Graid Technology Inc.

Data Speed & Resiliency Means Better AI Applications

RackLive AI Forum Presentation

Jason Harmon

September 2025

INTRODUCING

Graid Technology Inc.

— We Invented the Future of Storage — Now We’re Powering What Comes Next

- Creator of award-winning SupremeRAID™, the world’s first and fastest GPU-based NVMe RAID
- Eliminates traditional RAID bottlenecks to unlock full SSD performance
- Frees CPU resources by offloading RAID operations to the GPU
- Trusted by leading partners across AI, HPC, and enterprise infrastructure

HQ in Silicon Valley

Global R&D in Taiwan

Global Network of Partners, OEMs, Distributors & Resellers

> **“For the very first time, your storage system will be GPU-accelerated.”**
>
> JENSEN HUANG,
> NVIDIA CEO
> GTC 2025

AI Data Pipeline

FACTS

Winners and losers will be determined by who can truly harness the value of their data.

90%

OF THE WORLD'S DATA WAS GENERATED IN THE LAST 2 YEARS

THE VOLUME OF DATA STORAGE GLOBALLY IS DOUBLING EVERY 4 YEARS

2x

30%

OF C-SUITE EXECS CITE SLOW DATA INGESTION AS A CONCERN FOR BIG DATA ANALYTICS

UP TO **30%** OF ENTERPRISE IT BUDGETS ARE CONSUMED BY DATA STORAGE, BACKUP, AND DISASTER RECOVERY

49%

OF EXECUTIVES SAY THE CURRENT DATA SOLUTIONS AREN'T FLEXIBLE ENOUGH

AI SYSTEMS CAN SIT IDLE WAITING FOR DATA, FOR AS MUCH AS

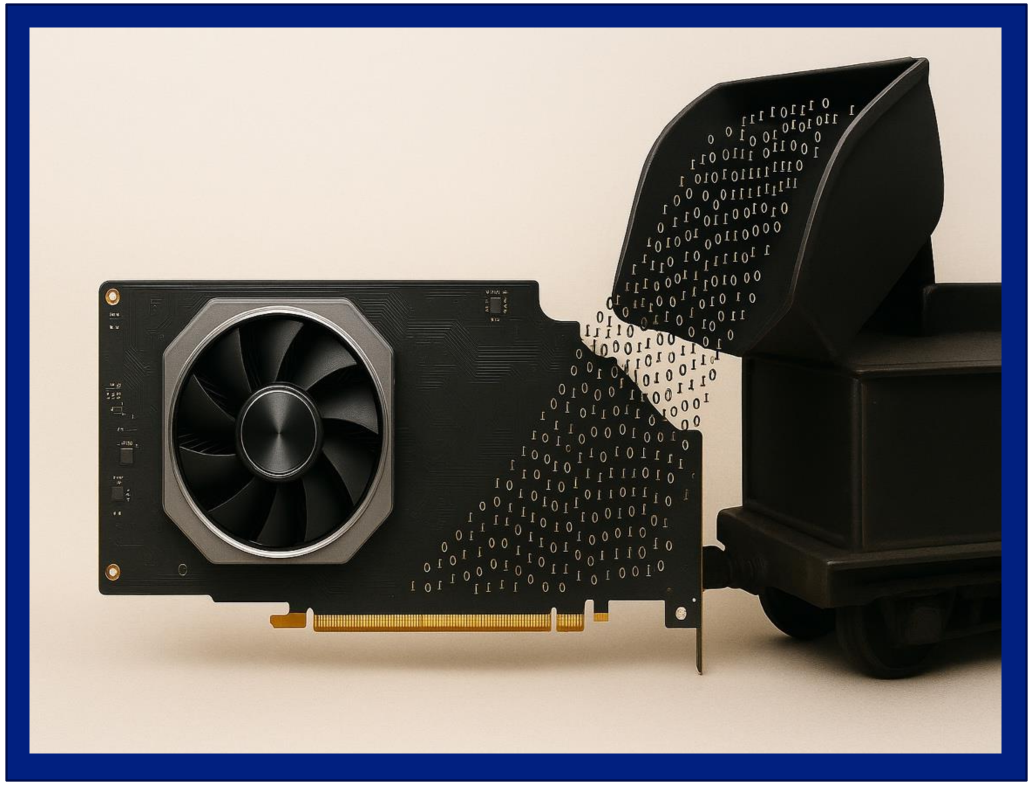
50%

OF THE TIME

The AI Storage Challenge

