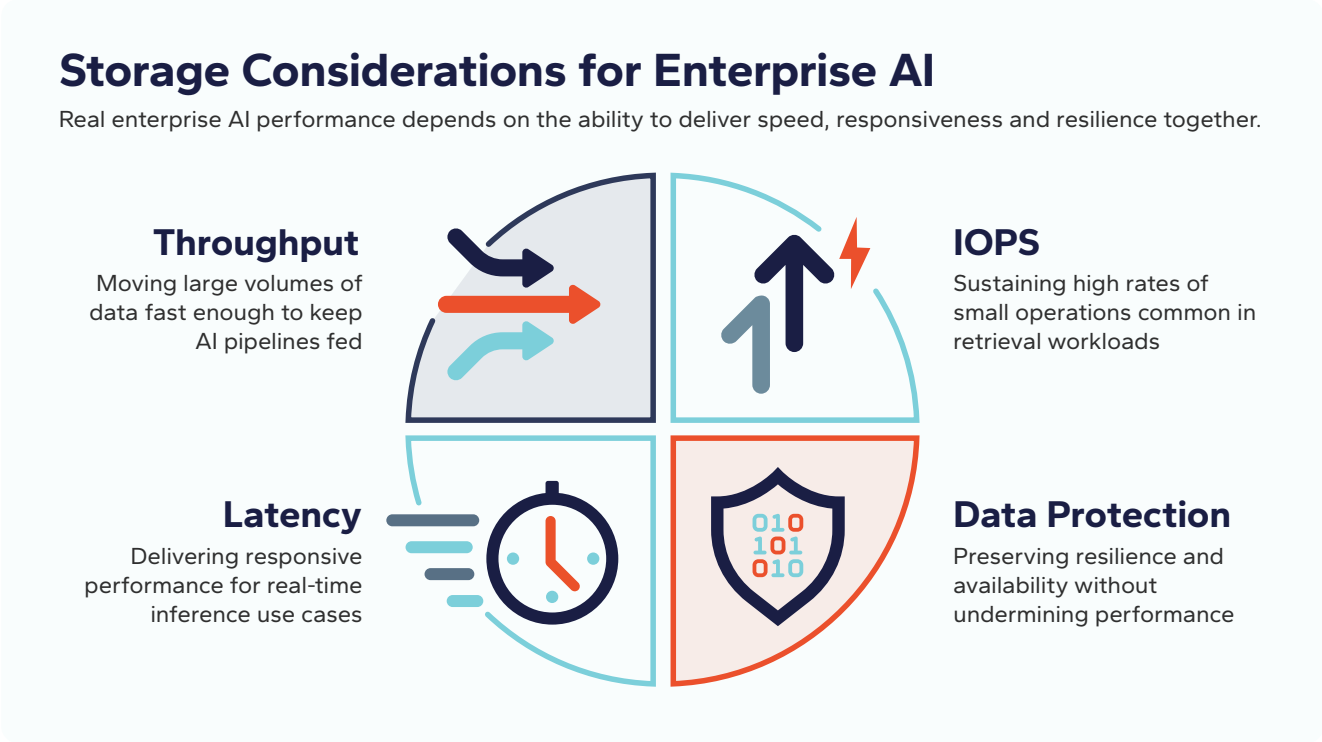


Figure 1: *Storage Considerations for Enterprise AI*

Among today's AI workloads, vector search stands out because it places simultaneous pressure on storage, compute, and response time. Each query can trigger multiple retrieval operations, move substantial volumes of data, and require intensive similarity calculations, all within production expectations. At smaller scales, memory and caching can conceal these demands. At 10-billion-vector scale and above, those buffers are no longer sufficient, and the true performance of the underlying storage infrastructure is revealed.

Vector Search in AI Workloads

Vector search workloads stress storage systems differently than traditional database workloads; for example, a single vector search query using IVF-PQ must read multiple inverted lists from disk, pulling in thousands of compressed vectors per request. At our tested scale of 10-billion vectors and 131,072 IVF cells, each list contains roughly 38,000 compressed vectors. With allocation padding, each list occupies approximately 4.4 MiB of contiguous on-disk storage. A query searching 16 lists, therefore reads ~70 MiB, larger than a typical database page read, but smaller than a sequential scan.

