PLiOPS LightningAI

vLLM *KV-Cache Offloading*

Pliops LightningAI

World's Leading KV-Cache Offload Solution

Pliops LightningAI – PB scale Memory Tier for GPU

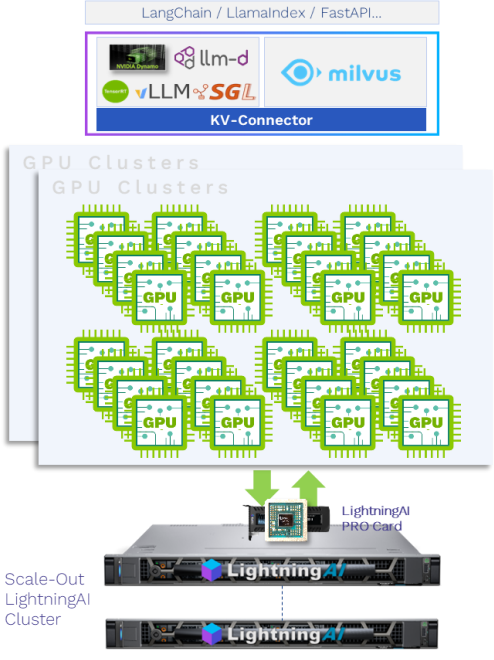
Pliops proudly presents our groundbreaking LightningAI, a revolutionary solution that optimizes GenAI inference infrastructure by reducing GPU compute per chat query I/O. This innovation unlocks seamless data retention, cost savings, and breakthrough performance for GenAI and LLM applications. Built as an Accelerated Key-Value distributed smart node, LightningAI introduces a new tier of memory below HBM for GPU compute workloads. It accelerates end-to-end performance by more than 3X and enhances efficiency across all leading inferencing frameworks, including vLLM and others. Leveraging Pliops' state-of-the-art technology, LightningAI substantially improves GPU utilization, enabling significant TCO reductions, lower power consumption, and decreased carbon emissions—delivering advanced AI solutions for the next generation of intelligent infrastructure.