How Do We:

- Feed the GPU enough data to fully utilize.
- Free up CPU resources for workloads.
  - Data preprocessing
  - Retrieval Augmented Generation (RAG)
  - Natural Language Processing (NLP) models
- Ensure your critical datasets and models are protected and are reliable.



These are all TCO Discussions!!!

AI Application ROI Requirements

AI application ROI is highly dependent on both data speed and resiliency at every stage of the data pipeline.

	DESCRIPTION:	RESULT:
PERFORMANCE	High-speed storage and networking maximize AI efficiency, enabling constant processing of massive datasets.	Bottlenecks from slow read/write or latency waste expensive compute cycles, delaying both training and inference.
+		
RESILIENCE	A resilient data pipeline ensures integrity, availability, and rapid recovery, keeping AI workloads running smoothly.	Advanced techniques—scalable indexing, failover, RAID, checkpointing—maintain consistent throughput and minimize retraining overhead.
=		
RETURN ON INVESTMENT	Eliminating delays reduces computational costs and accelerates time-to-insight, directly improving ROI.	Resiliency lowers TCO by preventing data loss, avoiding downtime, and enhancing customer experience, driving faster returns.

SW RAID can burden your CPU

High CPU overhead:

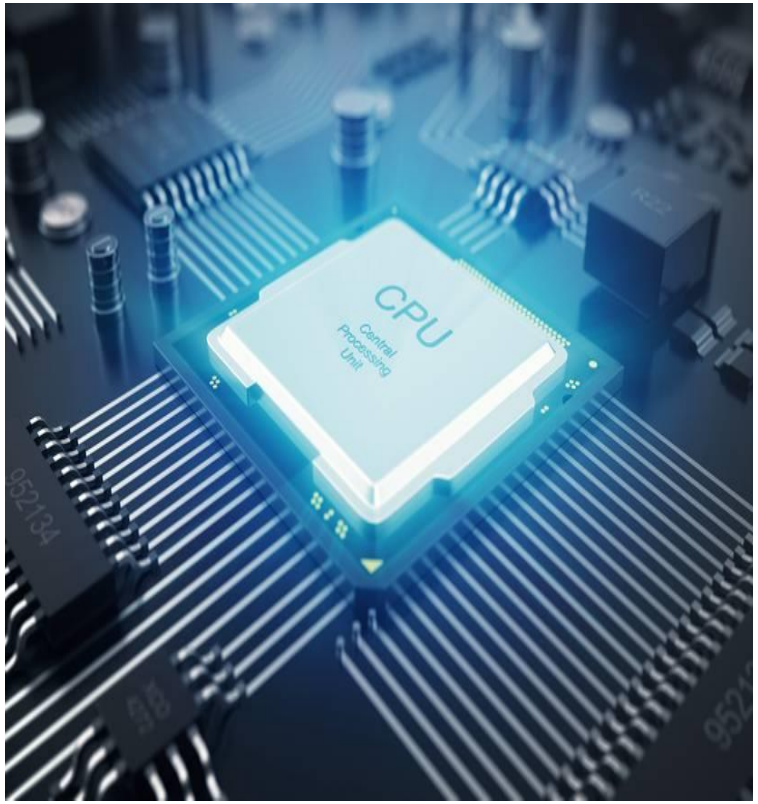
Software RAID uses CPU resources your application could be using.

