MANGOBOOST | Product Brief

Mango LLMBoost™ AI Inferencing Server

A ready-to-deploy, full-stack AI inferencing server offering unprecedented performance and flexibility

EXECUTIVE SUMMARY

The emergence of Large Language Models (LLMs) and their ability to both capture and generate contextual information has enabled applications such as advanced chatbots, visual analysis and generation, and enhanced programmer productivity tools. From scalable hyperscale inferencing to resource-constrained edge devices, the ability to run LLM inference has emerged as an essential capability. Regardless, the actual deployment of LLM inference in production faces multiple challenges. Once a model is ready for inference, developers face the productization tax. Integrating the ML processing pipeline into the rest of the application often requires additional time and engineering effort, leading to delays in productization.

Mango LLMBoost™ addresses all these challenges by creating an easy-to-use container that allows LLM experts to optimize their models and select the suitable GPUs on demand. Our inference engine employs three different forms of GPU parallelism, allowing GPUs to balance their compute, memory, and network resource usage to achieve maximum performance. Intelligent job scheduling also optimizes cluster-wide GPU resources, ensuring that load is equally balanced across nodes. Mango LLMBoost™ further ensures effective use of low-latency GPU caches and high-bandwidth memory through quantization, reducing data footprint without loss in accuracy.

HIGHLIGHTS

INDUSTRY-LEADING INFERENCE SERVING

- Cost-efficiency: Save up to 92% on inference costs
- High-performance: Speed up your inference up to 12.6x over other popular inference frameworks
- Flexible: Seamlessly deploy LLMBoost across all popular NVIDIA and AMD GPUs

MULTI-MODEL DEPLOYMENT AND MANAGEMENT

- Effortlessly deploy a diverse range of chat-based and multi-modal models, including Llama, Mixtral, Gemma, Qwen2, Llava, Phi3, Chameleon, MiniCPM, GLM-v4
- Deploy and manage multiple models in a single inference server with automated resource allocation

HASSLE-FREE DEPLOYMENT

- Deploy end-to-end with our web-serving and streaming APIs
- Easily tune kernel performance on your workloads with the push of a button
- Seamlessly integrate LLMBoost into your existing AI applications with our OpenAI-compatible endpoints

SPECIFICATIONS

CAPABILITIES

- OpenAI-compatible API
- Multi-model Deployment
- Intelligent Scheduling
- Advanced Resource Allocation

INTERFACE

- REST API
- Websocket API
- Native Programming Interface (Python)

PERFORMANCE SUMMARY

OpenOrca Tok/s (8xMI300x)

Llama3.1-70B (FP16)	13K
Llama3.1-70B (FP8)	21K

OpenOrca Tok/s (8xL40s)

Llama3.1-70B (FP16)	1.4K
Llama3.1-70B (FP8)	1.8K

I128-O128 Tok/s (8xMI300x)

Llama3.1-70B (FP16)	9K
Llama3.1-70B (FP8)	29K
Llama3.1-405B (FP16)	2.6K
Llama3.1-405B (FP8)	4.2K

I128-O2048 Tok/s (8xMI300x)

Llama3.1-70B (FP16)	5K
Llama3.1-70B (FP8)	27K
Llama3.1-405B (FP16)	1.5K
Llama3.1-405B (FP8)	3.4K

