signal65

June 2026
Dell Technologies

# Local Vector Search at 10 Billion Scale

## AI Performance on Dell PowerEdge R770 with Dell PERC H975i

Enterprise AI depends on storage infrastructure that delivers throughput, responsiveness, and resilience at scale. This study shows how a Dell PowerEdge R770 with dual PERC H975i controllers sustains high performance for retrieval-augmented generation, vector search, and real-time inference — without compromising data protection.

### 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.


Dell PowerEdge R770 platform

## Query Performance with and without Rebuild

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.

| Metric | 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 | **Modest** |

**Protected local NVMe storage can deliver production-scale AI retrieval performance without forcing a tradeoff between speed and resilience.**

Source: Signal65 Lab Insights — Local Vector Search at 10-Billion Scale

signal65.com