MANGOBOOST | Product Brief

# Mango LLMBoost™ AI Inferencing Server

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 LLM integration into 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. In many cases, once the model is ready to be used for inference, developers often have to work with the productization tax, where additional time and engineering effort is required to integrate the ML processing pipeline into the rest of the application, and these often lead to delay in productization.

Mango LLMBoost™ addresses all these challenges by creating an easy-to-use container that allows LLM experts to optimize their desirable models and select the suitable GPUs on demand. To enable all these benefits, Mango LLMBoost™ modified and optimized the LLM Inference Engine to take full advantage of the parallelism of the GPU cores and orchestrate the inference jobs to make effective use of all available GPUs in the cloud or cluster. Mango LLMBoost™ further ensures that all the data make an effective use of the faster GPU-side caches and memory through quantization, which utilize the smaller data format without loss in accuracy.

## HIGHLIGHTS

### BEST PERFORMING INFERENCE SERVING

- Cost-efficient: Up to 92% inference cost saving
- High-performance: Speed up your inference up to 12.6x over other inference frameworks
- Flexible: Mango LLMBoost™ works on all popular NVIDIA and AMD GPUs

### MULTI-MODEL DEPLOYMENT AND MANAGEMENT

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

### HASSLE-FREE DEPLOYMENT

- End-to-end deployment options with our Web-serving and streaming APIs
- Push-button performance tuning: Mango LLMBoost™ can intelligently select the best performing configuration given the GPU and the running models
- OpenAI API Compatibility: LLMBoost can be easily integrated into existing AI applications with OpenAI's API

## SPECIFICATIONS

### CAPABILITIES

- OpenAI-compatible API
- Multi-model Deployment
- Intelligent Scheduling
- Advance 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 |

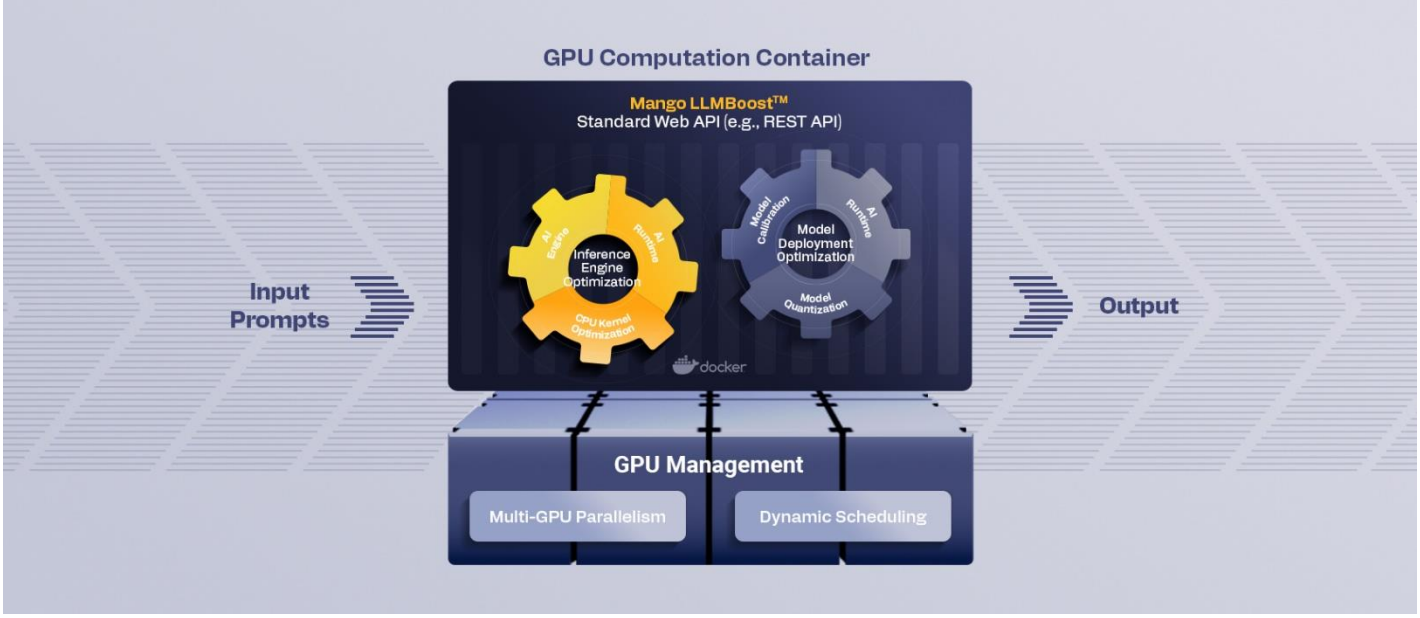
**I2048-O128** Tok/s (8xMI300x)

| | |
|---|---|
| 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™'s inference engine is optimized to take full advantage of modern GPUs. At its core, our optimization coordinates the system software schedule and ML computation to exploit parallelism existing in the hardware at all points to ensure all GPUs are fully utilized. Mango LLMBoost™'s kernel optimization allows the executed code to be configured automatically to yield the best performance on any given GPUs. To reduce the slowdown due to the data transfer between the host and the GPU's memory, Mango LLMBoost™ applies an intelligent prefetching mechanism that takes advantage of the LLM structure to allow data being used to always be available in the GPU's memory.

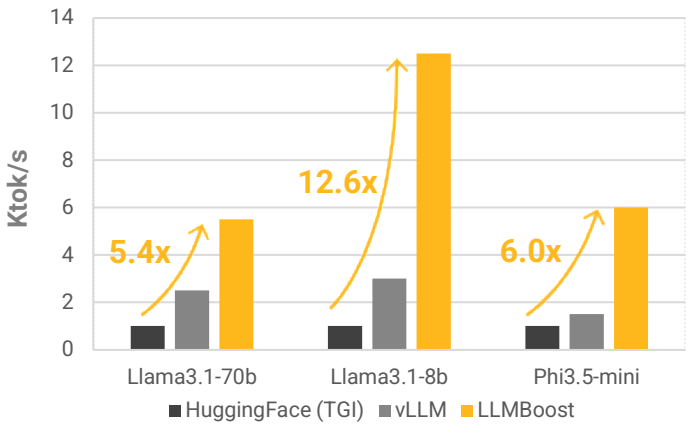
### Model Deployment Optimization

Mango LLMBoost™ performs additional optimization on the model deployment to take advantage of quantization, which utilizes smaller FP8 data formats available in modern GPUs. Additional calibration to the quantized model are applied to ensure no drop in accuracy. For large models, these techniques are critical to take advantage of the underlying GPU resources.

With the combination of model optimization and inference engine optimization, Mango LLMBoost™ is able to achieve higher performance than other state-of-the-art inference serving systems by enabling unprecedented level of parallelism, efficiently unlocking the power of the datacenter's resources from the caches, GPU's memory, host's memory and the network bandwidth.

## EVALUATION RESULTS

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




**Cost Saving**
AWS, g6e.48xlarge, 8xL40S