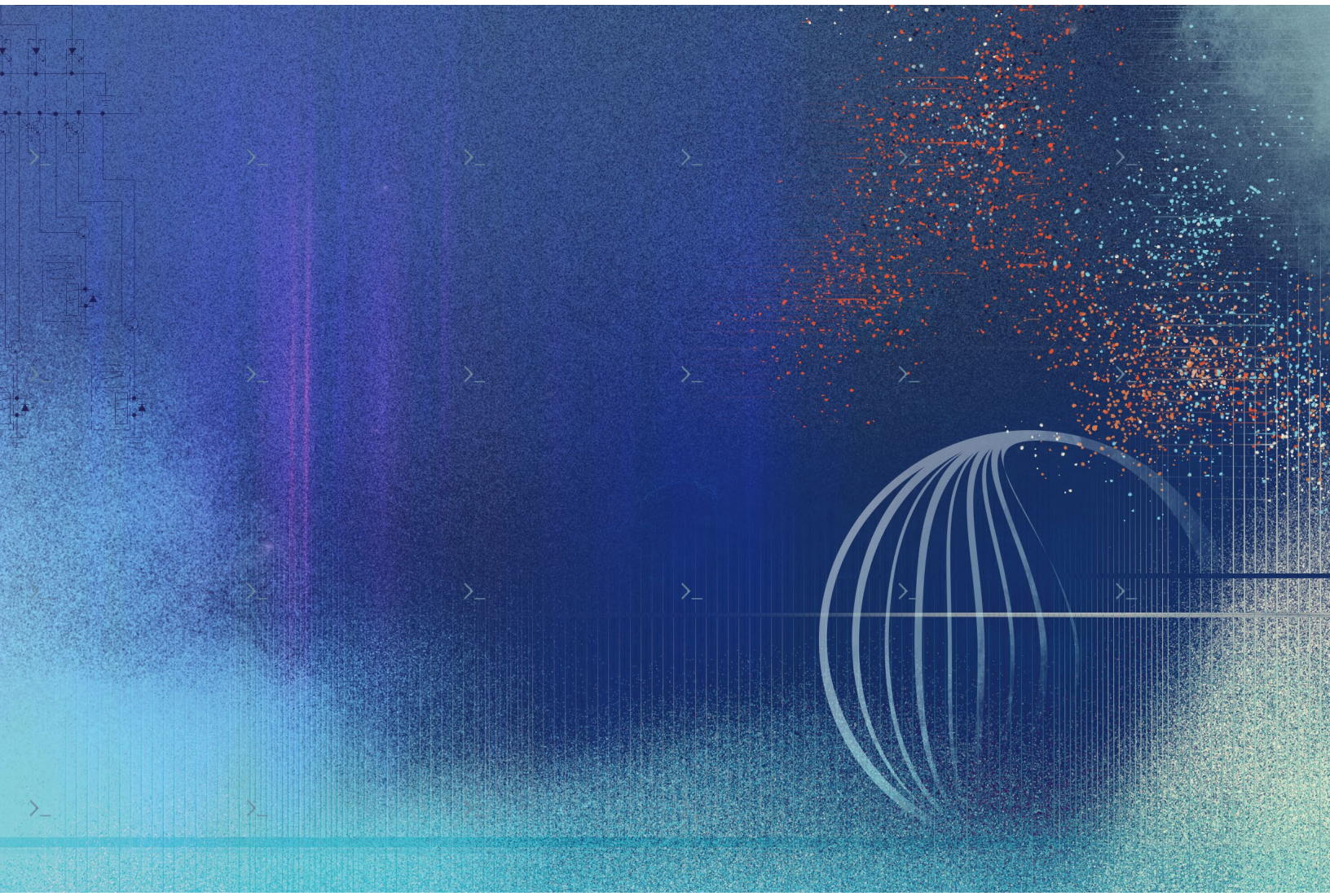
signal65



AI Storage Pipeline Acceleration with Dell PERC H975i (PERC13)

AUTHOR

Brian Martin
AI and Data Center Lead | Signal65

IN PARTNERSHIP WITH

DELL Technologies

MAY 2025

Executive Summary

AI Storage Challenges

As organizations increasingly adopt AI technologies, particularly large language models and generative AI, they face unprecedented demands on their storage infrastructure. AI workloads—particularly those involving fine tuning, inference, and vector operations—often require large datasets to be transferred, processed, and stored with minimal latency and consistently high bandwidth.

The scale of these demands is substantial. A typical enterprise AI deployment now requires storage systems capable of delivering sustained bandwidth exceeding 50 GB/s for efficient training, and tens of millions of small-block random read operations for inference serving. Additionally, retrieval-augmented generation (RAG) solutions demand storage systems that can manage both high-velocity vector database operations and rapid document retrieval.

Dell PERC13 Solution

The Dell PERC13 addresses these challenges through innovative architecture designed specifically for evolving AI workloads. Key RAID5 performance improvements include:

Breakthrough IOPs	Optimized Bandwidth	Exceptional Rebuild	Generational Improvement
Throughput exceeding 12.9M random read IOPS and 5M random write IOPS with an average 8uS response time.	Bandwidth of 56GB/sec for sequential read operations and 50GB/sec for sequential writes.	Throughput exceeding 10M random read IOPS and 2M random write IOPS during rebuild at 22 minutes per TB.	Demonstrating up to 20 times more write IOPS than PERC11 and 5.5 times more than PERC12.

