Graid Technology Inc. ATTO ScaleFlux

Next-Gen Innovation Starts Here: Deliver Unparalleled NVMe RAID Performance with Greater Flexibility and Precision Control

PARTNER SOLUTION BRIEF | AUGUST 2025

Redefine Agility at Scale in High-Performance Storage Solutions with ATTO Technology, Graid Technology, and ScaleFlux®

As enterprises and data centers adapt to meet the growing demands of AI, edge computing, and data-intensive workflows, IT leaders must redefine infrastructure priorities and rethink how performance and flexibility are delivered across a modern IT ecosystem - without adding complexity or operational overhead. Whether powering AI model training, enabling real-time analytics, supporting hi-res media ingest, or scaling private cloud environments, success depends on more than just raw speed. It requires direct, fast access to dense storage – combined with granular visibility, intelligent control, and uncompromised data integrity – to drive efficiency from the core to the edge.

The shift to NVMe has redefined expectations for speed, parallelism, and latency. Applications demanding faster data access are being underserved by traditional storage stacks under the weight of inefficient data paths, CPU bottlenecks, limited PCIe resources, and complexity with management. Configuring scalable NVMe systems often involves trade-offs that prevent organizations from fully leveraging its capabilities – impacting overall efficiency, time to revenue, and the ability to meet performance-driven business objectives. To overcome these challenges, IT organizations and data-driven businesses are seeking storage platforms that unlock the true performance and economic value of NVMe at scale.

Streamline Data Modeling and Orchestrate Workflow Automation with Low-Latency, Scalable and Efficient Solutions

Designing NVMe storage that keeps AI and ML continuously fed requires precise validation and tuning so data moves efficiently and GPUs stay busy. Complex RAID stacks, CPU contention, and unnecessary hops add latency that stretches deployments, raises costs, and caps throughput for training and inference while putting data integrity at risk. Teams also need a path to automate data movement and orchestration without adding more infrastructure cards or management burden.

The pressure is highest at the edge, in smart-city and IoT systems, and inside dense HPC clusters where space, power, and thermals are tightly constrained. Traditional RAID hardware and heavy cabling consume valuable slots and watts, slow time to deploy, and make it harder to deliver enterprise-class reliability, scalability, and real-time responsiveness.

This joint architecture establishes a low-latency pipeline to intelligent storage. ATTO ExpressNVM™ switch adapters create direct, high-speed NVMe connections between compute nodes and storage, eliminating unecessary hops and latency. Graid Technology's GPU-based SupremeRAID™ offloads RAID to the GPU, freeing CPU resources and sustaining full NVMe/NVMe-oF performance with strong data protection. ScaleFlux® computational NVMe SSDs compress and decompress data in-drive to reduce CPU overhead and accelerate I/O transactions.

Together, these components remove bottlenecks, shrink power and space footprints, and speed deployment while enabling automated, scalable data workflows—from edge inference to large-scale training—so models train faster and serve results with higher efficiency.

Fuel AI Workloads with GPU-optimized Storage – Built for Unmatched Reliability and Performance

As intelligent and data-intensive applications continue to push the boundaries of what modern infrastructure can handle, GPU servers are increasingly at the heart of enterprise and edge computing environments. However, achieving peak performance requires more than just powerful GPUs – it demands a storage architecture that is capable of transferring vast amounts of data at a rapid pace, instantaneous access to data by maintaining low-latency I/O transactions, and managing and optimizing data placement at scale.

Direct, High-Bandwidth Access to NVMe

- ATTO ExpressNVM™ PCIe Switch Adapters sit between the GPU server's CPU complex and a bank of NVMe drives, allowing for direct PCIe lane allocation to GPUs and NVMe storage reducing contention by bypassing bottlenecks in the CPU root complex.
- GPU servers can read/write directly to large pools of managed and self-contained NVMe storage with negligible latency, essential for high frame rate training data ingestion or rapid checkpoint writes during AI/ML training. With built-in FPGA intelligence, ExpressNVM offers granular control and visibility of storage assets not available on other commodity HBAs or RAID controller cards – making them an ideal choice for data-intensive workloads requiring rapid ingestion and checkpoint writes.

