SOLUTION BRIEF

Accelerating LLM Inference at Scale

UCloud MindForge AI Platform & Pliops LightningAI

Executive Summary

Enterprises running large-scale AI workloads face growing challenges in delivering fast, cost-efficient inference for large language models (LLMs). UCloud's MindForge AI Platform, combined with Pliops LightningAI, provides a breakthrough solution that eliminates GPU memory bottlenecks, dramatically improves performance, and scales seamlessly across heterogeneous environments.

Pliops LightningAI introduces a Deep Memory Layer that offloads and accelerates KV-cache operations, freeing GPU resources and unlocking higher throughput with better user experience.

Key Benefits

70% Faster TTFT

30% Reduction in TPOT

2X Increase in concurrency

- Broad GPU compatibility
- Seamless private cloud integration
- Cost efficiency

The Challenge

LLM inference at scale is constrained by KV-Cache bottlenecks in GPU high-bandwidth memory (HBM), resulting in high latency, limited concurrency, and escalating costs for GPU resources. Enterprises need a solution that accelerates inference without compromising accuracy or compatibility.

The Solution

Pliops LightningAI integrates natively with UCloud MindForge AI Platform to offload KV-Cache operations from GPU memory to Pliops high-performance KV acceleration hardware. This hardware delivers ultra-fast KV access, compression and quantization, reducing GPU load and enabling higher throughput and efficiency.

Ideal Use Cases

Enterprise AI platforms powering chatbots, copilots, and generative AI services; Financial services, e-commerce, and research institutions operating multi-tenant GPU clusters; AIGC workloads requiring long-context inference.

Deployment Model

Delivered as part of UCloud MindForge AI Platform with Pliops LightningAI acceleration suite. Requires Pliops hardware, supports private cloud deployments and multi-cluster scaling, and integrates with UCloud AI App Market for one-click enablement.

Next Steps

Contact UCloud or Pliops for architecture validation and sizing guidance. Request benchmark reproducibility pack for TTFT/TPOT gains and concurrency uplift. Explore co-marketing opportunities and AI App Market listing for streamlined adoption.

Learn More

UCloud MindForge AI Platform: ucloud-global.com/en | Pliops LightningAI: pliops.com

References

Setting the Standard: Multi-Turn Benchmarking in vLLM
Know how you ask your AI assistant to teach you about a new technology you've been hearing about but don't fully understand?

SNIA SDC 2025 Recap: Disaggregated KV Storage for LLMs
Technical tutorial on vLLM + LightningAI deployments.

Achieving Huge, Affordable Vector Databases
Vector search has emerged as a key part of AI application workflows. In retrieval-augmented generation (RAG), for example, it serves for reducing the LLM inference context window, for increased relevance and speed. Modern vector search applications scale to tens of billions of documents.

How LightningAI Redefines KV Cache Offloading for Scalable GenAI
Explore how Pliops LightningAI redefines KV cache offloading to enable scalable GenAI inference with higher efficiency, lower costs, and improved performance.