I/O bottlenecks:

Software RAID cannot deliver the IOPS of modern NVMe SSDs.

Degraded performance under load:

High IO loads can task the CPU or delay critical tasks.



AI Data Pipeline - Workloads and Needs

AI workloads are demanding, and the hardware is too expensive to sit idle or underutilized.

Data Preparation

Throughput	High
IOPS	Moderate
Latency	Moderate
Cost vs Perf	Balance/ Perf

Training/ Fine Tuning

Throughput	Very High
IOPS	Moderate
Latency	Low
Cost vs Perf	Performance

Inference

Throughput	Moderate
IOPS	Very High
Latency	Ultra-Low
Cost vs Perf	Balance

- Bottlenecks mean underutilized hardware (capex)
- Insufficient IOPS means each server will serve less users driving up costs
- High Latency causes a poor user experience and unpredictable application behavior
- All workloads are SSD dependent
- Performance is typically more critical overall than cost

Data Loss and Corruption Impact

- **Models are LARGE and growing** - data loss means long recovery times in an outage
- **Data errors can poison huge datasets** - meaning lost work and/or errors in results

Data Preparation

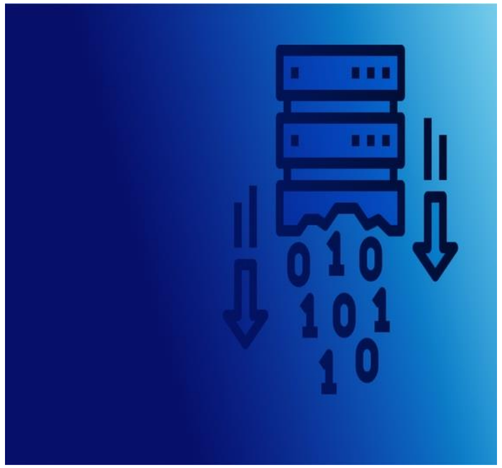
- Inference pipeline model loss will require a restore and redeployment
- Processed/ cleaned data loss means a restart of the ETL/ELT pipeline
- Metadata and label loss could mean hundreds of human hours to recreate

Training:

- Training data loss means a re-start of that training to at least the last checkpoint
- Trained model loss means a complete retrain effort

Inference

- Requests loss meaning poor user experience
- inference pipeline model loss will require a restore and redeployment



Cost

User Experience

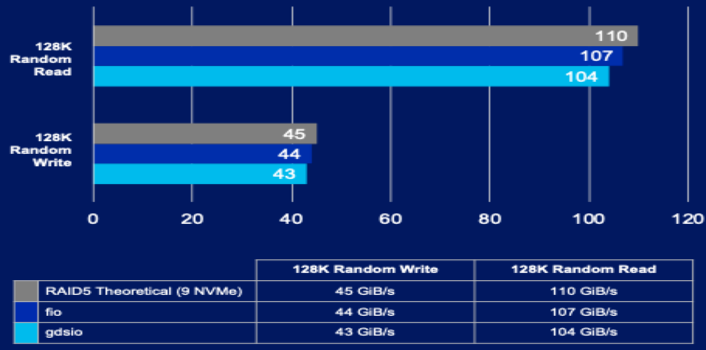
Time

SupremeRAID™ AE – Solving the AI Data Needs

SupremeRAID™ AE delivers IO performance that AI demands while also ensuring data resiliency and integrity.

