



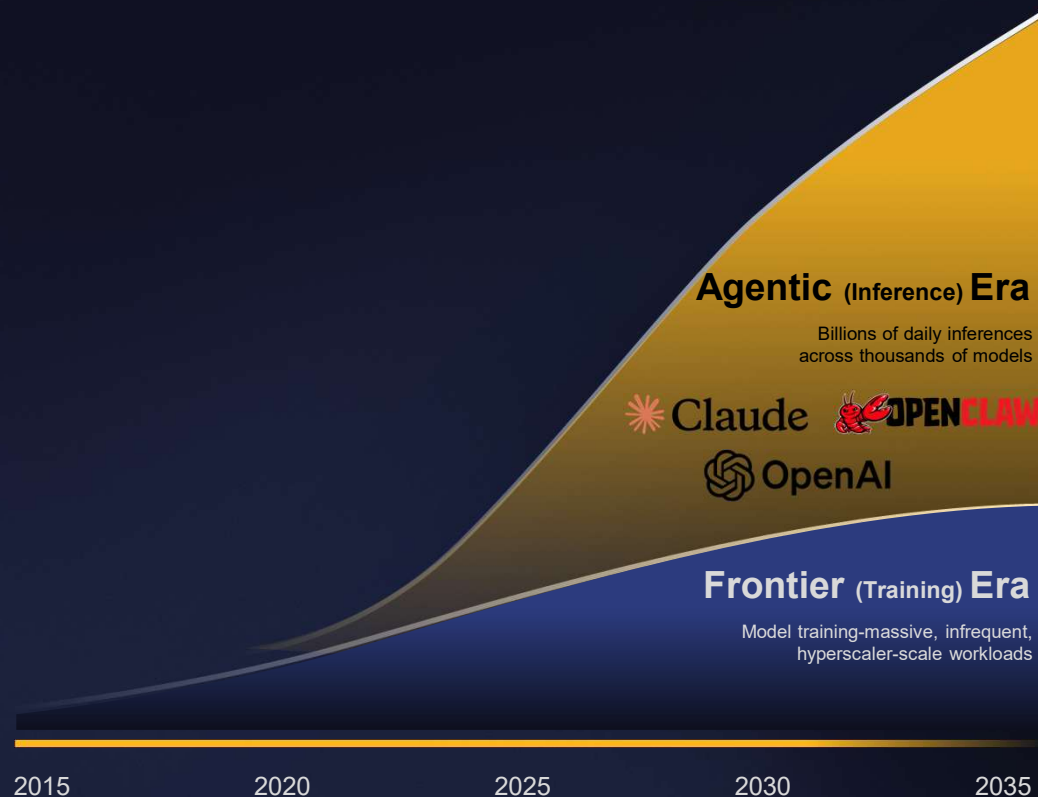
Architecting the Agent Factory for Scalable AI Intelligence



The Age of “Agentic AI” Has Begun



~80% of AI compute cycles in inference by 2028



Source: Gartner

© 2026 MangoBoost, Inc. All rights reserved. Do Not distribute without permission.

Driven by the rapid deployment of Agentic AI

FORTUNE

"AI Agents could automate 57% of all U.S. work hours" (Nov 2025)

McKinsey & Company

"SaaS Era is ending, Digital Workers Emerging" (Dec 2025)

ANTHROPIC

"Claude Code adoption skyrocketed by 400% in Jan 2026, shifting focus from 'Writing Code' to 'Agentic AI'" (Jan 2026)

Microsoft

"80% of Fortune 500 use active AI Agents" (Feb 2026)



"OpenClaw Did in 3 Weeks What Linux Took 30 Years to Achieve" (Mar 2026)

yahoo! finance

"AI agents will replace work performed by humans on a massive scale" (Mar 2026)

Forbes

"The Next Big AI Shift is Agentic AI-as-a-Service" (Apr 2026)

