Mango GPUBoost RDMA™ with MLPerf™ Training v4.1 on a Multi-Node System Powered by AMD MI300X GPUs

WHITE PAPER

REVISION 1.0 | Nov 14, 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

02 Introduction

03 Benefit #1 : Linear performance scale-up

04 Benefit #2 : Industry-standard and customizable features

01 | Key Points

- **Mango GPUBoost RDMA™** unlocks the power of state-of-the-art GPUs for AI training systems with multi-GPUs. As an example, MangoBoost demonstrates linear scaling using our RDMA solution, for MLperf Training v4.1 Llama2-70B-LoRA workload (BF16), using 24x AMD MI300X GPUs across multiple nodes.
- Mango GPUBoost RDMA™ supports **standard RoCEv2** to allow direct communication across GPUs to scale AI systems. Furthermore, MangoBoost HW-based in-line **customizable congestion control and custom headers** enable even more optimal solutions at scale.
- Mango GPUBoost RDMA™ is ready to ship. Pre-orders are open for customers who want to accelerate their AI infrastructure.
- Beyond RDMA/Roce-v2, MangoBoost is **a member of Ultra Ethernet Consortium**, and plans to offer an Ultra Ethernet solution in the near future.

02 | Introduction

- Efficient GPU-to-GPU communication across multiple nodes is essential for large-scale AI training workloads, where dozens to thousands of GPUs work together to train complex models. **RDMA (Remote Direct Memory Access)** has emerged as a key technology for enabling this scalability, as it allows a direct data transfer between GPU memory on different nodes without the help of the CPU host significantly reducing the latency of GPU-to-GPU data transfers. Among various RDMA solutions, **RoCEv2 (RDMA over Converged Ethernet)** has been shown to be a promising solution that offers unique advantages over traditional Infiniband for large-scale AI training environments[1].
- One of RoCEv2's primary advantages is its **scalability**. Because it uses an IP header for routing, RoCEv2 can scale over larger, more complex network topologies than Infiniband. This IP-based routing capability makes RoCEv2 especially appealing in hyperscale AI infrastructures, where scalability is crucial for managing massive AI models and multi-node GPU configurations.
- Another advantage of RoCEv2 is its **switch radix**, which directly impacts the physical layout of the AI networking infrastructure. Ethernet switches typically support more ports than Infiniband switches of the same generation, reducing the level of switches that introduces latency and potential oversubscription. With RoCEv2's use of high-port-count Ethernet switches, system architects can **create streamlined, low-latency network designs** that improve data flow across the entire GPU cluster.
- RoCEv2 also benefits from **a broad ecosystem, with multiple vendors supporting Ethernet-based solutions**. Unlike Infiniband, which is primarily produced by NVIDIA, RoCEv2 gives hyperscalers and enterprises the flexibility to choose from a variety of vendors, helping them to avoid vendor lock-in and maintain more control over their supply chain.