Which lists an individual query request accesses depends on the query vector itself. The access pattern is neither fully random (lists are contiguous) nor fully sequential (different queries hit different lists). This hybrid pattern challenges storage systems optimized for one extreme or the other.

Production vector databases serve many queries simultaneously. Each concurrent query generates its own I/O stream, creating high queue depths on the storage controllers. The system must maintain consistent latency under concurrent load.

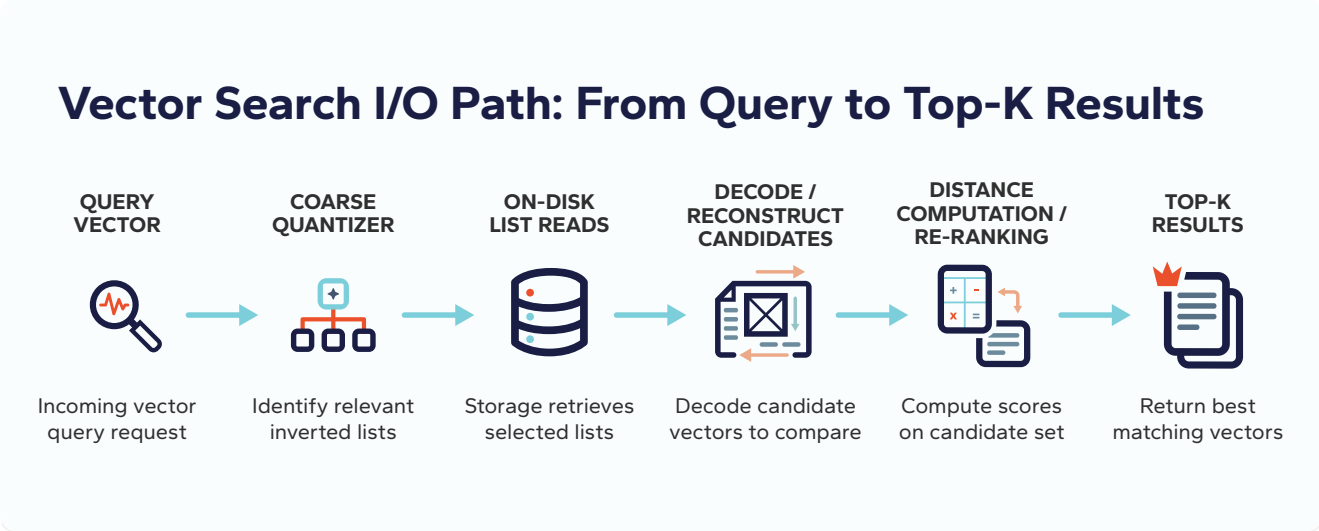


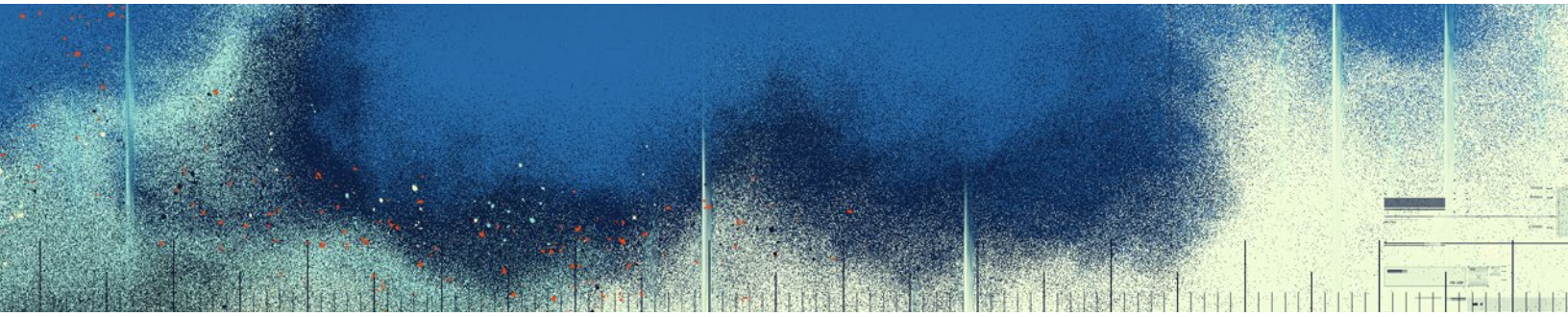
Figure 2: *Vector Search I/O Path*

Performance and Protection by Design

The architectural components in this study are designed and built to reflect enterprise priorities, where performance must be delivered alongside resilience and operational predictability. At the center are two Dell PowerEdge RAID Controller H975i modules, each attached through a 16-lane PCIe Gen5 host interface with Broadcom RAID technology. These controllers transform raw flash into managed, protected, policy-driven storage that preserves enterprise safeguards while sustaining high-performance access. This storage foundation is paired with the Dell PowerEdge R770, a 2U, two-socket platform designed to combine performance with power efficiency. With support for two Intel Xeon 6 processors, 32 DIMM slots, dense EDSFF E3.S NVMe topologies, and PCIe Gen5 connectivity, the R770 brings together the compute, memory, storage, and I/O flexibility required for this role in AI data infrastructure.

Figure 3: *Dell PowerEdge R770*

The earlier Signal65 PERC13 study demonstrated Dell hardware RAID delivers substantial gains in bandwidth, IOPS, latency, and rebuild behavior, creating a strong foundation for storage-intensive AI pipelines. In this study, configuring RAID5 is a deliberate continuation of that story. RAID5 offers an attractive balance for read-dominant AI workloads: it preserves far more usable capacity than mirrored layouts, maintains single-drive fault tolerance, and retains the strong read characteristics needed for large-scale retrieval and vector search.



NUMA Scaling Considerations

An important consideration for maximizing vector search performance is awareness of the system's NUMA (Non-Uniform Memory Architecture) layout. We configured each controller to operate in a dedicated locality domain, keeping memory access close to the compute resources serving the workload and reducing cross-socket contention. That improves consistency as well as raw performance, which matters because deterministic behavior under load is more valuable in production than a single peak result.