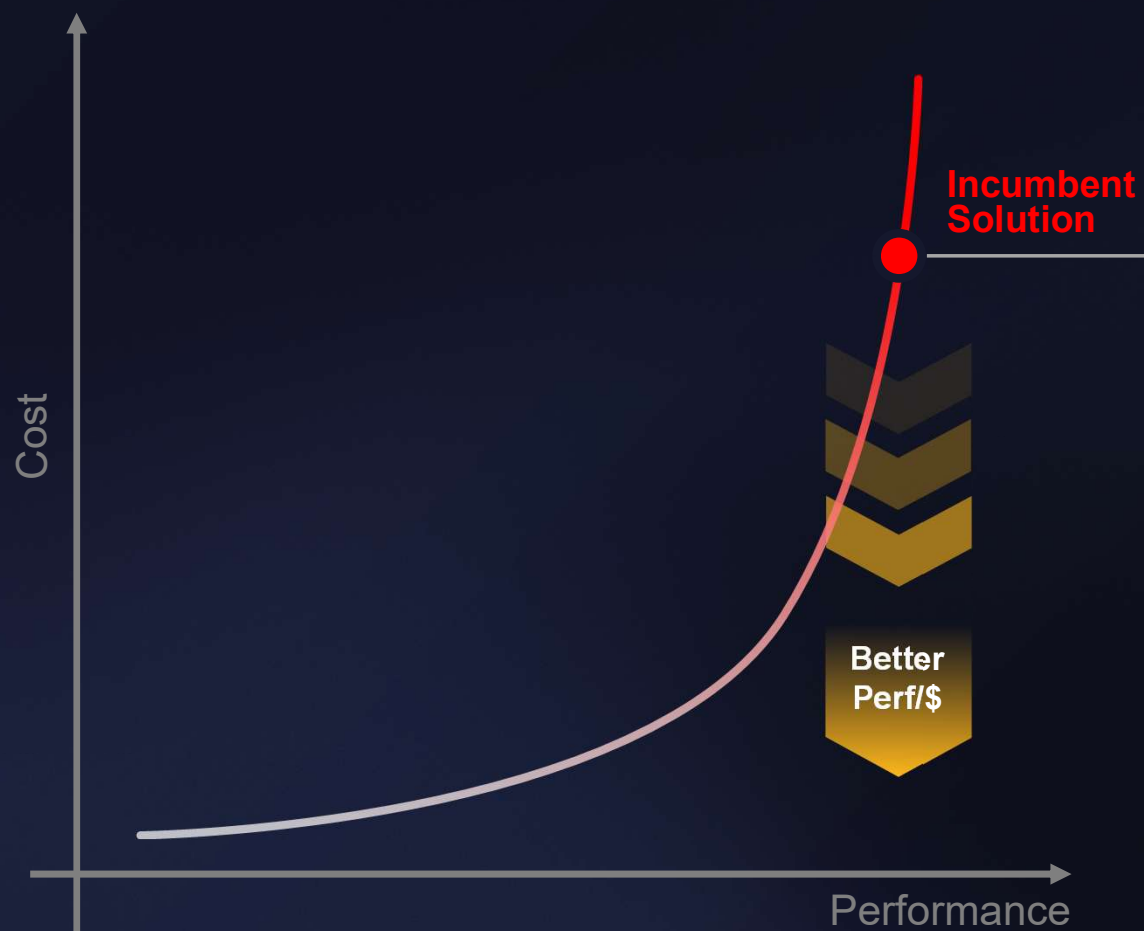
Agentic AI Workload Changes the Rules



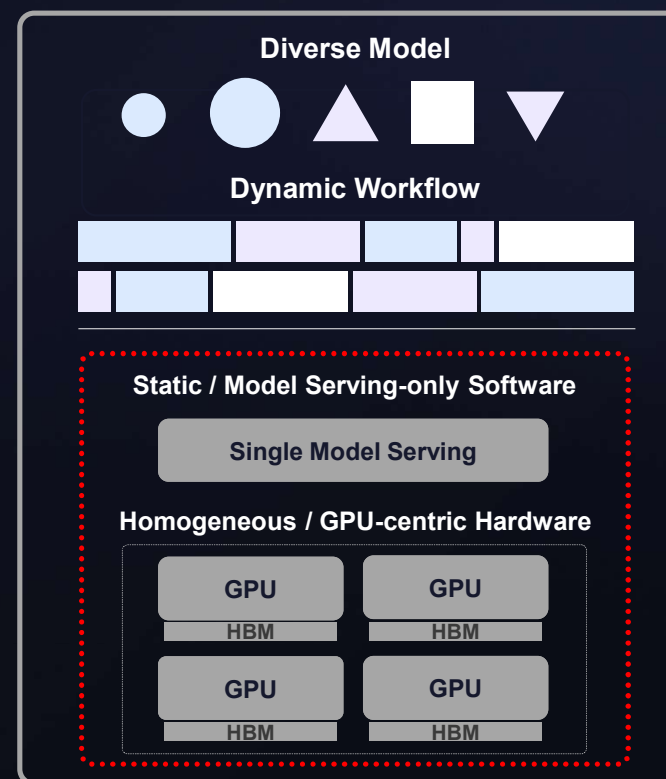
Source: Intel, Cerebras

© 2026 MangoBoost, Inc. All rights reserved. Do Not distribute without permission.