This dramatic performance improvement enables organizations to fully utilize GPU investments for AI operations while maintaining data resiliency.

Highlights and Value Proposition

The Dell PERC13 controller represents a significant advancement in local storage acceleration, particularly for AI workloads that benefit from high-performance direct-attached storage. Building on Dell's extensive experience with enterprise storage, PERC13 delivers unprecedented performance that aligns precisely with the demands of modern AI operations. Local RAID configurations with PERC13 address several critical needs for AI workloads running on single-server or edge deployments. First, PERC13's enhanced controller architecture delivers the raw IOPS, bandwidth, and low-latency responsiveness necessary for data-hungry training and inference tasks. Second, robust RAID options including RAID10 for maximum throughput or RAID5 for peak read and strong write throughput provide the flexibility to tailor storage to workload needs while maintaining critical resilience. Finally, offloading RAID processing and rebuild tasks to dedicated hardware reduces CPU overhead and ensures system resources remain focused on AI operations. Together, these features make PERC13 an optimal solution for environments that require high-performance local storage.

The Evolution of Storage Requirements in AI

Current Landscape of AI Workloads

The artificial intelligence landscape has evolved dramatically with the emergence of large language models (LLMs) and generative AI, fundamentally transforming storage requirements for enterprise infrastructure. Today's AI workloads can be categorized into four primary patterns, each placing distinct demands on storage systems.

Model Training/Fine-Tuning represents the three most storage-intensive workloads. Organizations increasingly perform transfer learning and fine-tuning of foundation models, requiring high-bandwidth sequential access to massive datasets. These operations typically involve reading hundreds of terabytes of training data repeatedly throughout the training cycle, often from local NVMe drives. The storage system must maintain sustained read throughput of 40-50 GB/s to keep pace with modern GPU clusters, while simultaneously supporting the periodic checkpointing of model weights that generate large sequential writes.

Model Training and Fine Tuning also include writing and reading checkpoints, the next two bandwidth intensive workloads. At regular intervals, training state is saved (a “checkpoint”) to avoid losing all previous effort should an error occur and interrupt training. All GPUs stop processing during this write, so it is critical it be completed as quickly as possible. Likewise, restoring a checkpoint to continue processing requires reading back the data as fast as possible. These high bandwidth write and read operations can consume up to 1GB/sec per GPU.

Inference Serving and RAG present a contrasting I/O pattern characterized by high volumes of small random reads. As organizations deploy AI models into production, storage systems must handle thousands of concurrent requests, each requiring rapid access to model weights and associated data. These operations demand exceptional IOPS performance and consistent low latency, typically requiring millions of low latency IOPS to prevent inference bottlenecks.

Data Preparation

Ingesting: High BW Write

Cleaning: High BW Read, High IOPs Write

Model Training

Feeding GPUs: High BW and High IOPs Reads

Checkpointing: High BW Write and Read

Model Inference

Loading Models: High BW Read

Vector Database: High IOPs Read

Storage Bottlenecks in AI Pipelines

Storage bottlenecks present significant challenges in AI pipelines, directly impacting the efficiency and effectiveness of AI operations. Understanding these challenges and their performance implications is crucial for designing effective AI infrastructure.

Modern AI workflows face several distinct storage-related bottlenecks. During training operations, insufficient storage bandwidth can leave expensive GPU resources underutilized, extending training times and increasing costs. A single Gen5 PCIe GPU can process data at rates approaching 58 GB/s; when storage systems cannot maintain this throughput, GPUs spend valuable cycles waiting for data, significantly reducing training efficiency.

In inference scenarios, storage bottlenecks manifest differently. Model serving often requires rapid access to weights and associated data across thousands of concurrent requests. When storage systems cannot deliver sufficient IOPS or maintain consistent low latency, inference response times become unpredictable, potentially violating service level agreements (SLAs) and degrading user experience.

Critical Performance Factors

Three key metrics determine storage system effectiveness for AI workloads. Each plays a crucial role in different aspects of AI operations:

Bandwidth becomes paramount during training operations and large-scale data preparation. Modern AI training pipelines require sustained read bandwidths of 40-50 GB/s to keep pace with GPU processing capabilities. This requirement increases proportionally with the number of GPUs being utilized, making bandwidth a critical scaling factor for training operations.

IOPS performance primarily impacts inference and RAG operations. Production AI systems often need to support thousands of concurrent users, each generating multiple storage requests. Systems must deliver millions of IOPS with consistent performance to maintain responsive AI applications. For RAG implementations specifically, high IOPS capabilities support rapid vector database operations and concurrent document retrieval.

Latency affects the responsiveness and reliability of AI applications. While bandwidth and IOPS measure maximum throughput, latency determines how quickly individual requests are serviced. AI operations require consistent low latency, frequently under 1 millisecond, to maintain acceptable performance. Variable or high latency can create unpredictable application behavior and poor user experience.

Understanding AI Storage Workload Patterns

Modern AI operations encompass distinct workflows, each generating unique storage access patterns and performance requirements. Understanding these patterns is essential for optimizing storage infrastructure and ensuring efficient AI operations.[1]

Training Data Ingestion

Training data ingestion represents one of the most demanding storage workflows in AI operations. This process involves reading extensive datasets repeatedly during training cycles, often processing hundreds of terabytes of data. The storage pattern is characterized by large sequential reads, typically accessing data in blocks of 1MB or larger. Additionally, data ingestion frequently requires random access to multiple files simultaneously to support data augmentation and shuffling operations.