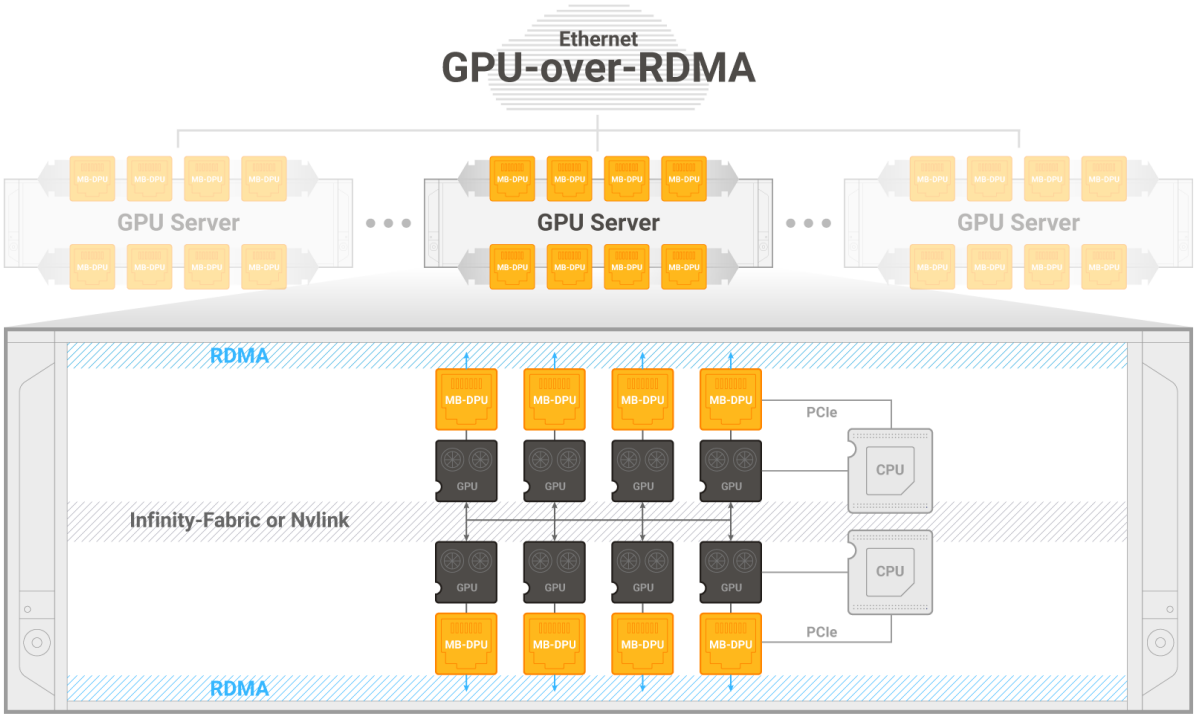

***FIGURE 1**: Large-scale AI Training system using MangoBoost DPU*

- Figure 1 shows how MangoBoost set up a large-scale AI Training system using MangoBoost DPUs. Each server is equipped with eight MangoBoost DPUs and eight GPUs, linked via PCIe for GPU-NIC communication. In this system, high-bandwidth interconnects like AMD's Infinity Fabric or NVIDIA's NVLink manage intra-node communication. For inter-node communication, MangoBoost DPUs enable RoCEv2, connecting multiple nodes over Ethernet switches and creating a high-performance and scalable setup optimized for large-scale AI workloads.

03 | Benefit #1 : Linear performance scale-up

- Our Llama-70B LoRA training leveraged from 8 to 24 MI300X GPUs with MLPerf datasets and was conducted in full compliance with the MLPerf Training v4.1 standards using BF16 precision (code provided by MLPerf). The setup focused on **maximizing GPU-to-GPU communication efficiency to boost throughput**. We used AMD's Infinity Fabric for intra-node communication, while inter-node communication was powered by RoCEv2 RDMA, which is crucial for seamless scaling across multiple servers.

Server Configuration

Component	Description
CPU	• 2 x Intel® Xeon® Gold 6438N
Memory	• 16x 32GB DDR5 (512 GB in total)
GPU	• 8 x AMD Instinct™ MI300X
Network Switch	• Dell PowerSwitch Z9432F-ON
GPU-to-GPU Interconnect	• AMD Infinity Fabric™ Link
Network Card	• 8x MangoBoost DPU cards (HHHL)

- To ensure high-efficiency RoCEv2 performance, we installed MangoBoost DPUs, assigning one DPU per GPU in each server. This **1:1 mapping enabled optimized data transfer across nodes**, taking advantage of RDMA's capabilities to **support high-speed distributed training**. Each server used an identical configuration, providing a robust and scalable infrastructure capable of handling the intense demands of large language model (LLM) training.