The Limits of Current Infrastructure



Hard to Scale Perf / Cost Efficiency

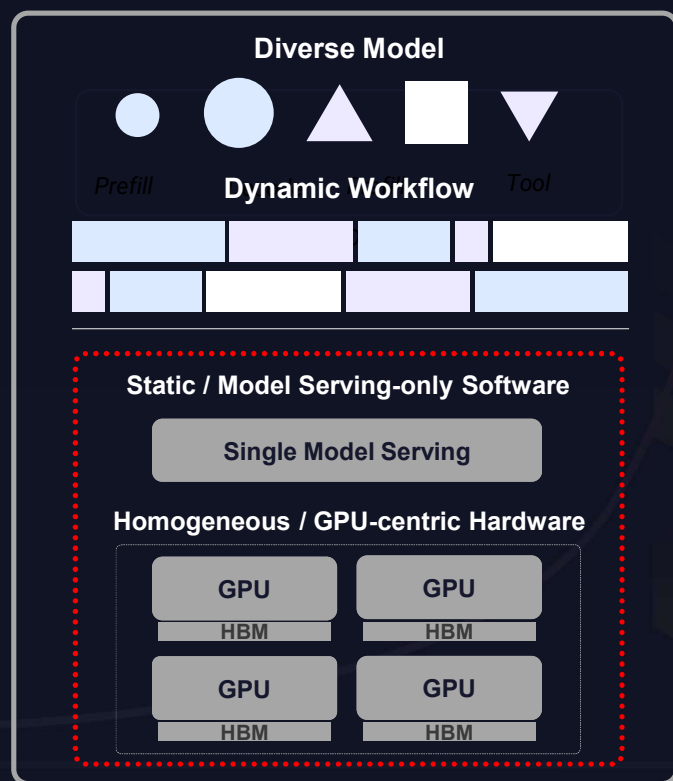


100-1000× more tokens, same infra = economically broken

Note: Not actual numbers, approximate tokens per task based upon external sources and internal estimates

© 2026 MangoBoost, Inc. All rights reserved. Do Not distribute without permission.

Three Requirements of Agentic AI Infrastructure



① AI Agent OS

*Not Static Model-based Serving Only;
OS can understand more diverse models and dynamic workflows*

② Heterogeneous Compute

*Not Just a Flagship GPU for all;
More cost-effective options*

③ Agent Memory Expansion

*Not Just HBM-only;
More extended memory hierarchy*

Note: Not actual numbers, approximate tokens per task, based upon external sources and internal estimates.

