XDP LightningAI

The Next Memory Tier for Gen-AI Infrastructure

Solving AI Infrastructure Challenges:

Pliops is enabling AI-powered businesses and hyperscalers achieve impressive performance with significantly less power, reduced costs, and a smaller carbon footprint. Pliops' innovative XDP LightningAI solution enables sustainable, high-efficiency AI operations when paired with GPU servers. By increasing GPU inference throughput by up to 8X and significantly improving efficiency, this breakthrough product reduces data center TCO by more than 50%, lowers cooling costs and CO2 emissions by half, and empowers businesses to sustainably expand their AI capacities.

5X

Lower cost per million tokens

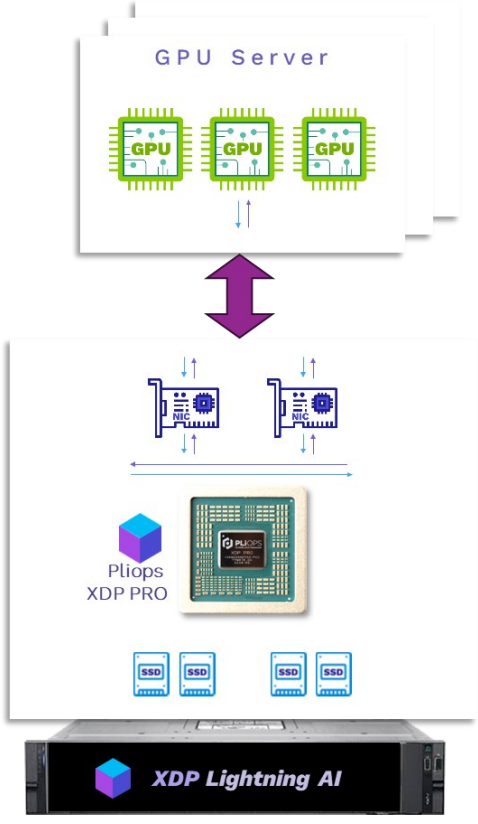
Organizations are increasingly concerned about the rising power consumption in data centers, particularly as AI infrastructure and emerging AI applications lead to higher energy footprints and strain cooling systems. As they scale their AI operations and add GPU compute tiers, the escalating power and cooling demands, coupled with significant capital investments in GPUs, are eroding margins. ***A monumental challenge looms as data centers struggle to secure essential power,*** creating significant pressure for companies striving to manage costs while expanding their AI capabilities.

Pliops XDP LightningAI
A Peta Byte Memory Tier for GPUs:

Pliops knows that efficient infrastructure solutions are essential to address these issues – and the company's newest Extreme Data Processor (XDP), XDP-PRO ASIC – plus a rich AI software stack and distributed nodes – address GenAI challenges by utilizing a GPU-initiated Key-Value I/O interface as a foundation, creating a memory tier for GPUs, below HBM.

Pliops XDP LightningAI easily connects to GPU servers by leveraging the mature NVMe-oF storage ecosystem to provide a distributed Key-Value service.

Pliops has focused on LLM inferencing, a crucial and rapidly evolving area within the GenAI world that demands significant efficiency improvements. The company's demo at SC24 is centered around accelerating LLM Inferencing applications. This same memory tier is seamlessly applicable for other GenAI applications that Pliops plans to introduce over the next few months.



