WHITE PAPER

MangoBoost's Turn-Key AI LLM Inference Serving solution on AMD Instinct™ MI300X GPUs with MLPerf® Inference v4.1 Llama Benchmark

REVISION 1.0 | Aug 28, 2024

MANGOBOOST

Disclaimer

The performance claims in this document are based on the internal cluster environment. Actual performance may vary depending on the server configuration. Software and workloads used in performance tests may have been optimized for performance only on MangoBoost products. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. Results that are based on pre-production systems and components as well as results that have been estimated or simulated using MangoBoost reference platform for informational purposes only. Results may vary based on future changes to any systems, components, specifications, or configurations. Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. MangoBoost does not guarantee any specific outcome. Nothing contained herein is, or shall be relied upon as, a promise or representation or warranty as to future performance of MangoBoost or any MangoBoost product. The information contained herein shall not be deemed to expand in any way the scope or effect of any representations or warranties contained in the definitive agreement for MangoBoost products.

The information contained herein may not be reproduced in whole or in part without prior written consent of MangoBoost. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. MangoBoost assumes no obligation to update or otherwise correct or revise this information and MangoBoost reserves the right to make changes to the content hereof from time to time without any notice. Nothing contained herein is intended by MangoBoost, nor should it be relied upon, as a promise or a representation as to the future.

MANGOBOOST MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

TABLE OF CONTENTS

01 | Key Points

- **New MLPerf® result on AMD Instinct™ MI300X GPU.** First MLPerf Inference v4.1 results on AMD MI300X GPU shows 23.5K token/second on Llama-70B benchmark! [1]
- **The productization challenge**. Going from MLPerf inference benchmark to production deployment is not easy, as it requires a tremendous effort (e.g., bring-up, co-optimize, and validate of full-stack AI inference serving software).
- **MangoBoost demo**. MangoBoost demonstrates two solutions to address the productization challenge and unlock the full power of GPUs, such as AMD MI300X, in AI inference.
- **Solution 1: Mango LLMBoost™**. Ready-to-deploy full-stack AI inference serving container, containing MangoBoost's inference serving software that complements popular vLLM, packaged with everything needed to run optimized FP8 Llama on AMD MI300X.

  > **Turn-key**. Out of the box, easily reproduce valid MLPerf Inference v4.1 Llama2-70B results with FP8 on AMD MI300X systems. [6]
  > **Production ready**. Mango LLMBoost container includes AI-inference server software optimized with vLLM, with added features for production (e.g., data parallelism and tensor parallelism, web API).
  > **Case studies**. MangoBoost reports results with Llama-70B from MLPerf inference and Llama Guard-7B from Meta, measured on MangoBoost's server with AMD MI300X GPUs.

- **Solution 2: Mango WebBoost™.** A hardware-accelerated web server for frontend gateway server for AI GPU inference clusters, which offers unparalleled throughput and reduced latency.

  > **DPU accelerated**. Mango WebBoost leverages MangoBoost Hardware Data Processing Unit (DPU) solution on an AMD Alveo™ U45N FPGA card to offer stateful offload of complete TCP/IP networking tasks, to achieve reliable low average latency and tail latency.
  > **Nginx case study**. Integrated and demonstrated with Nginx, one of the most widely used web servers. Can be also integrated into other web servers without modification.

Try it now. If interested in the above solutions, please contact us at contact@mangoboost.io

02 | Introduction

- The recent MLPerf Inference v4.1 publications [1] and news [2] have reported the first published result of a system with 8x AMD's MI300X GPUs on Llama-70B benchmark delivering state-of-the-art performance of 23.5K tokens/second. As reported by AMD, this result is by leveraging vLLM, which has been optimized for AMD GPUs [3].
- However, if one were to download vLLM, and run it out-of-the-box on MI300X GPUs, the AMD reported MLPerf performance level can be difficult to achieve. One needs to tune vLLM and ROCm, as well as quantize the Llama-70B model properly to FP8. Furthermore, vLLM does not support certain key features out-of-the-box, such as data parallelism to scale inference across multiple GPUs.
- Beyond the MLPerf benchmark, deploying MI300X in a production AI inference serving solution needs additional features, such as support for various web API endpoints, multi-model serving, hooks for cluster management, etc. Within the inference serving production cluster, one also needs a front-end web server to manage multiple back-end GPU servers.
- MangoBoost introduces two turn-key solutions to address above said challenges.