MangoBoost Solution for Agentic AI Infrastructure



Enterprise
Agentic AI

MANGOBOOST
Orchestrated
End-to-End Stack

HW-SW
Codesigned,
Unified System



① AI Agent OS

Dynamic Orchestration natively aware of Agentic AI workloads



Agent Orchestrator | Workflow & Planning

Agent Runtime | Model Serving & Tooling
& Context Mgmt. & Guardrail

Hardware Interface | Optimized Kernel & Fabric

**Infrastructure
Observability**

HW-SW Codesign

Agent Coordination Offload
HW-aware Tuning

② Heterogeneous Compute

The right accelerator for each step

Flagship GPUs
Heavy Decode

Value GPUs
Prefill

NPU
Light Decode

CPUs
Tool/Sandbox

Unified Architecture

Open Ethernet Network
Compute + Storage Fabric

③ Agent Memory Expansion

HBM/DRAM/Storage(SSD/HDD) as a tiered Agent memory pool

HBMs | Live Context

DRAMs | Short-term Context

SSDs/HDDs | Long-term Context & Persistent

Validation.1 AI Workload Orchestration & Optimization



Superior Token Performance & Auto-tuning in Action

Performance Results: Auto-tuning & Simple Orchestration



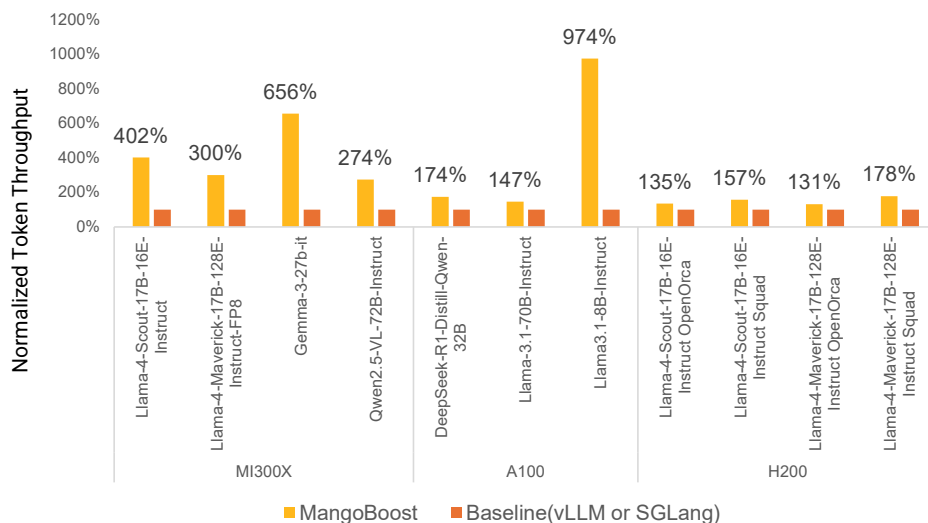
Maximum Device Flexibility



Multi-model Deployment and Management



Hassle-free Deployment



Performance Results: Next-gen GPU's Token Throughput & TTFT

Deployment Setup

GPU : On 4x MI355X GPU
Model : Deepseek v3.2

Baseline : vLLM v0.14.0
Dataset : OpenOrca

Generation Throughput (tok/s)

800
600
400
200
0



■ vLLM ■ LLMBoost

161%
Higher (better)

TTFT Latency (s)

100
80
60
40
20
0



■ vLLM ■ LLMBoost

416.29x
Lower (better)

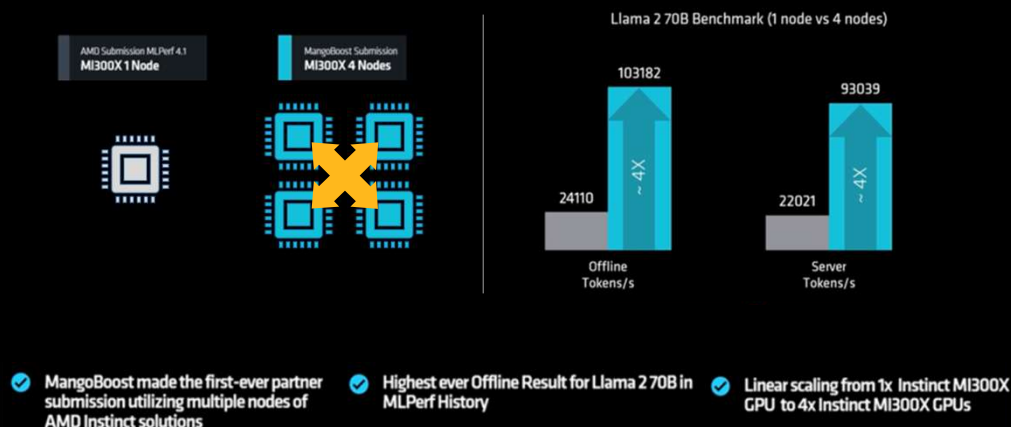
Validation.2 Heterogeneous Compute Clustering