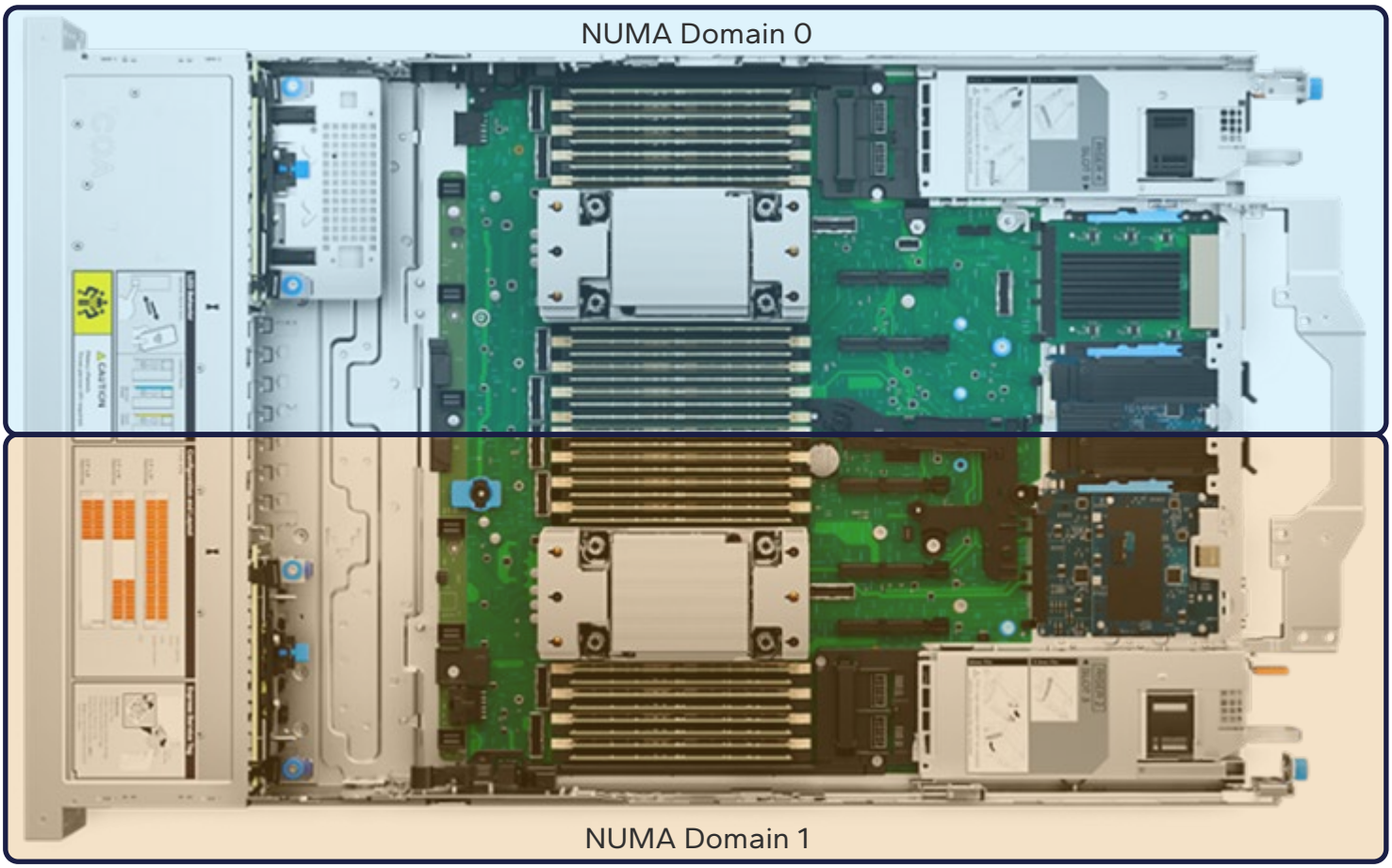


Figure 4: *Separate NUMA domains for storage, CPU, memory, and GPU*

By sizing the index beyond available system memory, and configuring direct IO, the testing ensures measured results reflect genuine storage behavior rather than DRAM-assisted caching effects. This more demanding workload is a more useful representation of how enterprise AI systems behave as deployments move beyond proof-of-concept scale and memory supply chains remain congested.

Hardware Configuration

Component	Specification
Server	Dell PowerEdge R770
Processors	Two Intel Xeon 6760P (64 cores/128 threads each)
Memory	512 GiB DDR5-5200 (16 x 32 GB DIMMs)
Storage Controllers	Two Dell PERC H975i (NVMe)
NVMe Drives	32 x 3.2 TB NVMe (102.4 TB, 76.8 TB formatted)
RAID Configuration	8 x RAID5 sets, 64 KB stripe
GPUs	Two NVIDIA L40S

NUMA Architecture

Domain	Threads	Memory	Controller	GPU
Domain 0	128	256 GB	PERC 0	L40S 0
Domain 1	128	256 GB	PERC 1	L40S 1

Each socket and attached storage forms an independent NUMA domain. Software performs index construction, search, and result processing most efficiently on the same domain as the storage controller, minimizing memory traffic across the inter-socket link.

Software Environment

Component	Version
Operating System	Ubuntu 24.04 LTS
Kernel	6.8.0-106-generic
FAISS	1.14.1
Python	3.12.1
Filesystem	XFS (4 KB blocks)

Vector Index Configuration

Parameter	Value
Total Vectors	10,000,000,000
Dimensions	768
Raw Size	29 TB
Index Type	IVF131072 + PQ96x8
On-Disk Size	1 TB
Compression Ratio	~30x

The index uses IVF (Inverted File) partitioning into 131,072 coarse quantizers with PQ (Product Quantization) at 96 dimensions × 8 bits for fine-grained refinement. This configuration balances search quality, inference speed, and storage footprint.

From Raw Capacity to Realized Performance

A key finding in this study is the gap between default and optimized performance. With out-of-the-box system settings, throughput was limited to about 1 GB/s, well below what synthetic benchmarks suggested was possible. That gap underscores an important reality: hardware potential does not automatically become application performance. In AI infrastructure, results are determined by end-to-end system behavior, not by the specification of any single component.

Through targeted tuning of kernel behavior, runtime concurrency, and compute efficiency, the same hardware ultimately reached 51.6 GB/s of sustained throughput. The controller, drives, and server did not change; the software stack did. Infrastructure performance is not simply purchased; it is achieved through careful design, tuning, and validation. It also reframes the earlier benchmark study: those synthetic results were not just impressive peak numbers, but a realistic ceiling that could be approached once the surrounding system was tuned to expose, rather than hide, the controller's capabilities.