  > **Mango LLMBoost -** Ready-to-deploy full-stack AI inference serving container, containing MangoBoost's inference serving software that complements vLLM, packaged with everything needed to run optimized FP8 Llama on MI300X.

  > **Mango WebBoost -** A hardware-accelerated web server for frontend gateway server for AI GPU inference clusters, which offers unparalleled throughput and reduced latency.

- We detail these solutions in the next sections, and report case studies that quantify their benefits.

03 | Solution 1: Mango LLMBoost™

Ready-to-deploy AI inference software, demonstrated on AMD MI300X GPUs with MLPerf Inference v4.1 Llama-70B and Meta Llama Guard-7B.

- MangoBoost's LLMBoost AI inference server software is an optimized turn-key solution, ready-to-deploy as a container that is packaged and optimized with open-source vLLM and ROCm. It offers unparalleled LLM inference performance on AI servers with GPUs, such as AMD's MI300X. Designed for the next generation of high-memory capacity GPUs and LLMs, LLMBoost employs tensor and data parallelism schemes optimized for accelerating a variety of models, from larger models such Llama-70B, to more compact ones like Meta Llama Guard-7B.

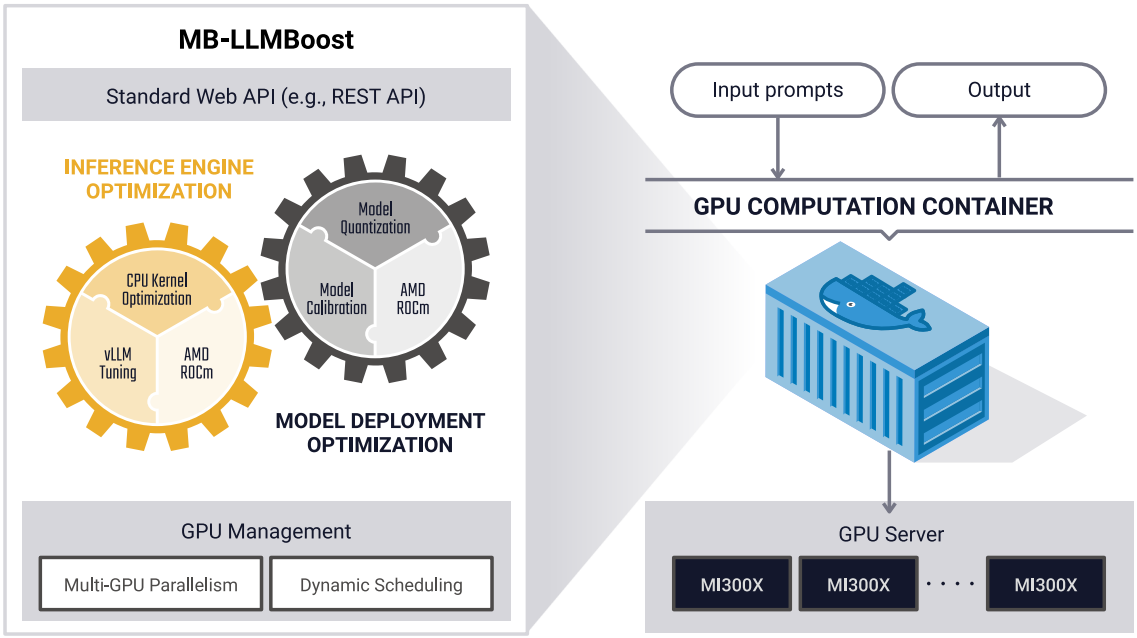

***FIGURE 1**: LLMBoost container - turn-key solution for LLM-serving based on vLLM*

- Fig. 1 highlights two key areas of optimizations that are applied in an LLMBoost container, namely the inference engine and model optimization. Our inference engine employs advanced GPU Kernel techniques allowing for multidimensional parallelism. This is coupled with domain-specific tuning of the ROCm software stack for low-precision computation. This approach enhances the efficiency of generalized matrix multiplication (GEMM) kernels and further economizes GPU shared memory usage and minimizes costly data transfers between the on-chip shared memory and the off-chip global memory. Asynchronous data transfer, along with data prefetching, PagedAttention, and KV-caching, has been optimized in vLLM to enhance memory efficiency and boost performance.