MLPerf Benchmark: GPU Multi-node Scaling & 3-GPU Integration

MLPerf Inference 5.0

Partner Showcases Multi-Node Performance with AMD Instinct™ GPUs

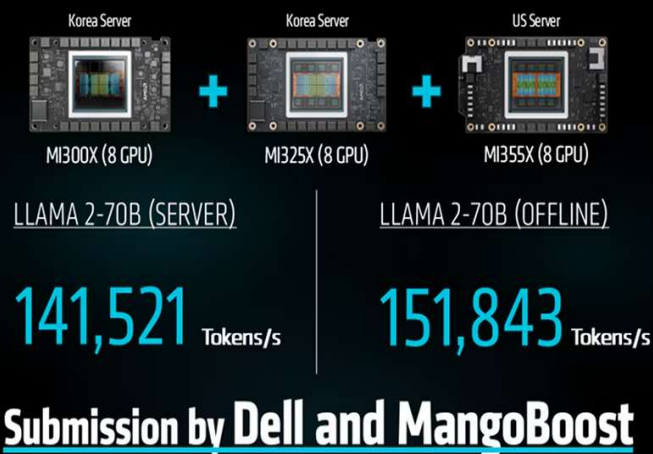


MANGOBOOST + AMD MI300X
vs
JUNIPER NETWORKS + NVIDIA H100

**~ 3X Higher
Cost Efficiency**

MLPerf Inference 6.0

First 3 GPU MLPerf Heterogeneous Submission AMD Instinct™ MI300X, MI325X and MI355X GPUs



Submission ID: 6.0-0031

Demonstrates Flexible Inference Across Geographies



"An especially important detail is geography. **The AMD Instinct MI355X platform was located in Dell's lab in the United States, while the Instinct MI300X and MI325X platforms were in Korea.** That makes this more than a mixed-generation inference story, it is also a proof point for orchestration across systems in different geographies."

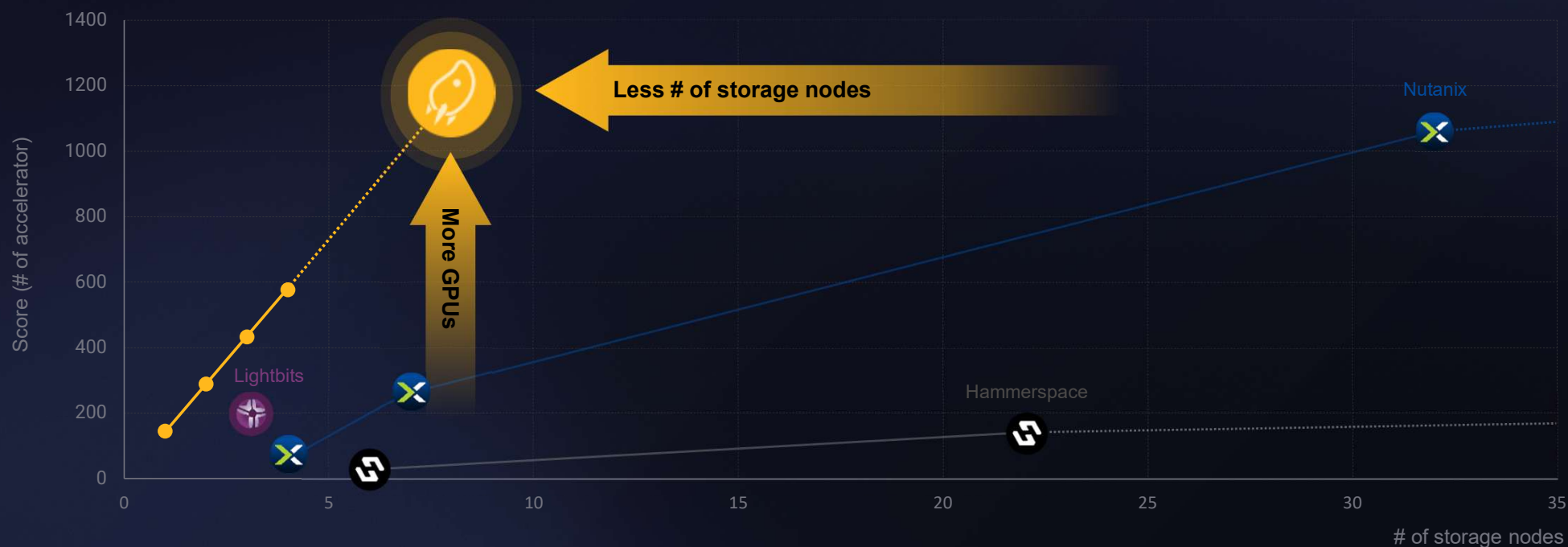
Validation.3 Fabric-accelerated AI Storage System