Modern training pipelines require storage systems capable of delivering sustained read bandwidths of 40-50 GB/s to maintain GPU utilization. These operations often exhibit a bursty nature, with periods of intense I/O activity followed by computational phases. Storage systems must maintain consistent performance during peak periods while supporting concurrent write operations for logging and checkpointing.

Fine-tuning Operations

Fine-tuning workflows present a more nuanced storage pattern than initial training. While still requiring high-bandwidth sequential reads, fine-tuning operations typically work with smaller datasets but demand more frequent model checkpointing. This creates an interleaved pattern of sustained reads and periodic large sequential writes.

Storage requirements for fine-tuning include both high read bandwidth (30-40 GB/s) and write performance capable of handling checkpoint operations without disrupting the read pipeline. The system must maintain low latency during these mixed operations to prevent training interruptions. Typical checkpoint sizes range from several gigabytes to hundreds of gigabytes, requiring write bandwidth of at least 10 GB/s to minimize checkpoint overhead.

Inference Serving

AI inferencing, the process of using a trained model to make predictions on new data, places distinct demands on storage performance, primarily centered around low latency and high read IOPS. Unlike AI training, which often involves sequentially processing massive datasets and benefits from high throughput, inferencing frequently deals with smaller, individual data points or batches requiring rapid responses. The core requirement is minimizing the time it takes to load the necessary model parameters and input data to generate a prediction. This makes low-latency storage crucial, especially for real-time applications like fraud detection or recommendation engines.

Furthermore, storage performance is crucial during initialization and for operational flexibility. Even if a specific model fits comfortably within GPU memory (VRAM) once loaded, it must first be transferred from storage to the GPU(s). Additionally, inference environments may serve requests for multiple different models. If available VRAM is insufficient to hold all currently active models, the system needs to dynamically swap models in and out of GPU memory from storage. Reliable high-performance storage significantly reduces model load time.

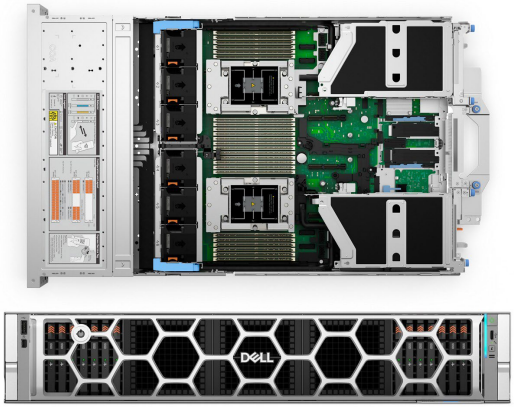
RAG Vector Operations

Retrieval-Augmented Generation introduces complex storage patterns combining aspects of both inference and data retrieval workflows. Vector operations involve frequent small random reads and writes as the system updates and queries vector databases. Simultaneously, the system must support efficient document retrieval, creating a mixed workload of random and sequential access patterns.

These operations require storage systems capable of delivering both high IOPS (3-5 million) for vector operations and sustained bandwidth (20-30 GB/s) for document retrieval. The storage system must maintain consistent performance across these different access patterns while supporting the high concurrency typical of production RAG implementations.

Performance Analysis

The PERC13 controller supports multiple RAID levels, each balancing performance, resilience, and usable capacity differently. We explore the four most relevant RAID types—RAID 0, RAID 5, RAID 10, and RAID 6 using the FIO performance benchmarks tool to drive representative workloads. The test system is a Dell PowerEdge R7725 with PERC13 and 16 Dell 3TB mixed workload NVMe drives configured as four volumes with four drives each or two volumes with eight drives each. All PERC13 test results are from Signal65 testing.



Dell PowerEdge R7725 Rack Server

PERC13 Performance

The PERC13 controller delivers breakthrough performance across a range of RAID configurations, balancing throughput, fault tolerance, and latency to meet the diverse needs of AI workloads. The table below summarizes performance metrics for RAID 0, 5, 10, and 6, providing a snapshot of how each level performs under common scenarios. These results highlight the exceptional IOPS, bandwidth, and rebuild efficiency made possible by the PERC13 advanced architecture. The following sections explore each RAID level in more detail, outlining where each configuration excels and how organizations can best match RAID strategy to specific AI pipeline demands. Small block random write performance in this paper is limited by the configuration of 16 NVMe drives; systems supporting additional drives will benefit from even higher performance.

Metric	Definition	Units	R0	R10	R5	R6
Read Bandwidth	Storage bandwidth for 100% 64KB sequential read	GB/s	56	56	56	56
Write Bandwidth	Storage bandwidth for 100% 64KB sequential write	GB/s	54	45	50	40
Read IOPs	Random 4KB Read Operations per second	IOPs	13M	13M	13M	13M
Write IOPs (limited by drive count of 16)	Random 4KB Write Operations per second	IOPs	10M	5M	2.9M	2M
Write Latency	Average time to complete a storage operation up to 75% of maximum IOPS	us	8	8	8	8
Perf Under Rebuild	Storage Subsystem performance during Rebuild (100% RR)	IOPs	n/a	10M	10M	9M
Rebuild Under Load	Minutes to rebuild failed device in RAID array	Min/TB	n/a	30	31	45

RAID 0: Maximum Performance, Zero Redundancy

What it is:

RAID 0 stripes data evenly across all drives without any redundancy. It offers the **highest performance possible** by utilizing the full bandwidth and IOPS of all disks simultaneously. However, it provides **no fault tolerance**—if a single drive fails, all data is lost.