- Our inference server provides multiple endpoint options for seamless integration with existing solutions. Synchronous and asynchronous REST APIs offer comprehensive flexibility in request management, while WebSockets provide a streamlined interface for streaming applications. Our adaptable dynamic scheduling algorithms ensure peak GPU efficiency across diverse deployment scenarios by intelligently assigning incoming requests to the most suitable GPU. Customers can fine-tune their deployments to prioritize throughput, latency, or a hybrid approach, creating a fully customizable inference server that consistently delivers state-of-the-art performance on cutting-edge GPU hardware to meet a broad spectrum of customer needs.
- Furthermore, our containers expose standard Prometheus endpoints for monitoring performance, allowing seamless integration with existing Kubernetes deployments and enabling Horizontal Pod Autoscaling. Multiple models can be served in a single container with various degrees of flexibility offered to the user, tailoring the inference server to their specific needs.

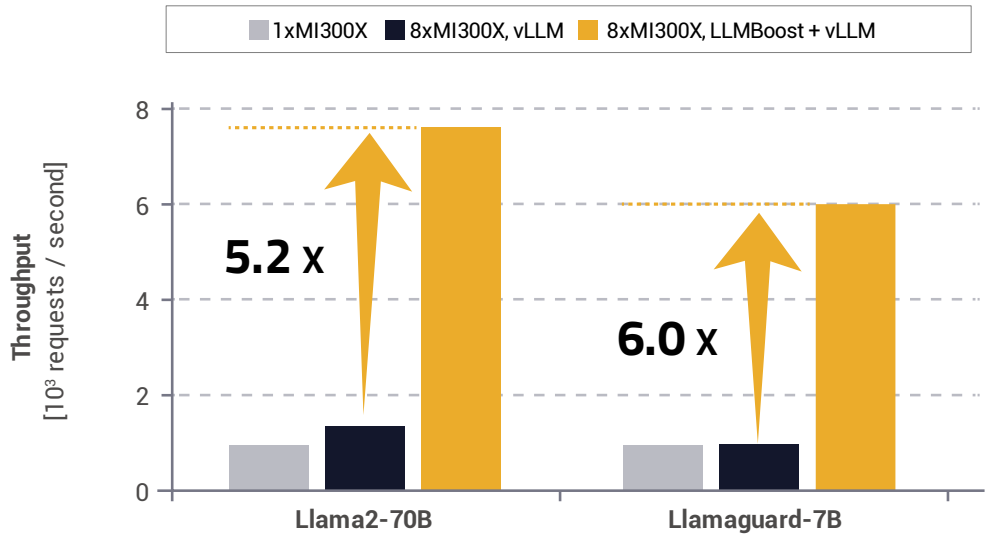

***FIGURE 2**: Throughput (Token/Second) normalized comparison between standard vLLM and vLLM enhanced by LLMBoost on a server with 8x AMD MI300XGPUs.*
LLMBoost offers data parallelism on top of vLLM, resulting in ~6x token/sec uplift.

- **Case Study with MLPerf Inference v4.1 Llama-70B and Meta Llama Guard-7B**. We integrated LLMBoost with vLLM and compared it with vLLM-only solution [4] [5]. We selected two prominent LLMs for this assessment: Llama2-70B from the MLPerf Inference v4.1 Benchmark and Llama Guard-7B from Meta. Our testing environment featured a high-performance Supermicro server equipped with 8x AMD MI300X GPUs. The details are in the table below.

GPU-Compute Server Configuration

Component	Description
CPU	• 2 sockets, AMD EPYC™ 9534 64-Core Processor
GPU	• 8 x AMD Instinct™ MI300X
Memory	• 8 x 192GB (total of 1,536GB) GPU HBM3 • 24 x 96GB (total of 2,304GB) server DDR5
OS	• Ubuntu 22.04.4 LTS
Workload	• Llama2-70B from MLPerf and LLamaGuard from Meta