GPU-powered RAID Protection

- With SupremeRAID™ parity calculations, striping, and rebuild operations are handled by your existing GPU using SupremeRAID™ AE (AI Edition). This replaces the need for CPU-bound RAID controllers and avoids stealing compute cycles from AI workloads.
- RAID calculations occur at GPU speeds and are significantly faster than traditional RAID implementations – benefiting from the aggregated performance of NVMe SSDs to deliver high-performance data protection with faster rebuild times at scale.

Computational Storage for Preprocessing

- When GPU servers pull raw datasets from storage, compression reduces I/O latency, accelerating the transfer of data between the GPU and the SSD to increase GPU utilization and efficiency. In-drive compression on ScaleFlux® NVMe CSSDs, reduces data volumes written to the flash, completing checkpoints faster and accelerating reads.
- With data compression offloaded from the GPUs, GPUs are freed up to focus on their core tasks such as modeling and training, cutting idle times. The result -- dramatically improved overall system performance and responsiveness for more efficient and scalable workflows, particularly when dealing with massive AI, ML, or high-resolution media datasets.

Use Cases

Edge AI & Industrial IoT

Space and power constrained deployments such as autonomous systems, smart factories, and IoT nodes benefit from compact, high-throughput NVMe storage with GPU-offloaded RAID and on-drive pre-processing, enabling real-time analytics at the edge.

High-Performance Computing & Research

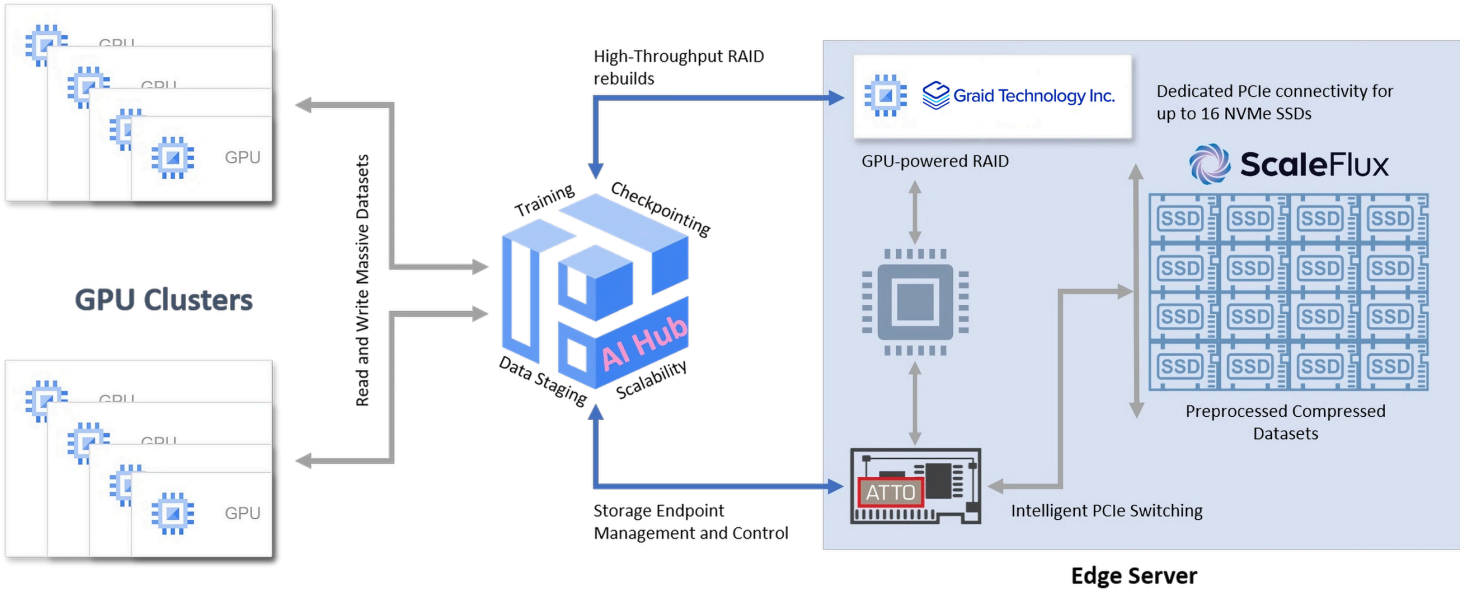
Simulation, AI modeling, and computational research workloads require rapid dataset staging and low latency access to data. Accelerate research cycles by enabling multiple GPU servers to share NVMe pools without CPU bottlenecks.