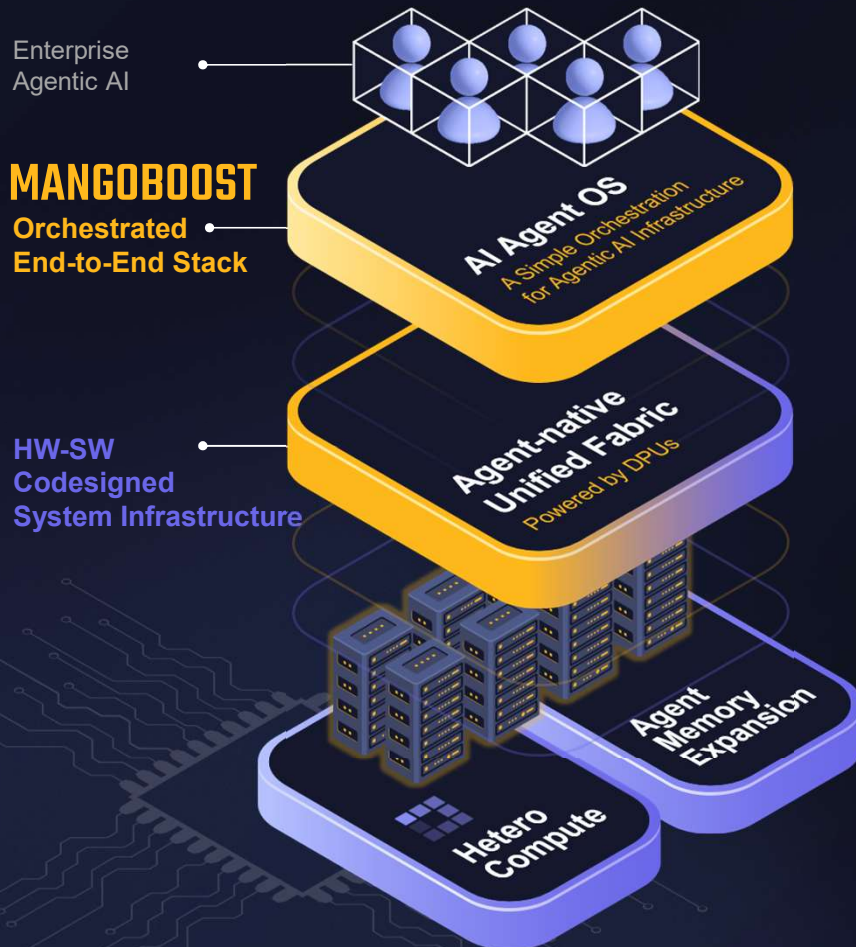
MLPerf Benchmark: Storage as a GPU Memory

MLPerf Storage v1.0

No. 1 MANGOBOOST



Values of MangoBoost Agentic AI Infrastructure

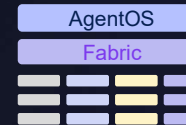


© 2026 MangoBoost, Inc. All rights reserved. Do Not distribute without permission.

① Configurable

Purpose-built for every AI agent

Agentic System for
Coding



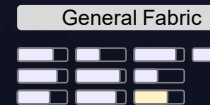
Agentic System for
Personal Assist



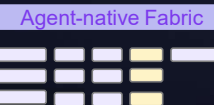
② Cost-effective

Maximum utilization, minimum waste

Static Scheduler



Dynamic AgentOS



③ Trustworthy

*Zero-trust security & safety
built into the hardware*

Software Guardrail

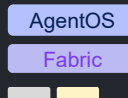
Hardware Guardrail

High Latency
with CPU occupation

Low Latency,
freeing CPU for app.