Why a customer might choose it:

RAID 0 may appeal to customers with **non-critical, high-throughput workloads** such as short-lived scratch data, temporary caches, or intermediate datasets that can be regenerated easily. In AI, this might include ephemeral preprocessing buffers or staging areas where speed matters more than durability.

Performance Highlights:

- **4K Random Read:** Up to **13M IOPS**, with average latency under 0.3mS
- **4K Random Write:** Up to **10M IOPS**, with average latency of 8µS
- **64K Sequential Read/Write:** 56 GB/s read, 54 GB/s write

RAID 0 sets the bar for pure throughput, but its lack of protection makes it suitable only for transient data.

RAID 5: Best All-Around Value

What it is:

RAID 5 stripes data with a single parity block distributed across drives. It can tolerate one drive failure and offers a strong balance of read/write performance, protection, and usable capacity (N-1 of N drives).

Why a customer might choose it:

Ideal for AI training, inference, and RAG workloads where read performance is critical, write speeds are important but not dominant, and some level of fault tolerance is required. Customers choose RAID 5 when they want to maximize usable storage capacity without compromising performance—especially for single-server AI nodes that cannot rely on network-based redundancy.

Performance Highlights:

- **4K Random Read: 13M IOPS,** matching RAID 0 with average latency under 0.5mS
- **4K Random Write: 2.9M IOPS,** with average latency of 8µS
- **64K Sequential Read/Write:** Up to **56 GB/s read, 50 GB/s write**
- **Rebuild Under Load:** 35min/TB sustaining **10M IOPS**

RAID 5 offers near-RAID 0 performance for reads and significantly better protection—making it the smart choice for most AI use cases.