Media & Entertainment

4K/8K video ingest, VFX, and rendering pipelines rely on fast storage and preprocessing. ATTO, SupremeRAID™, and ScaleFlux® work in unison to streamline offloading and preparation of massive media assets while maintaining compute focus on creative workloads.

Enterprise IT

Organizations running large-scale AI and machine learning pipelines gain the advantage of high throughputs for ingesting massive datasets, performing frequent checkpointing, and enabling rapid iteration for model optimization and computation.



Conclusion

Unlock unprecedented performance, reliability, and efficiency for AI, ML, and data intensive workloads with this turnkey solution from ATTO, SupremeRAID™, and ScaleFlux®.

Simplify deployment, reduce bottlenecks, and maximize utilization of storage resources while keeping compute efficiency under check – allowing IT teams and system integrators to focus on innovation rather than bare metal SKUs.

Accelerate your AI infrastructure today: explore how this integrated solution can transform your workflows and deliver measurable results at scale.

Contact Graid Technology at info@graidtech.com to learn more.

Key Highlights

Zero CPU Tax for RAID & Storage I/O
More CPU resources available for orchestration, data loading, or other system tasks.

Optimized PCIe Lane Allocation
Ensuring NVMe drive groups operate at full bandwidth simultaneously with intelligent PCIe switching.

Faster Dataset Preparation
Offload preprocessing and management tasks, allocating time and resources for AI computation and modeling.

High-density, Low-latency Architecture
Essential for real-time analytics, inferencing, and time-sensitive training.

Resilient Enterprise Data Protection
Ensure data availability and prevent interruptions to your workloads.

Maximize Storage with Limited PCIe Slots
ATTO ExpressNVM™ allows up to 16 drives from a single PCIe x16 slot and SupremeRAID™ AE will enable data protection and maximize performance of high performance ScaleFlux® CDS5000 Series SSDs.

About Graid Technology, Inc.

Graid Technology is led by a dedicated team of experts with decades of experience in the SDS, ASIC, and storage industries and continues to push boundaries in data storage innovation by protecting NVMe-based data from the desktop to the cloud. Cutting-edge SupremeRAID™ GPU-based RAID removes the traditional RAID bottleneck to deliver maximum SSD performance without consuming CPU cycles or creating throughput bottlenecks, providing unmatched flexibility, performance, and value. With headquarters in Silicon Valley supported by a robust R&D center in Taiwan, we are globally committed to spearheading advancements in storage solutions. For detailed product information, visit our website or connect with us on LinkedIn.

About ATTO Technology, Inc.

Since 1988, ATTO Technology has been a global leader in network and storage connectivity and infrastructure solutions for the most data-intensive environments. We manufacture host adapters, network adapters, intelligent bridges, Thunderbolt™ adapters, and software. ATTO's exclusive technologies provide an exceptional level of performance to all storage interfaces, including Fibre Channel, SAS, SATA, iSCSI, Ethernet, NVMe and Thunderbolt. Proudly offering products engineered and manufactured in the USA. Learn more at https://www.atto.com/.

About ScaleFlux

The enterprise Flash Storage and Memory innovators. ScaleFlux products unlock unprecedented performance, efficiency, security, and scalability for your data infrastructure. From AI/ML, enterprise, and data center to edge computing and inferencing, ScaleFlux NVMe and CXL solutions make your data fly. Learn more at https://scaleflux.com/.