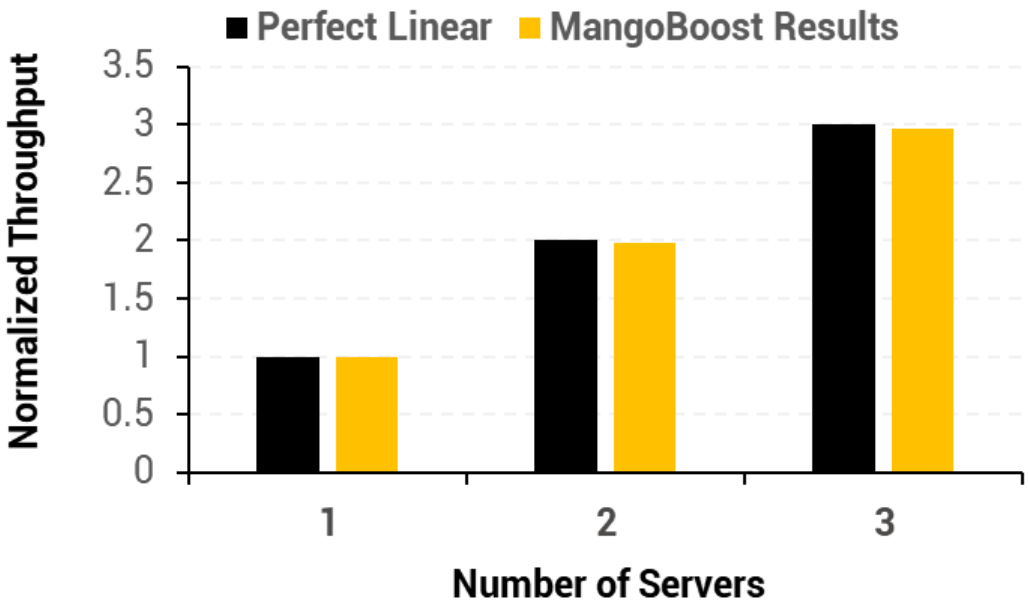

***FIGURE 2**: Scalability test: 1-3 GPU servers (8x AMD MI300X GPUs per server), running Llama-70B LoRA training*

- Our scaling tests demonstrated **near-linear performance scaling** as we increased the number of servers, which is paramount in large-model training where hardware limitations and network overhead often constrain scalability. With our setup, we maintained nearly optimal performance even as additional resources were added.
- Throughout the process, **we ensured strict adherence to the MLPerf v4.1 validation requirements**. This included following guidelines for hyperparameters, evaluation methodology, and quality targets, such as achieving a 0.925 evaluation loss threshold. We validated our setup against the Reference Convergence Point (RCP) rules, ensuring accuracy and reliability in producing high-quality training results for large language models.

04 | Benefit #2 : Industry-standard and customizable features

- Mango GPUBoost RDMA™ is a **high-performance RoCEv2-based RNIC solution** that unleashes the full potential of ethernet-based large-scale AI workloads. It provides greater network scalability alongside peer-to-peer communication between GPU and RNIC (e.g., GPUDirect, ROCmRDMA) to enable optimized connectivity among GPUs at scale.