Reaching that level required removing bottlenecks across the stack. Default page-cache behavior was poorly matched to the application's large, irregular read pattern, making direct IO the more efficient path. The Python GIL constrained concurrency, so multiprocessing became essential to generate enough parallel I/O. As storage performance improved, vector decode emerged as the next bottleneck, requiring optimized AVX-512 implementations to keep pace. Block-layer settings also mattered, influencing how efficiently large requests reached the controller. The broader point is that each bottleneck masked the next. That is typical of real AI systems: gains in one layer only matter if the rest of the pipeline can absorb them. Full performance comes from system-level alignment, where controller capability, kernel policy, application architecture, and compute efficiency work together.

Optimization Progression

Configuration	Throughput (GB/s) Improvement
Baseline (default)	1.0x
+ Readahead (16 MB)	8.2x
+ O_DIRECT	18.4x
+ PQ AVX-512	29.4x
+ max_sectors_kb	51.6x

Final Results

Queries Per Second (QPS) and system utilization metrics were captured and are presented below across various nprobe values and number of workers per node.

Workers/ node	nprobe	QPS	CPU%	sys%	iowait%	GB/s
16	4	468	12.8%	0.6%	0.5%	22.0
16	16	175	12.7%	0.7%	1.0%	33.1
16	64	48	12.7%	0.8%	1.2%	36.4
32	4	764	24.7%	1.1%	1.3%	36.0
32	16	266	24.5%	1.3%	4.3%	50.5
32	64	68	24.7%	1.3%	4.8%	51.6
64	8	540	48.9%	3.6%	13.8%	50.8
96	4	860	75.1%	6.1%	2.5%	40.3
96	16	264	70.2%	6.0%	29.1%	50.1

Peak Performance Summary

Metric	Value
Peak QPS	860
Peak Throughput	51.6 GB/s
CPU Utilization	75.1%

QPS-Throughput Tradeoff

The index uses IVF partitioning into 131,072 coarse quantizers with PQ at 96 dimensions × 8 bits for fine-grained refinement. This configuration balances search quality, inference speed, and storage footprint.

QPS in Context

Published vector database benchmarks report QPS at smaller scale and lower dimensionality, and typically allow in-memory caching:

System	Scale	Dims	QPS	Cached?	Storage BW
DiskANN (Microsoft)	1B	128	~2,000–5,000	Partial (SSD)	2.85 GB/s
FlashANNS (2025)	1B	128	~5,000–15,000	No (SSD)	~17–35 GB/s
ScyllaDB (2026)	1B	128	252,000	Likely fully cached	Not measured
Couchbase (2025)	1B	128	703 @93% recall	Not disclosed	Not measured
This study (cached)	**10B**	**768**	**2,004**	**Yes (mmap, ra=16M)**	**~0 (cached)**
This study (O_DIRECT)	**10B**	**768**	**860**	**No (guaranteed)**	**40.3 GB/s**

RAID5 Resilient Performance

Enterprise storage must continue serving queries when hardware fails. We tested the two key operating points, peak throughput and peak QPS — under a simulated single NVMe drive failure and during active RAID5 rebuild. Response times are shown for the 50th percentile (mean), 95th, and 99th (outliers) percentiles.

Condition	Config	QPS	p50	p95	p99	Throughput
Healthy	32w, nprobe=64	68	862 ms	1,207 ms	1,342 ms	51.6 GB/s
Degraded	32w, nprobe=64	57	1,010 ms	1,450 ms	1,657 ms	43.8 GB/s
Rebuilding	32w, nprobe=64	57	1,019 ms	1,492 ms	1,666 ms	43.3 GB/s
Healthy	96w, nprobe=4	860	204 ms	222 ms	231 ms	40.3 GB/s
Degraded	96w, nprobe=4	788	208 ms	296 ms	343 ms	37.2 GB/s
Rebuilding	96w, nprobe=4	791	210 ms	306 ms	358 ms	37.4 GB/s

