signal65

EXECUTIVE SUMMARY

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

Overview

Enterprise AI has moved from experimentation into operational deployment, where success depends on more than model accuracy alone. For retrieval-augmented generation, vector search, and real-time inference, infrastructure must deliver data to models quickly, consistently, and reliably at production scale. When storage cannot keep pace, user-facing latency rises, GPU and CPU resources sit underutilized, and the business value of AI investments is reduced.

Earlier Signal65 work showed that the Dell PERC H975i (PERC13) controller can reach roughly 56 GB/s of sequential throughput and 13 million IOPS in synthetic testing, establishing the platform's technical ceiling. The more important question for enterprise buyers is whether that ceiling is reachable 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.

Key Highlights

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

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

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

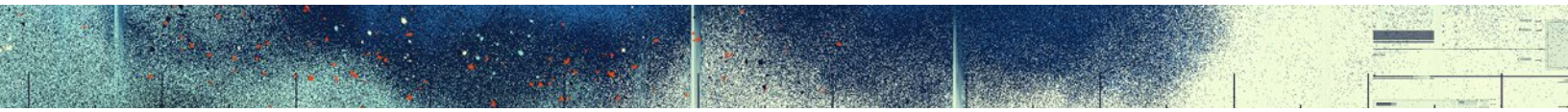
The result is clear: properly configured local NVMe storage can sustain production-scale vector search performance while preserving resilience. The tested system delivered up to 51.6 GB/s of sustained throughput and 860 queries per second from a storage-based FAISS index, while maintaining RAID5 protection. During RAID failure and rebuild conditions, performance degradation remained modest, demonstrating that high-performance AI infrastructure does not require sacrificing data protection.

Why This Matters for Enterprise AI

AI infrastructure is often evaluated through the lens of compute: GPUs, accelerators, model size, and inference throughput. Those factors matter, but they do not operate in isolation. Retrieval-intensive AI workloads also depend on the ability to move large volumes of data quickly, respond with predictable latency, sustain high I/O concurrency, and preserve availability when components fail.

At smaller scales, memory and caching can conceal storage limitations. At 10-billion-vector scale, those buffers are no longer enough. The true behavior of the storage architecture becomes visible. In this study, the index was sized beyond system memory and tested using direct I/O to ensure results reflected real storage performance rather than DRAM-assisted caching effects.

The findings show that storage is not a secondary consideration in enterprise AI. It is a primary component of the inference pipeline. For RAG, vector search, and other data-intensive workloads, storage performance directly affects end-user responsiveness, service-level confidence, infrastructure efficiency, and total cost of ownership.

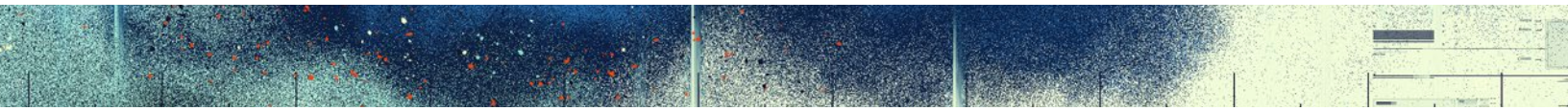


Strong Failure-Mode Performance

The key production question is not whether the platform can deliver peak performance in a healthy state, but whether it can continue serving AI workloads when the storage system is recovering. In RAID5 rebuild testing, the Dell PowerEdge R770 with Dell PERC H975i controllers continued to sustain strong vector-search performance with only modest impact to throughput and query rate.

Operating Point	Healthy System	During RAID5 Rebuild	Impact
Peak Throughput	51.6 GB/s	43.3 GB/s	~16% lower
Peak Query Rate	860 QPS	791 QPS	~8% lower
Peak-QPS p50 Latency	204 ms	210 ms	~3% higher
Peak-QPS p99 Latency	231 ms	358 ms	Modestly wider

This result is important for enterprise AI because rebuild events are not theoretical. Drives fail, arrays rebuild, and production workloads still need to run. The platform's ability to maintain more than 43 GB/s during rebuild and nearly 800 queries per second at the peak-QPS operating point shows that protected local NVMe storage can deliver performance and resilience together, reducing the traditional tradeoff between speed and availability.



The Bottom Line

With the right architecture and proper tuning, organizations can accelerate retrieval-intensive AI workloads, protect critical data, and improve infrastructure efficiency while reducing deployment risk. The R770 / PERC H975i platform delivered solid performance on a realistic 10-billion-vector workload, sustained that performance under a degraded RAID5 rebuild, and did so with modest memory overhead. For enterprise buyers, that translates into faster time to value, stronger service-level confidence, and a foundation that can scale without compromising availability or trust.

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

View the full report on signal65.com: Local Vector Search at 10 Billion Scale

signal65

CONTACT INFORMATION
Signal65 I signal65.com