vLLM Demo

Most Efficient GenAI LLM Inferencing

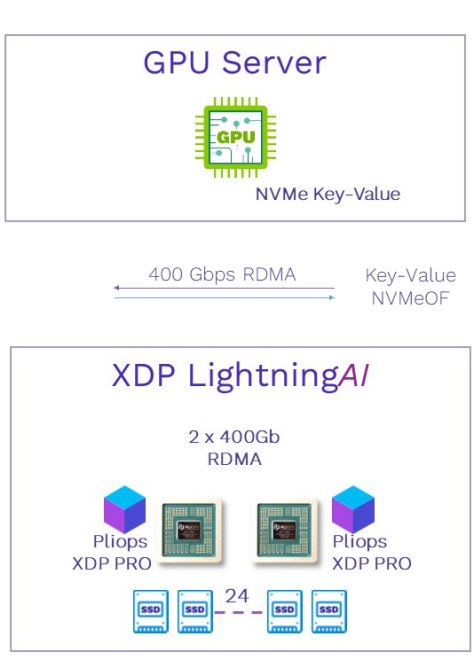
8X

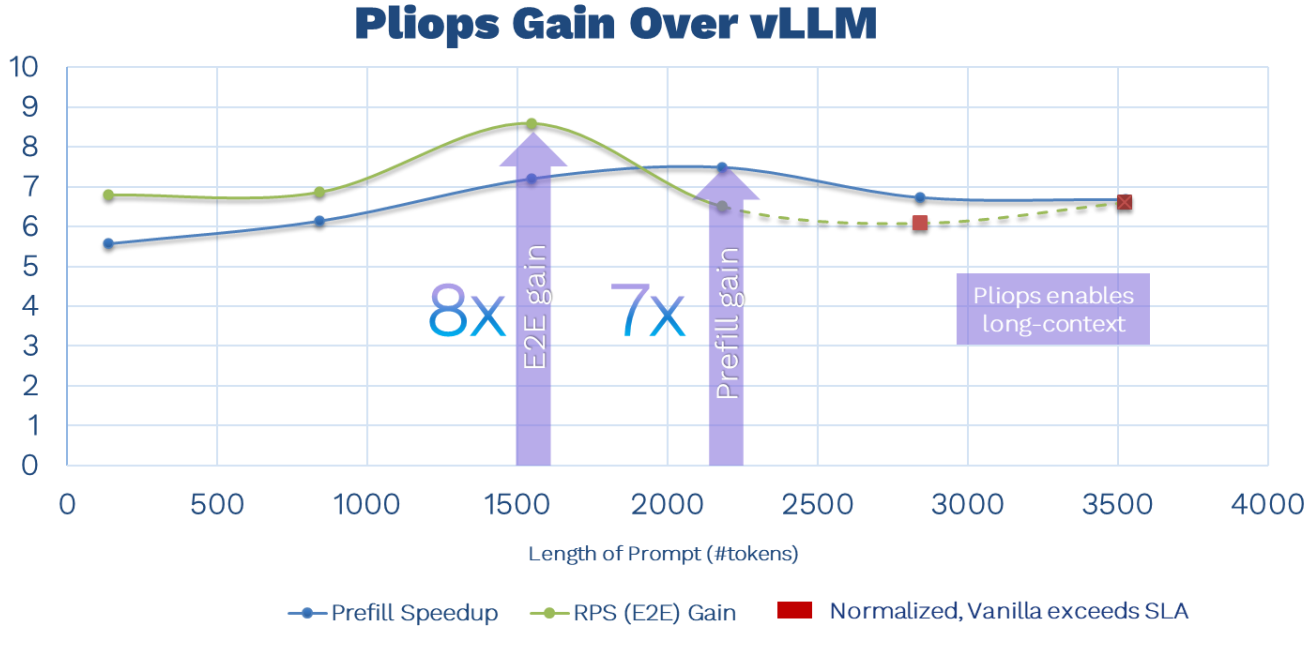
E2E LLM Inferencing Gain

> 8X LLM Inference Performance – vLLM Demo

Pliops' groundbreaking XDP LightningAI, a revolutionary Accelerated Key-Value distributed smart node, introduces a new peta-byte tier of memory below HBM for GPU compute applications.

Pliops is currently demonstrating XDP LightningAI accelerating E2E performance by up to 8X and enhancing efficiency for vLLM, a leading inferencing solution. Pliops XDP LightningAI is currently being tested and exceeding performance targets at key partner sites.



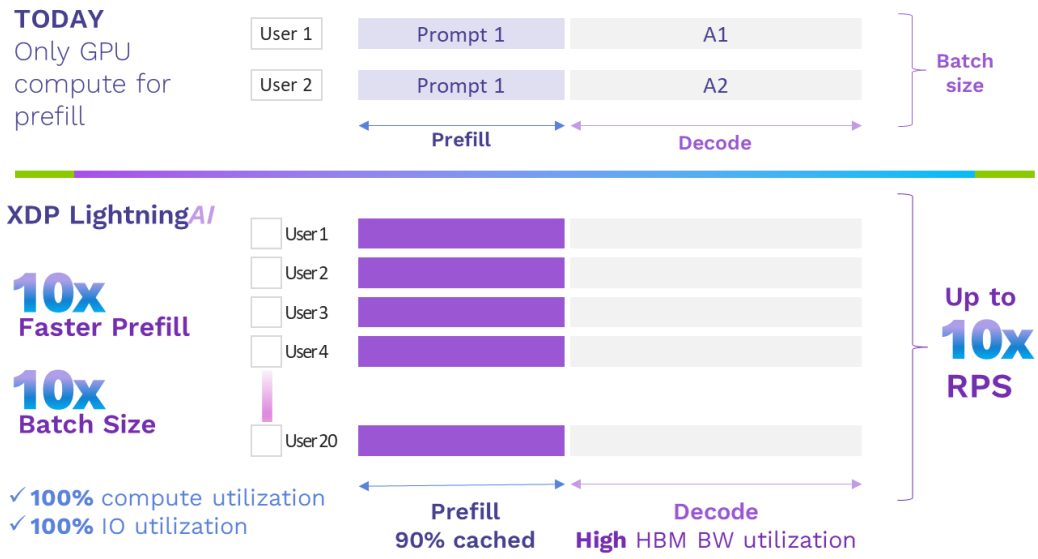


Most Efficient GenAI LLM Inferencing

All About Improving Prefill & Decode Operations

Prefill Acceleration Drives E2E Performance Gains

In today's LLM inferencing computing, GPU prefill operations are heavily compute-bound and critically determine the batch size. While prefill can fully utilize GPU resources, increasing the batch size beyond a certain point only increases the Time to First Token (TTFT) without improving prefill rate. On the other hand, GPU decode operations are HBM bandwidth-bound and mainly influenced by model and KV cache sizes, benefiting significantly from larger batch sizes through higher HBM bandwidth efficiency. Pliops' solution improves prefill time, allowing for larger batch sizes without violating user SLA for prefill operations. This enhancement directly affects decode performance as well, as it gains greatly from the increased batch size. As a result, by improving prefill time, the system achieves nearly proportional improvements in end-to-end throughput.



Pliops XDP LightningAI enables 10x faster prefill operations and batch sizes with 90% data caching and improved HBM bandwidth utilization, ensuring optimal resource use and substantial end-to-end gains.

Pliops XDP LightningAI

- Peta-Bytes Scale memory for AI (100x HBM)
- Accelerated Key-Value store
- Highly parallel
- Distributed and natively virtualized
- Cost-efficient (1/100 of HBM)

Applications

- Replacing compute with 100x lower cost memory
- Accelerated LLM Inference and Training
- Personalized memory layer for users
- Accelerated Vector Databases
- Many more

A PARADIGM SHIFT FOR GEN-AI