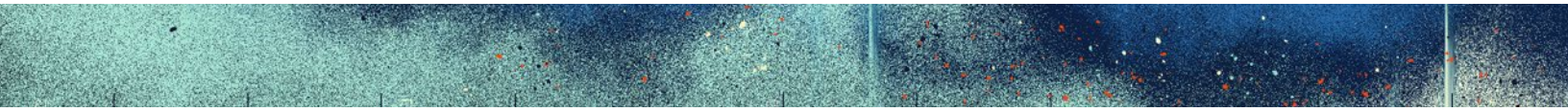
One drive in the 4-drive RAID5 group backing NUMA node 0 was taken offline via perccli2, degrading one of eight volumes while the other seven remained healthy. After the degraded benchmark, the drive was brought back online and the same tests rerun during active rebuild.

Throughput impact is modest: 15% at peak bandwidth, 8% at peak QPS. The degraded volume must reconstruct reads from parity across the remaining three drives, adding latency to NUMA node 0 I/O while NUMA node 1 operates normally. The throughput-heavy configuration (nprobe=64) is more affected because it issues more I/O per query through the degraded volume.

Rebuild adds no measurable overhead. Degraded and rebuilding performance are within measurement variation — the PERC H975i's background rebuild does not compete with foreground read traffic at these throughput levels.

Tail latency widens more than median. At the peak QPS config, p50 increases only 2% (204 -> 208 ms) but p99 increases 48% (231 -> 343 ms). This reflects occasional reads that hit the degraded volume's parity reconstruction path, while the majority of queries — those routed to healthy volumes or to cached parity — see minimal impact.

Condition	Peak BW (GB/s)	Peak QPS
Healthy	51.6	860
Degraded	43.8	788
Rebuilding	43.3	791



Key Insights for Enterprise Deployment

PERC13 Performance Translates to Real Workloads

The PERC H975i achieves 91% of synthetic peak (51.6 GB/s vs 56 GB/s theoretical) on a realistic 10-billion-vector search workload. This validates that enterprise storage specifications provided by Dell reflect genuine capability on production AI workloads.

Software Configuration is Dominant Variable

A 51x performance range exists between default and optimized configurations without changing hardware. This demonstrates that procurement success depends primarily on deployment expertise and system tuning rather than component selection.

Block I/O Sizing Matters More than Raw IOPS

Tuning max_sectors_kb from 192 KB to 1024 KB increased throughput 74% while the system sustained only 55,000 IOPS throughout. The key metric is bytes per I/O operation, not the count of operations.

NUMA Isolation is Non-Negotiable

Binding storage controllers, memory, and computation to single NUMA domains eliminated cross-socket memory traffic and improved performance by eliminating memory bandwidth contention.

Memory Requirements are Modest

The 10-billion-vector search uses less than 3 GB per NUMA domain in working memory despite the index exceeding 900 GB on disk. The system achieves high performance through efficient buffer reuse rather than large in-memory caches.

The Storage Subsystem is Genuinely Capable

Enterprise storage is often over-provisioned for traditional database workloads. Vector search reveals the true performance envelope: a properly configured R770 with PERC13 controllers can sustain >50 GB/s on realistic AI workloads.

Scaling Beyond 10 Billion Vectors

The R770 base configuration addresses the 10 billion vector scale. Scaling to 30 billion or 40 billion vectors requires only larger drives, enabling 4x the vector count without additional server investment.

Capacity Scaling

Vector Count	Drive Size	Drives	Raw	RAID5 Usable	Raw Vectors	Index Size
10B	3.2 TB	32	102.4 TB	76.8 TB	28.6 TB	~1 TB
20B	3.2 TB	32	102.4 TB	76.8 TB	57.2 TB	~2 TB
30B	6.4 TB	32	204.8 TB	153.6 TB	85.8 TB	~3 TB
40B	6.4 TB	32	204.8 TB	153.6 TB	114.4 TB	~4 TB

Expected Search Performance

Search performance scales linearly with vector count at constant QPS. A 30 billion vector index requires 3x the I/O bandwidth to achieve the same query rate, resulting in 3x lower QPS. Barring improvement to the current 51 GB/s sustained throughput, the expected QPS decreases proportionally.

