Apr 2, 2025 11:00 AM Eastern Daylight Time

MangoBoost Achieves Record-Breaking MLPerf Inference v5.0 Results for Llama2-70B Offline on AMD Instinct™ MI300X GPUs

Share

BELLEVUE, Wash.--(BUSINESS WIRE)--MangoBoost, a provider of cutting-edge system solutions designed to maximize AI data center efficiency, has set a new industry benchmark with its latest MLPerf Inference v5.0 submission. The company's Mango LLMBoost™ AI Enterprise MLOps software has demonstrated unparalleled performance on AMD Instinct™ MI300X GPUs, delivering the highest-ever recorded results for Llama2-70B in the offline inference category.

> *This milestone marks the first-ever multi-node MLPerf inference result on AMD Instinct™ MI300X GPUs.*
>
> Share

This milestone marks the first-ever multi-node MLPerf inference result on AMD Instinct™ MI300X GPUs. By harnessing the power of 32 MI300X GPUs across four

Hardware – H100 vs. MI300X Pricing)— Mango LLMBoost™ delivers up to 62% cost savings while maintaining industry-leading inference throughput.

In terms of cost-efficiency, the Mango LLMBoost™ + MI300X system delivers approximately 2.8× more inference throughput per $1,000 spent than the H100-based system, making it the clear choice for high-performance, budget-conscious deployments.

Mango LLMBoost™: A Scalable and Hardware-Flexible MLOps Solution

Mango LLMBoost™ is an enterprise-grade AI inference software that provides seamless scalability and cross-platform compatibility. It supports over 50 open models, including Llama, Qwen, and DeepSeek, with one-line deployment via Docker and built-in OpenAI-compatible APIs. The software is cloud-ready—available on AWS Marketplace, Microsoft Azure Marketplace, and Google Cloud Platform—and is also available for on-premise deployment for enterprises requiring full control and security.

Key capabilities of Mango LLMBoost™ include:

- **Auto Parallelization** – Efficiently distributes large models across GPUs and nodes.
- **Auto Config Tuning** – Optimizes runtime parameters based on workload characteristics.
- **Auto Context Scaling** – Dynamically adapts memory usage to maximize GPU utilization.
- **Auto Disaggregated Deployment** – Ensures flexible deployment across multiple inference stages.

LLaMA3.1-8B. In terms of cost-efficiency, Mango LLMBoost™ also leads the pack with the lowest GPU cost per million tokens, reducing inference cost by over 99% compared to Ollama, and by over 30% even compared to vLLM on high-throughput workloads.

Expanding AI Infrastructure Solutions

In addition to the Mango LLMBoost™ software, MangoBoost offers hardware acceleration solutions based on Data Processing Units (DPUs) to enhance AI and cloud infrastructure, including:

- **Mango GPUBoost™** – RDMA acceleration for multi-node inference and training via RoCEv2.
- **Mango NetworkBoost™** – TCP/IP stack offloading for enhanced CPU efficiency.
- **Mango StorageBoost™** – High-performance NVMe/TCP initiator and target solutions for scalable AI storage.

For more information, please refer to our technical blog or reach out to contact@mangoboost.io.

About MangoBoost

MangoBoost delivers cutting-edge, full-stack system solutions that maximize AI data center efficiency. The company offers a high-performance DPU that seamlessly integrates with general-purpose GPUs, accelerators, and storage products, enabling cost-effective, standardized AI infrastructure. In addition, MangoBoost provides AI inference optimization software that enhances GPU efficiency for large-scale LLM workloads, accelerating deployment and reducing operational costs.