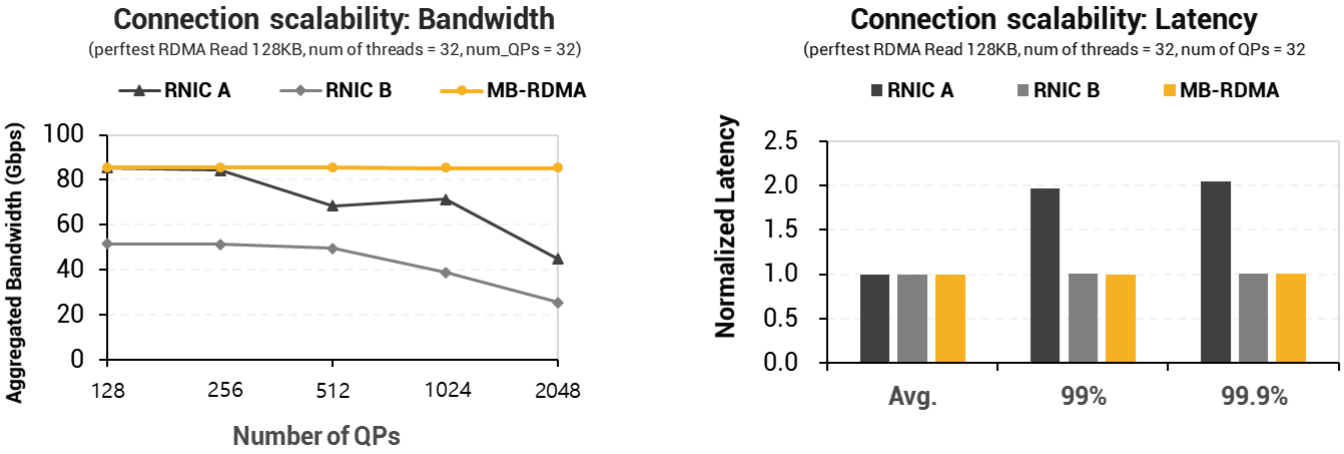

***FIGURE 3**: Scalability test: provide better throughput and latency with many QP connections*

- Mango GPUBoost RDMA™ **achieves impressive performance, sustaining 90% of line-rate throughput** even as the number of QP (Queue Pair) connections increases. It also delivers **significantly lower latency** compared to other standard RDMA network interface cards (RNICs) available on the market. Figure 3 demonstrates these advantages using results from Perftest, a widely-used RDMA benchmarking tool, where Mango GPUBoost RDMA™ achieves 3.4 times higher throughput and reduces 99.9% latency by 52%.

- Mango GPUBoost RDMA™ also provides **two important features (multipath and custom congestion control)** to resolve well-known problems within RoCEv2.

  > First, due to the nature of the RoCEv2 protocol — which enforces packets to be delivered in the order they were sent — network hotspots often occur because of path collision [2]. To alleviate these hotspots and take advantage of multiple rich paths, Mango GPUBoost RDMA™ supports a **unique out-of-order packet handling mechanism and provides multipath extension** with custom headers.

  > Moreover, no single congestion control solution fits all use cases for RoCEv2. Although DCQCN has become the de-facto standard for RoCEv2, it is hard to configure in practice [3] and has been shown to be less effective for AI workloads [4]. Instead, each workload demands a tailored congestion control algorithm depending on its unique characteristics [5] (e.g., traffic pattern, incast, etc). To maximize workload performance, Mango GPUBoost RDMA™ offers a **placeholder for custom congestion control**, allowing users to deploy their own algorithms directly on our RNIC.

- We are ready for customers who want to innovate their AI infrastructure by accelerating network communication with Mango GPUBoost RDMA™. Our DPUs have been tested on various AI training and inference workloads with top-of-the-line GPUs, demonstrating its flexibility along with industry-leading performance. We support standard RoCEv2 with customizable features and compatibility with common server products.
- We continue to innovate performance and scalability in AI infrastructure. As a member of the Ultra Ethernet Consortium, we also plan to showcase our Ultra Ethernet DPU in the near future.

References

[1] RDMA over Ethernet for Distributed Training at Meta Scale (Meta, SIGCOMM'24)

[2] Multi-Path Transport for RDMA in Datacenters (Microsoft, NSDI'18)

[3] HPCC: High Precision Congestion Control (Alibaba, SIGCOMM'19)

[4] Impact of RoCE Congestion Control Policies on Distributed Training of DNNs (Meta, HOTI'22)

[5] Datacenter Ethernet and RDMA: Issues at Hyperscale (Microsoft/Broadcom/HPE/Google, IEEE Computer'23)