Vector Count	Max Throughput (GB/s)	QPS at Throughput Saturation
10B	51.6	860
20B	51.6	428
30B	51.6	285
40B	51.6	214

Incremental Index Updates

Real-world deployments require adding new vectors to existing indexes. The incremental update mechanism reuses the tuning parameters established for initial index construction. Small batches can be added without full index rebuilds, providing flexibility for production systems.

GPU-Accelerated PQ Decode

Currently, vector reconstruction occurs on CPU. L40S or RTX Pro 4500 GPUs in the system could potentially handle PQ decode operations, freeing CPU resources for other inference tasks. A hybrid approach—streaming compressed vectors from NVMe to GPU for decode—could provide an additional 2-3x performance improvement without additional storage hardware.

The main challenge is the index exceeds GPU memory by approximately 20x, requiring a continuous streaming pipeline rather than in-GPU caching. This engineering effort remains future work.

Implications for Enterprise AI Infrastructure

Storage is often viewed as secondary to compute in AI systems. This work demonstrates that storage itself is a primary component of the AI inference stack. For vector search and similar data-intensive workloads, the storage subsystem performance directly influences end-user query rates and system throughput.

Local NVMe storage with RAID protection offers a compelling alternative to network storage. The R770 baseline configuration provides superior performance to network storage at comparable cost, with the added benefit of eliminating network dependencies and providing fault isolation.

Procurement specifications alone do not guarantee performance. Successful deployments require tuning expertise, understanding of OS-level I/O scheduling, and willingness to deviate from default configurations. Organizations building AI infrastructure must invest in systems expertise to realize the full potential of hardware investments.

> Peak performance matters less than repeatable, protected performance. Optimize for production conditions: RAID5 resilience with validated tuning parameters and procedures.

References

- Signal65, “AI Storage Pipeline Acceleration with Dell PERC H975i (PERC13),” May 2025. https://signal65.com/research/ai/ai-storage-pipeline-acceleration-with-dell-perc-h975i-perc13/
- Meta AI, “FAISS: A Library for Efficient Similarity Search,” https://github.com/facebookresearch/faiss.
- SNIA, “Storage Requirements for AI,” CMSS24-Cardente, 2024. https://www.snia.org/educational-library/storage-requirements-ai-2024

Conclusion

Ultimately, this architecture shows that enterprise AI infrastructure can deliver **performance, protection, and predictability together.** For buyers, that matters because AI success at scale depends on more than model quality alone. It depends on whether the underlying platform can support **realistic scale**, sustain **high performance** under production conditions, and do so with strong **energy and infrastructure efficiency**. In this study, realistic scale is not an abstract claim; it is demonstrated through a 10-billion-vector deployment sized beyond system memory, reflecting the kinds of data footprints enterprises will increasingly need to manage in production. High performance is not presented as an isolated peak number, but as throughput that remains strong under a demanding real-world vector search workload while preserving RAID5 protection. Resource efficiency also becomes part of the business case, because a modular server-based architecture with onboard compute and storage helps organizations do more with existing power, space, and cooling budgets while improving utilization of costly AI infrastructure.

Taken together, these outcomes translate into clear business value. Enterprise buyers gain lower operational risk, better use of capital-intensive compute and storage resources, and greater confidence that AI deployments will perform reliably outside the lab. Just as important, they gain assurance that scaling AI does not require sacrificing resilience or overspending on infrastructure to compensate for bottlenecks elsewhere in the stack. The result is a more practical path to production AI: faster time to value, stronger service-level confidence, improved infrastructure efficiency, and a platform foundation that can support growth without compromising availability, trust, or return on investment.

Acknowledgments

The author would like to thank Dell Technologies for providing access to the PowerEdge R770 hardware platform and technical expertise. Special recognition goes to the engineering teams at Broadcom for their collaboration on storage optimization strategies.

Important Information About this Report

CONTRIBUTORS

Brian Martin
AI Data Center Performance | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is a leading research organization specializing in enterprise AI infrastructure optimization and deployment strategies. Our lab focuses on evaluating and optimizing AI hardware and software solutions for real-world enterprise applications, with particular expertise in large language models, retrieval-augmented generation systems, and distributed AI architectures.

For more information, visit signal65.com
or contact research@signal65.com

signal65

IN PARTNERSHIP WITH

DELL Technologies

signal65

CONTACT INFORMATION
Signal65 I signal65.com