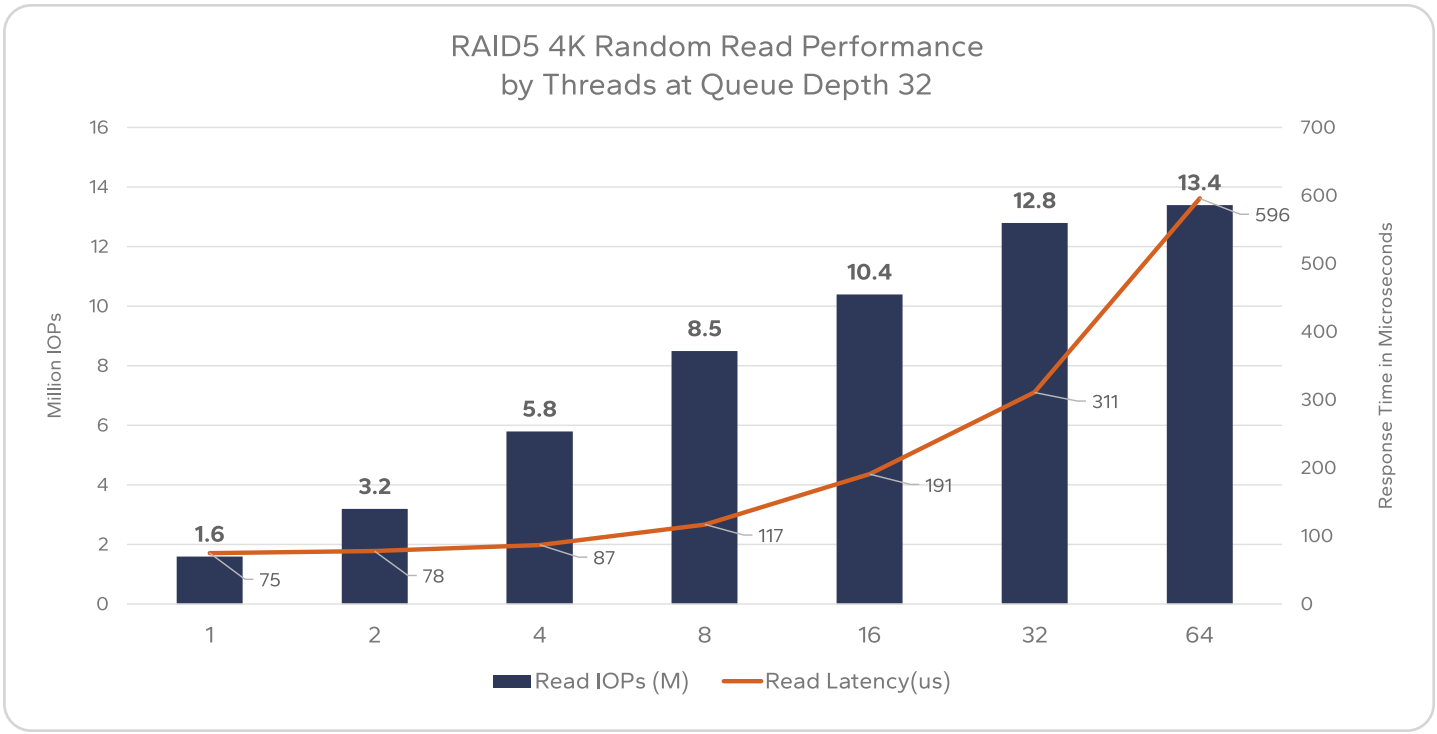


Figure 1: *RAID 5 Random Read Performance*

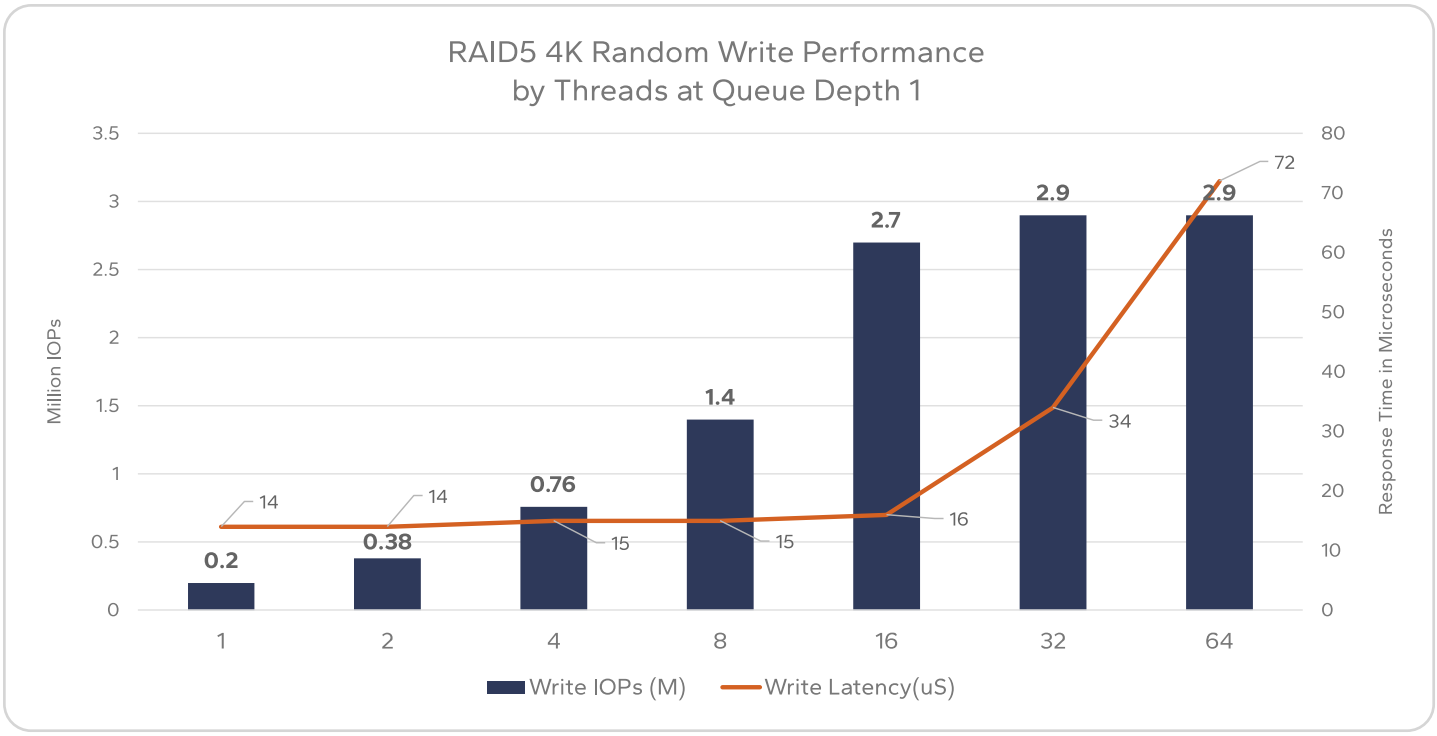


Figure 2: *RAID 5 Random Write Performance with Average Latency*

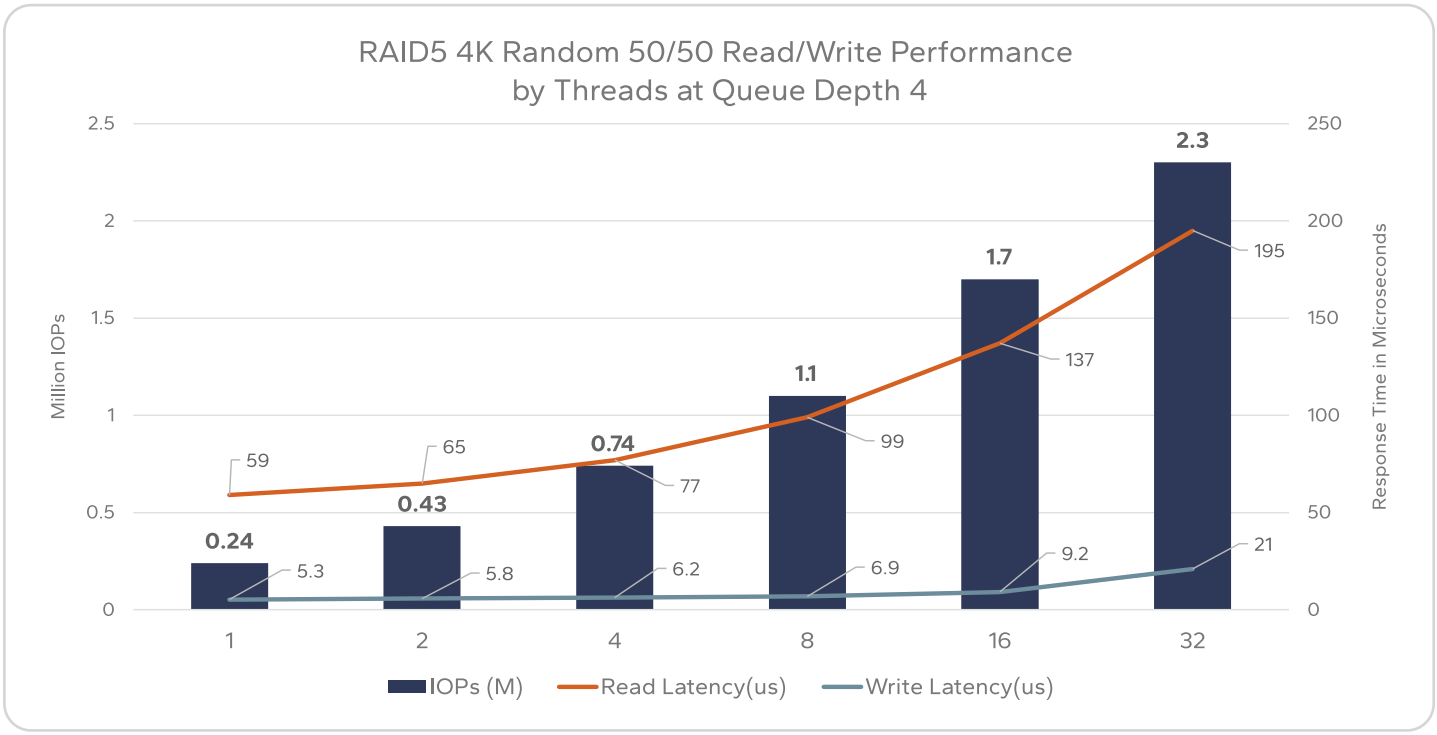


Figure 3: *RAID 5 Random 50/50 Read/Write Performance*

RAID 10: Premium Write Performance and Fast Recovery

What it is:

RAID 10 mirrors each drive (RAID 1) and stripes data across mirrored pairs (RAID 0). This allows for fast reads, better write performance than parity-based RAID, and high resilience.

Why a customer might choose it:

Best suited for AI inference workloads with heavy write demands, or situations where fast recovery time is a priority. RAID 10 provides low latency, high throughput, and excellent rebuild speeds, but at the cost of 50% usable capacity. Customers with smaller working sets or higher budgets may opt for RAID 10 for predictability and simplicity.

Performance Highlights:

- **4K Random Read:** Up to **13M IOPS** with average latency under 0.2mS
- **4K Random Write: 5.2M IOPS,** double that of RAID 5
- **64K Sequential Read/Write: 56 GB/s read, 45 GB/s write**
- **Rebuild Under Load: : ~93–96 minutes,** ~32 min/TB

RAID 10 shines in workloads where write latency, IOPS, and quick recovery are critical—albeit with a capacity trade-off.

RAID 6: Extra Redundancy, Moderate Penalty

What it is:

RAID 6 uses **dual parity**, allowing two simultaneous drive failures. It provides better fault tolerance than RAID 5 but incurs more overhead, especially on writes.

Why a customer might choose it:

RAID 6 is best for customers who prioritize **data durability** over write speed—such as deployments in **remote, unattended, or harsh environments** where recovery may be delayed. It's a fit for **archival AI models, RAG knowledge bases,** or other AI pipelines that involve large datasets and **require higher fault tolerance**.

Performance Highlights:

- **4K Random Read:** Matches RAID 5 at **13M IOPS,** ~3µs latency
- **4K Random Write:** Drops to **~2M IOPS,** latency ~8µs
- **64K Sequential Read/Write: 56 GB/s read, 40 GB/s write**
- **Rebuild Under Load: 123–130 minutes,** ~43–45 min/TB, performance sustained at **~8.7M IOPS**

RAID 6 is a conservative but reliable option when dual-drive fault tolerance is worth the performance tradeoff.

Generational Improvement (RAID 5)

PERC 13 demonstrates superior performance across all metrics compared to PERC 12 and PERC 11 as shown below:

RAID 5 read IOPS on the PERC13 controller consistently reach 13 million across test scenarios, matching the throughput of RAID 0 while retaining fault tolerance. This demonstrates the controller's ability to deliver high random read performance critical to inference and RAG workloads without sacrificing data protection.