- ✓ Best Efficiency
- ✓ Best Performance

3X

Faster for multiple GenAI Application

Any GPU
Any Model



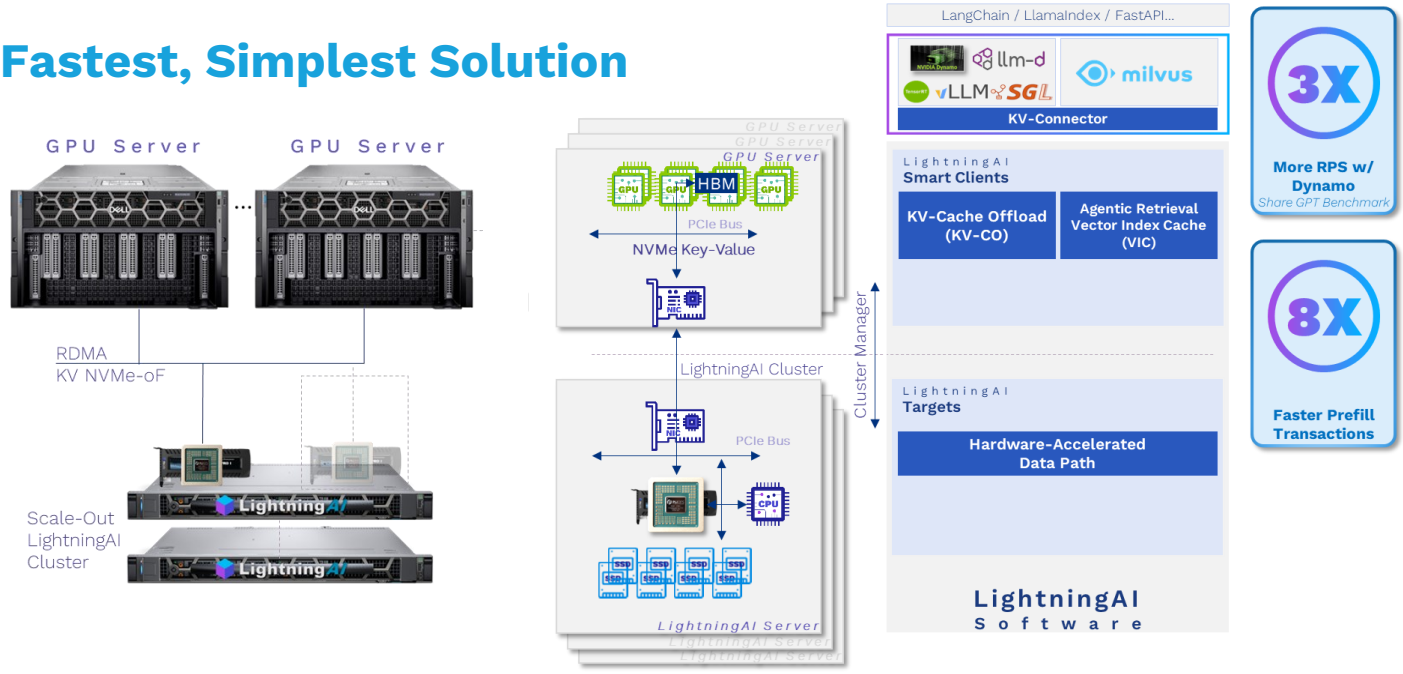
- Scale-out, multi-rack solution
- Supported out-of-box by vLLM and more
- Deep (PB scale) & affordable Long-Term Memory
- Simple integration with distributed inference frameworks
- KV-Cache Offload:
  - at memory level performance.
  - at 3X E2E LLM Inference efficiency
- KV-Vector-Index-Cache:
  - Huge # of vectors at a fraction of the cost
- >30X More IOPS density vs best-in-class FS
- Global Name space, No indexing overhead
- Unified context management and data flow
- Self healing + HW compression

Pliops vLLM KV-Cache Offload in Production, Multiple OEMs + SIs

PLiOPS LightningAI

vLLM *KV-Cache Offloading*

Fastest, Simplest Solution



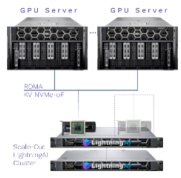
LightningAI vs. GPU Host Cache Solutions

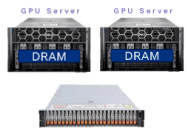
LightningAI delivers HBM-like performance, unlimited scalability, and cost efficiency by leveraging low-cost SSDs in a streamlined, single-tier architecture. It offers a smaller footprint, optimal load balancing, and simplified operations.

In contrast, GPU Host Cache solutions are performance-sensitive, relying heavily on L2 hit rates. They suffer from limited scalability, high DRAM/SSD costs, and complex multi-tier designs that increase power consumption and create routing inefficiencies.

Bottom line: LightningAI provides a high-performance, scalable, and cost-effective alternative to traditional GPU caching approaches, with simpler management and better efficiency.

NVIDIA Dynamo + llm-d vLLM Production Stack + LightningAI		NVIDIA Dynamo + llm-d vLLM Production Stack
Field-Proven, HBM-like Performance	Performance	Highly depends on the L2 hit rate Degradation without a high hit rate
Infinite. A single high-performance, low-latency, petabyte-scale tier	Scalability	Limited. Performance degrades with L2 misses.
Low-cost SSD	Cost/Capex	50X DRAM/SSD cost ratio. Expensive, low-performance FS
Smaller footprint.	Power / Opex	Additional L2(DRAM) layer, high-footprint for shared storage
Streamlined, Single-tier, Single-flow design	Simplicity	Complex, multi-tier, hard-to-tune
Optimal load balancing	Efficiency	Cache-aware routing creates hot-spots





102825 || Copyright© 2025 Pliops. All rights reserved