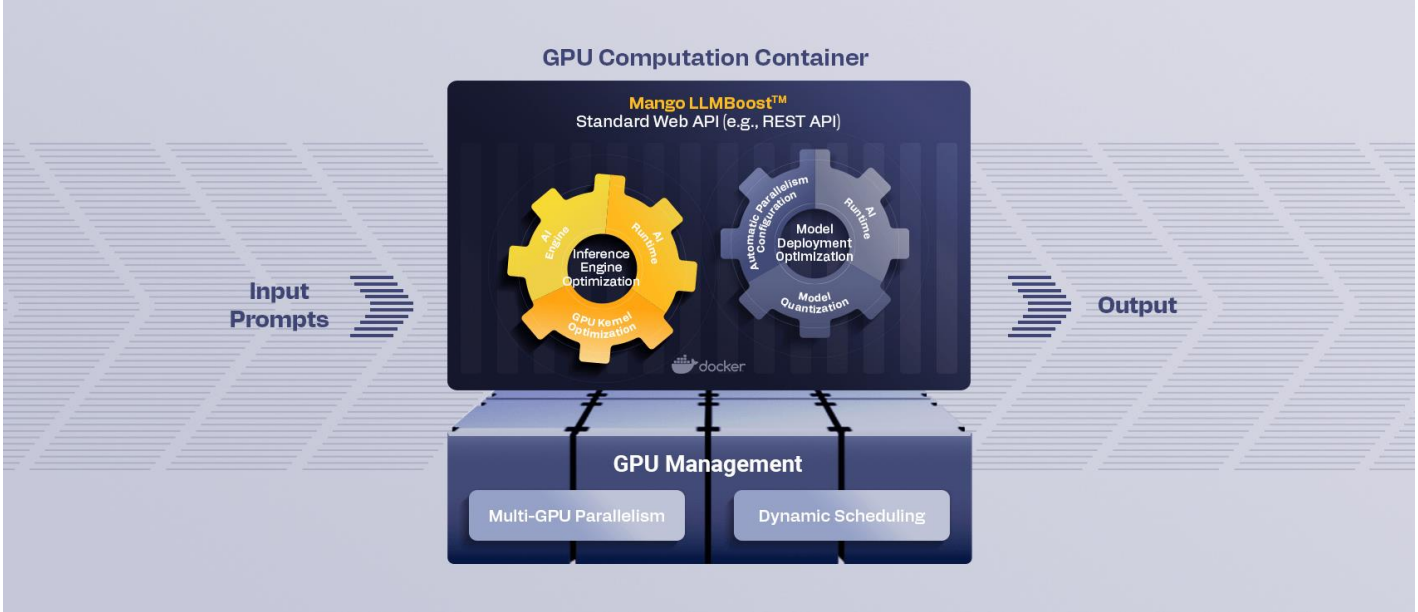
I2048-O128 Tok/s (8xMI300x)

Llama3.1-70B (FP16)	20K
Llama3.1-70B (FP8)	47K
Llama3.1-405B (FP16)	3.8K
Llama3.1-405B (FP8)	5K

I2048-O2048 Tok/s (8xMI300x)

Llama3.1-70B (FP16)	6K
Llama3.1-70B (FP8)	27K
Llama3.1-405B (FP16)	1.9K
Llama3.1-405B (FP8)	3.2K

DESIGN OVERVIEW



Inference Engine Optimization

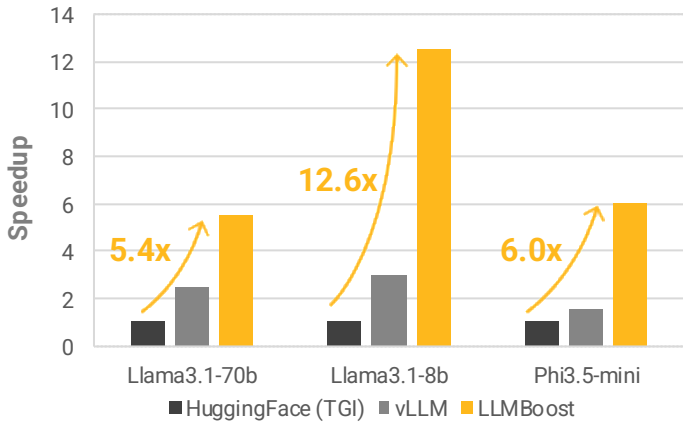
Mango LLMBoost™ is optimized to take full advantage of the capabilities of modern GPUs. At its core, our inference solution combines intelligent job scheduling and enhanced parallelism to reduce hardware bottlenecks and ensure that all GPUs are fully utilized. Our push-button kernel optimization feature automatically selects the highest performance GPU kernels for given workloads, yielding higher inference performance on any given GPU. To fully utilize GPU compute capability, Mango LLMBoost™ employs an intelligent memory management mechanism that reduces memory fragmentation within the GPU memory. This allows more useful data to be present on the device, unlocking unprecedented throughput and latency improvements.

Model Deployment Optimization

Mango LLMBoost™ allows users to easily take advantage of model quantization, which utilizes smaller FP8 data formats available in modern GPUs. Our FP8-optimized kernels enable modern GPUs to achieve much higher peak throughput by fully utilizing their compute units and reducing memory costs with no drop in accuracy. For large models, these techniques are critical to take advantage of GPU hardware resources. Additionally, Mango LLMBoost™ automatically discovers the highest-performance parallelism configuration for a multi-GPU deployment. With the combination of deployment optimization and inference engine optimization, Mango LLMBoost™ offers unprecedented performance and cost efficiency for your application.

EVALUATION RESULTS

Relative Performance Improvement
AWS, g6e.48xlarge, 8xL40S



Cost Saving
AWS, g6e.48xlarge, 8xL40S