- For the Llama2-70B model, we adhered to the MLPerf Inference v4.1 Closed Division rules, ensuring the validity and integrity of the results on the LLMBoost + vLLM configuration with all 8 GPUs active. This setup provided a consistent and rigorous foundation for comparison. Subsequently, identical conditions were applied across all other experiments to maintain fairness in our evaluation.
- For the MLPerf Inference v4.1 Llama-70B test case, we are able to achieve 22.8K tokens per second, which is within 3% of AMD's published MLPerf results of 23.5K tokens per second. Note that our system uses a different CPU than what AMD uses in their MLPerf official submission, which may account for the slight difference in performance.
- Since the MLPerf Inference Benchmark does not yet include validation rules for Llama Guard, we applied the same configuration used for Llama2-70B to ensure consistency in testing.
- As shown in Fig. 2, LLMBoost achieved a 5.2x to 6.0x improvement over the native vLLM thanks to three features of LLMBoost; (1) enable data parallelism, (2) dynamic scheduling across 8 GPUs, and (3) lightweight streamlined interface. This remarkable enhancement underscores the potential of LLMBoost to accelerate inference workloads, making it a crucial tool for optimizing large-scale AI deployments.

04 | Solution 2: Mango WebBoost

Accelerating Nginx web server with MangoBoost stateful TCP/IP hardware acceleration on an AMD U45N FPGA card

- WebBoost is a web-serving container for running Nginx on a front-end server equipped with hardware acceleration using MangoBoost DPU (data processing units) solution. Our solution can support multiple FPGA cards, and in this paper we demonstrate it on AMD U45N. The LLMBoost container and the WebBoost container are designed to work together seamlessly in a coordinated operation to stream input data to the GPUs and to return generated output tokens back to the clients. Fig. 3 below illustrates the network processing layers between arriving LLM requests and an LLM server's GPU. Clients submit their LLM requests into a REST API and then send them to the Nginx front-end gateway. Nginx collects all incoming requests and then distributes the request to the back-end GPU computation containers. In a front-end server with WebBoost, MangoBoost DPUs solution offers stateful hardware acceleration of the TCP/IP networking stack, to boost web-server software such as Nginx. WebBoost offers reliable low average latency and tail latency, as well as boosting throughput; these are all important for LLM inference. Additionally, WebBoost hooks the standard POSIX socket API calls, ensuring that it requires no modifications to the user application, making it an easy-to-use solution for seamless integration into existing systems.

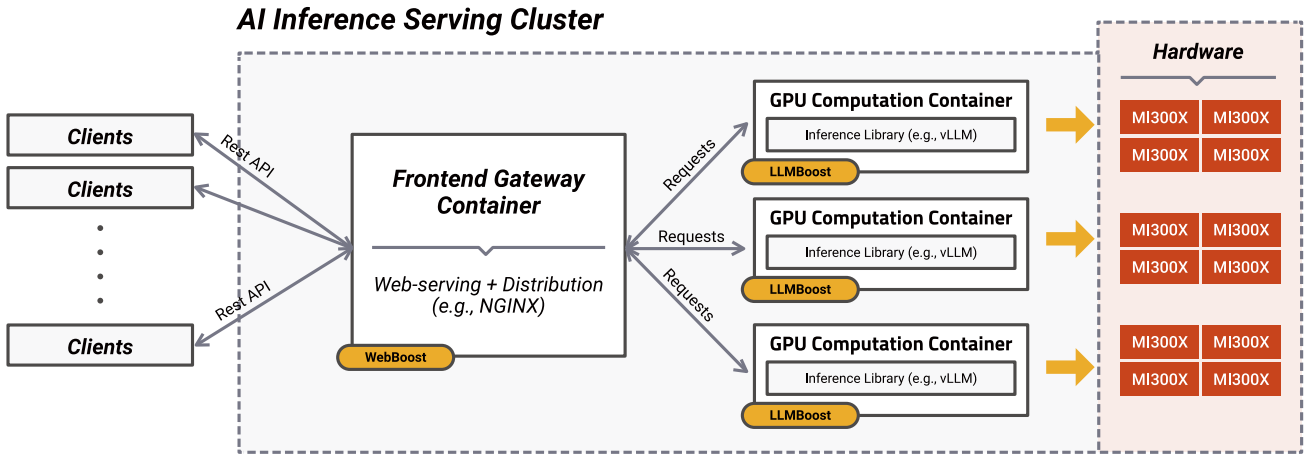

***FIGURE 3**: End-to-End Inference Serving Service*

- **Case Study with WebBoost**. To quantify the performance benefits of MangoBoost's WebBoost when integrated with Nginx, we conducted a series of benchmarks comparing our solution against a setup using Linux's standard TCP kernel software stack. TCP kernel software stack relies on the NVIDIA® ConnectX®-6 network card, which is widely used in typical data center environments.