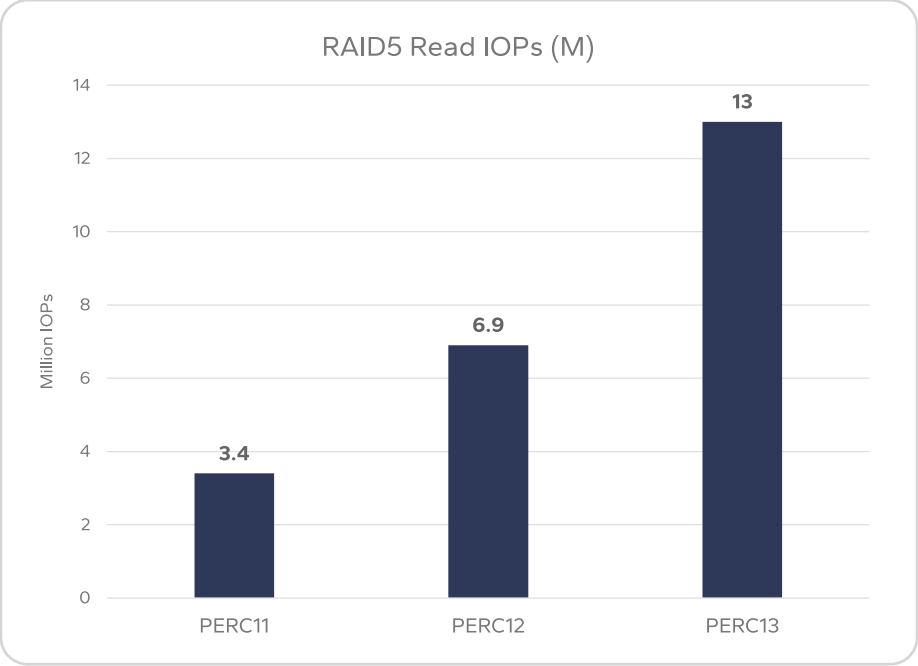


Figure 4: *RAID 5 Read IOPs in Millions*

Write performance under RAID 5 reaches up to 2.9 million IOPS with 16 drives and scales over 5 million with 32 drives, showcasing the PERC13's robust write engine. This level of performance enables responsive vector updates.

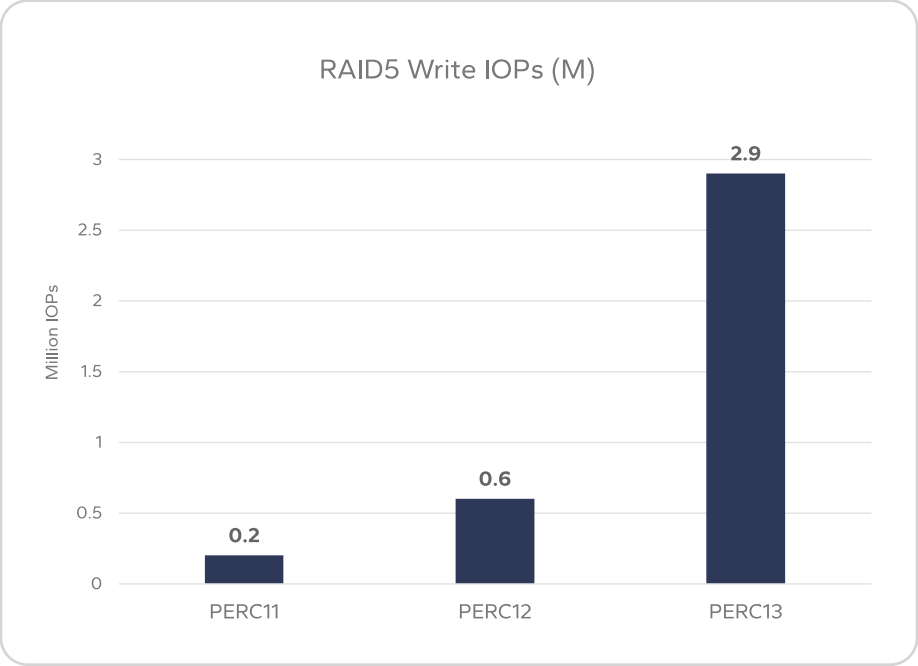


Figure 5: *RAID 5 Write IOPs in Millions*

Sequential throughput for RAID 5 peaks at 56 GB/s read and 50 GB/s write, demonstrating that parity overhead does not limit performance for modern AI training and fine-tuning operations. This level of bandwidth ensures the PERC13 can support model checkpointing and model loading.

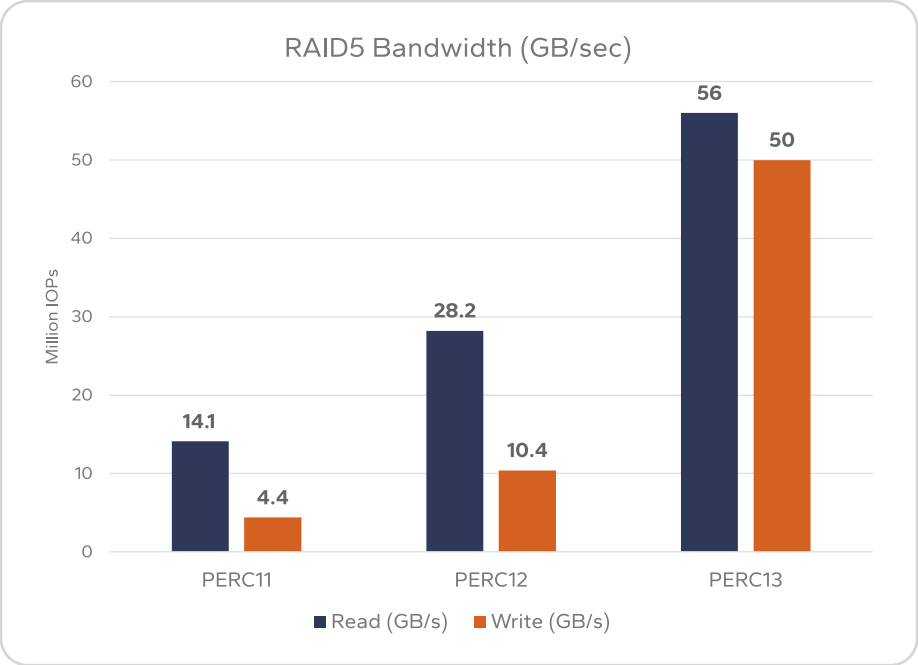


Figure 6: *RAID 5 Bandwidth (GB/sec)*

The following chart illustrates the rebuild efficiency of RAID 5 under load, with the PERC13 sustaining up to 10 million IOPS during the rebuild process. The rebuild rate of approximately 31 minutes per terabyte minimizes downtime and reduces vulnerability windows, making RAID 5 a practical choice for AI environments that prioritize resilience without compromising active performance.