Enabling Performance

- Near theoretical NVMe performance
- Leverages GPU Direct Storage (GPU) for strong performance getting data in and out of GPUs
- Prevents GPU Underutilization and idling by ensuring maximum data flow
- Offloads RAID off the CPU and onto a GPU
- Minimizes Latency and Maximizes IOPS



		128K Random Write	128K Random Read
■	RAID5 Theoretical (9 NVMe)	45 GiB/s	110 GiB/s
■	fio	44 GiB/s	107 GiB/s
■	gdsio	43 GiB/s	104 GiB/s

Results above are with 9 drives in RAID 5

Ensuring Data Resiliency and Integrity

- Protects the data from routine drive failures
- Mitigation of read errors preventing data corruption/ loss
- Automated and transparent bad block recovery
- End-to-end Data integrity preventing silent data corruption
- Continuous health monitoring of drives
- Fast Rebuild and recovery magnitudes better than HW RAID or SW RAID
- Non-Intrusive data flow protecting from controller failures

CUSTOMER SUCCESS STORY

University Accelerates Visualization Workflows By Removing Storage Bottlenecks With SupremeRAID™

- The µ-VIS X-Ray Imaging Centre at the University of Southampton needed a high-speed server for microfocus Computed Tomography to overcome bottlenecks in 3D visualization workflows.
- They built a custom server using a SupremeRAID™ solution, which replaces legacy RAID cards with GPUs for storage management.
- This setup provides high performance, low latency, and enhanced data redundancy, significantly improving efficiency for demanding workloads.
- The new server achieves up to 20GB/s local sequential write speeds -- four times faster than standard SSDs -- eliminating workflow delays and boosting overall performance.

https://business.novatech.co.uk/blog/differing-storage-needs-in-higher-education/

University of Southampton

4x

FASTER THAN STANDARD SSDS

5 GB/s → 20 GB/s

LOCAL SEQUENTIAL WRITES

Summary

Your AI needs performance, but it also needs reliable data.

AI has high storage performance requirements

- Large data sizes
- High IOPs demands
- Low Latency needs

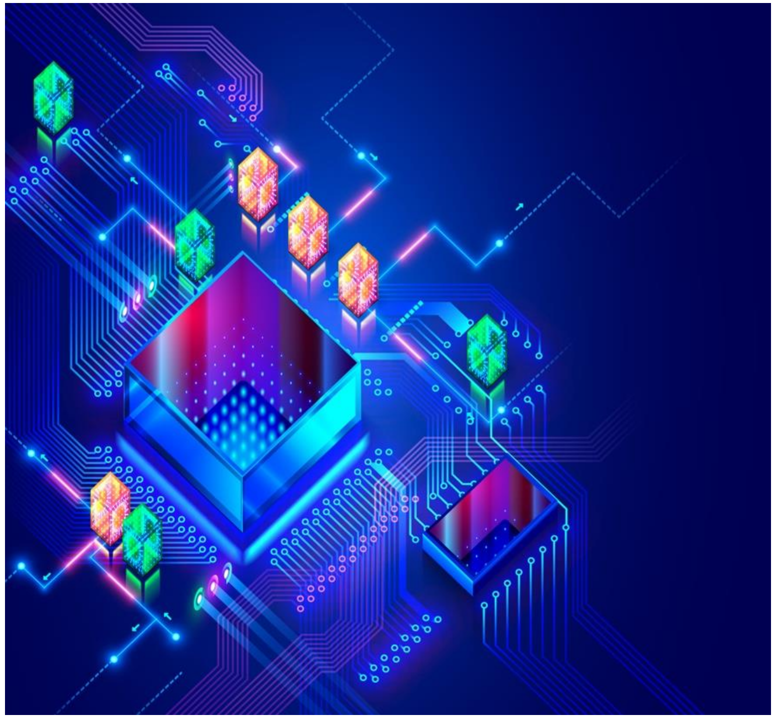
Data lakes and models are becoming larger and more complex

- This means every piece of data is critical
- It takes time, energy, and money to create these models

Data errors and loss WILL occur

- Fixing the minor errors when they happen are cheaper and better for your customers
- Recovery from lost data is expensive – shorter is better!

We Invented the Future of Storage – Join Us!



To learn more, visit: www.graidtech.com

We Invented the Future of Storage.

LEARN MORE AT GRAIDTECH.COM