- We utilized the wrk benchmarking tool to generate network requests to the Nginx web server. This tool is designed for measuring HTTP performance, making it ideal for evaluating the throughput and latency of our solution. In this evaluation, wrk sends HTTP POST requests to the Nginx web server. The requests contain sentences typically input by users of LLM AI inference serving. The Nginx then responds with sentences as part of the HTTP responses. The configuration details of our server hardware and software are in the table below.

Web-Serving Front-End Server Configuration

Component	Description
CPU	• 2 sockets, AMD EPYC™ 9534 64-Core Processor
Memory	• 24 x 96GB (total of 2,304GB) server DDR5
Network Card	• Linux TCP: NVIDIA® ConnectX®-6 • Mango WebBoost™: AMD Alveo™ U45N
Network Switch	• Dell EMC™ PowerSwitch S6100-ON
OS	• Ubuntu 22.04.4 LTS
Workload	• wrk: 512 connections, 32 threads • Nginx: 4 worker process • HTTP post and response size: 4KB

- The evaluation results are summarized in the graphs in Fig. 4.

> **Throughput**: WebBoost outperforms the Linux TCP stack significantly, achieving a 2x improvement in throughput. This means that Nginx can handle 100% more requests per second when using the WebBoost solution compared to the standard TCP stack.
>
> **Latency**: WebBoost also demonstrates 0.53x and 0.54x reductions on the 50th and 90th percentile latency, respectively.

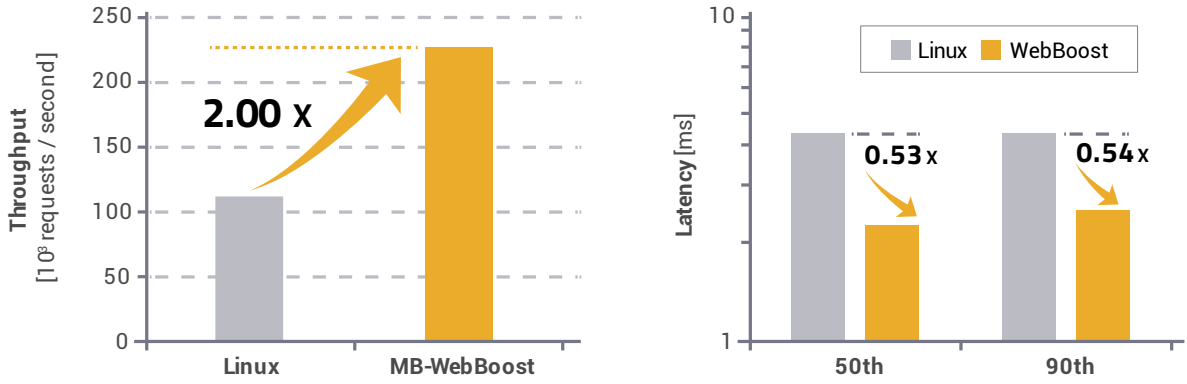

***FIGURE 4**: Throughput and Latency comparison between Linux and WebBoost*

- These results clearly illustrate the efficiency gains from offloading TCP packet processing to our DPU. By handling TCP connections more effectively, the WebBoost solution not only increases throughput but also significantly reduces the time it takes to process requests, leading to a more responsive and scalable Nginx deployment.
- Nginx is a cornerstone of modern web infrastructure, and optimizing its performance is crucial for maintaining the efficiency and reliability of web services. By offloading TCP packet processing to a DPU, we can significantly enhance Nginx's capabilities, allowing it to handle more traffic with lower latency. Our solution represents a leap forward in web server performance, setting a new standard for what is possible in high-demand environments.

Try our solutions today!

If you are interested in LLMBoost and WebBoost, please email contact@mangoboost.io.

References

[1] MLPerf Inference: Datacenter Benchmark Suite Results

[2] Unveiling MLPerf® Results on AMD Instinct™ MI300X Accelerators

[3] Smith, Alan, et al. "Realizing the AMD Exascale Heterogeneous Processor Vision: Industry Product." 2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA). IEEE, 2024.

[4] vLLM in AMD MLPerf Inference v4.1 submission

[5] AMD ROCm vLLM

[6] Result not verified by MLCommons Association.