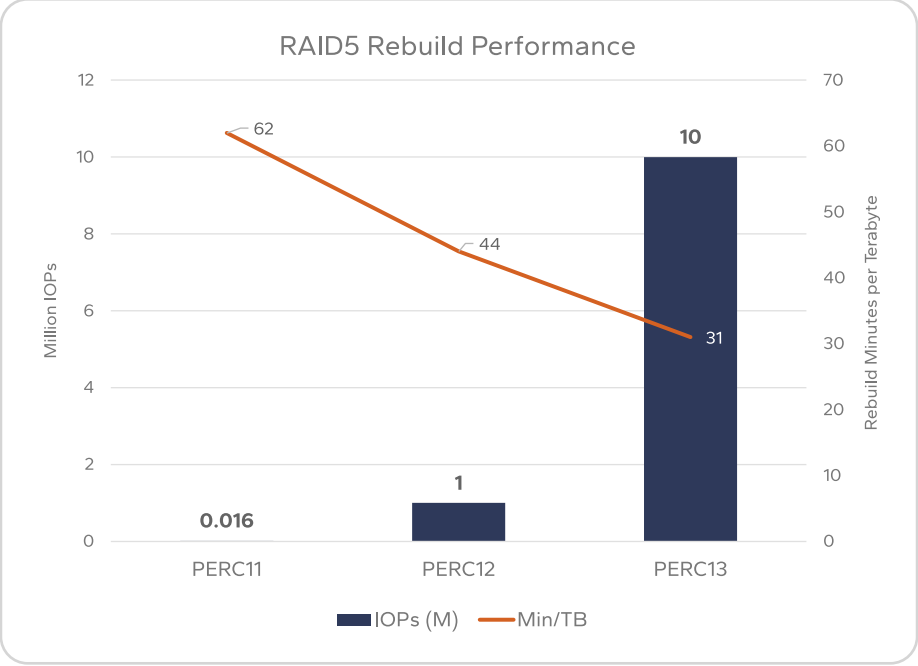


Figure 7: *RAID 5 Rebuild Performance*

RAID5 Performance Comparison

Metric	Definition	Units	PERC11	PERC12	PERC13
Read Bandwidth	Storage bandwidth for 100% read	GB/s	13	27	**56**
Write Bandwidth	Storage System Data Rate	GB/s	4	10	**50**
Read IOPs	Storage Read Operations per second	IOPs	3.4M	6.9M	**13M**
Write IOPs	Storage Write Operations per second	IOPs	240K	650K	**2.9M**
Write Latency	Average time to complete a storage operation up to 75% of maximum IOPS	us	>200	8	**8**
Perf Under Rebuild	Storage Subsystem performance during Re-build (100% RR)	IOPs	16.5K	1.0M	**9.8M**
Rebuild Under Load	Minutes to rebuild failed device in RAID array	Min/TB	90	35	**31**

Note: *PERC11 and PERC12 performance numbers from*
https://infohub.delltechnologies.com/it-it/p/perc-12-generational-performance-boosts/

Conclusion and Future Outlook

Dell PERC13 delivers key benefits tailored for demanding AI workloads by providing breakthrough storage performance and enhanced operational efficiency. It achieves this through dramatically improved RAID5 IOPS, reaching 13M random 4KB reads and 5M random 4KB writes, and high sequential RAID5 bandwidth up to 56 GB/s read and 50 GB/s write. The controller offers exceptional performance even during rebuilds, minimizing vulnerability windows, and provides flexibility through robust RAID options like RAID 5 for balanced performance and capacity or RAID 10 for maximum write throughput, catering to diverse AI pipeline needs such as training, inference, and RAG. PERC13 represents a significant generational leap in performance and efficiency.

Looking ahead, Dell PERC13 is positioned to play a vital role, particularly in edge and single-server AI deployments. As AI models continue to grow in complexity and inference tasks become more distributed, the need for localized low-latency, high-IOPS storage solutions like PERC13 will intensify. The ability to deliver significant bandwidth and IOPS directly to the compute and GPU resources makes it well-suited for accelerating fine-tuning, demanding inference workloads, and RAG operations locally. By maximizing the performance potential of NVMe drives while ensuring data protection, PERC13 provides a robust foundation for future AI infrastructure.

Appendix

References:

[1] https://www.snia.org/sites/default/files/SSSI/CMSS24/CMSS24-Cardente-Storage-Requirements-for-AI.pdf

Testing Configurations

Testing performed on Dell PowerEdge R7725 with two AMD EPYC 9755 128-Core Processors, 768GB 6400MT/s registered DDR5, Dell PERC H975i RAID controllers, and 16 Dell 3.2TB Data Center NVMe Mixed Use E3s Gen5 drives. Drives were configured as four volumes of four drives each.

Important Information About this Report

CONTRIBUTORS

Brian Martin
AI and Data Center Lead | 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.

IN PARTNERSHIP WITH

DELL Technologies

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION
Signal65 I signal65.com