Scalable Agent Intelligence

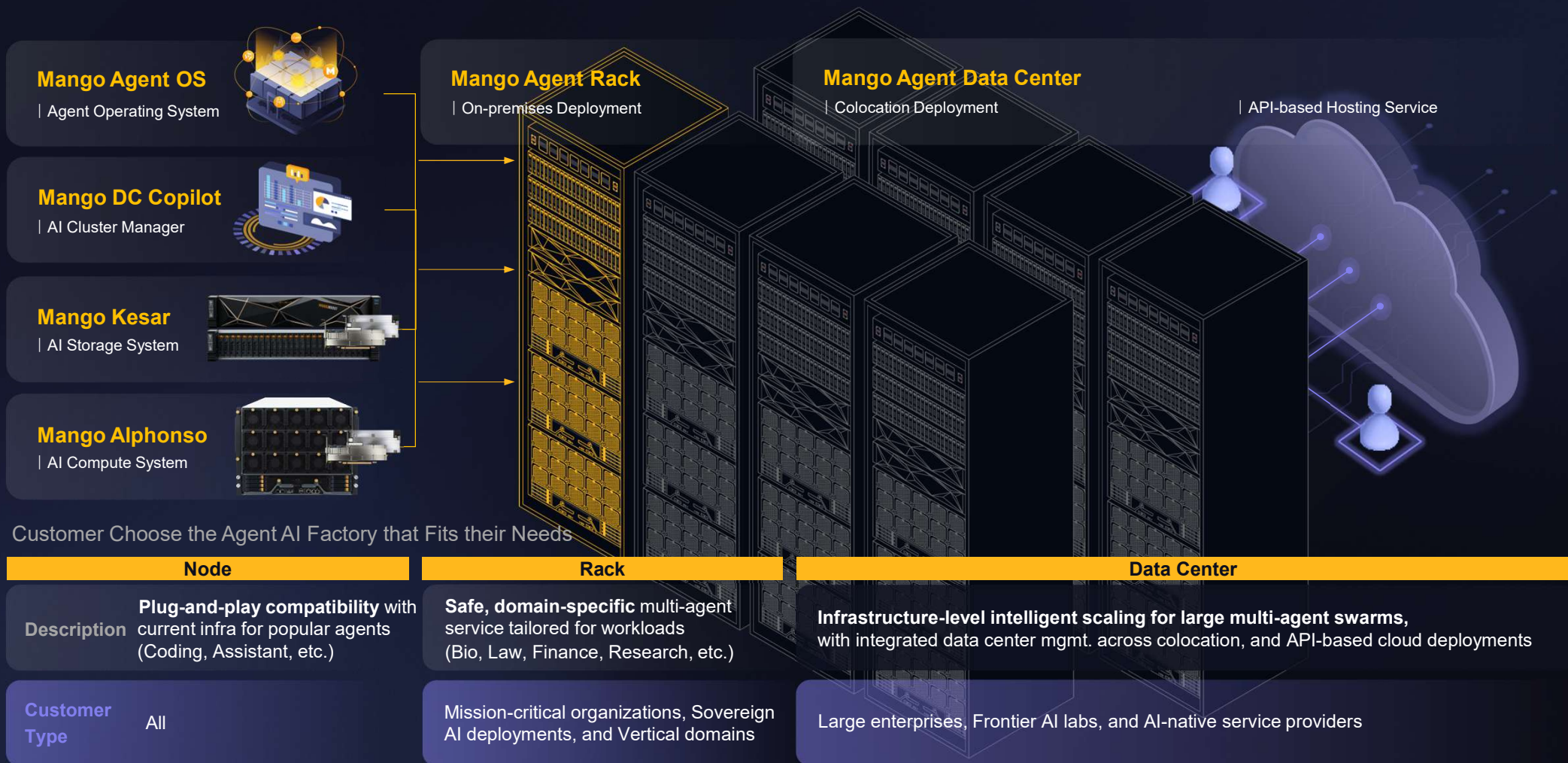
10 Agents



Scale, more intelligence
1000 Agents



Product Portfolio: Agent AI Factory



MangoBoost mini-AI Factory



World's Smartest & Most Cost-effective AI DC



MANGOBOOST AI SERVER VIRTUAL DEMO [REGISTER FOR DEMO](#)

End-to-End AI, Optimized by MangoBoost

Del. 35 AI at Lightning Speed From Power On to Production

[Try a Demo](#)

**NVIDIA
Comparison Metrics**
Start your trial request instantly

**CapEx & OpEx
Savings Breakdown**
Our engineers will configure your workload environment

**Optimal
Deployment Sizing**
24-hour access to your AI server testbed



Collaborating with Best-in-Class AI Ecosystem Partners



First 3 GPU MLPerf Heterogeneous Submission AMD Instinct™ MI300X, MI325X and MI355X GPUs



LLAMA 2-70B (SERVER)

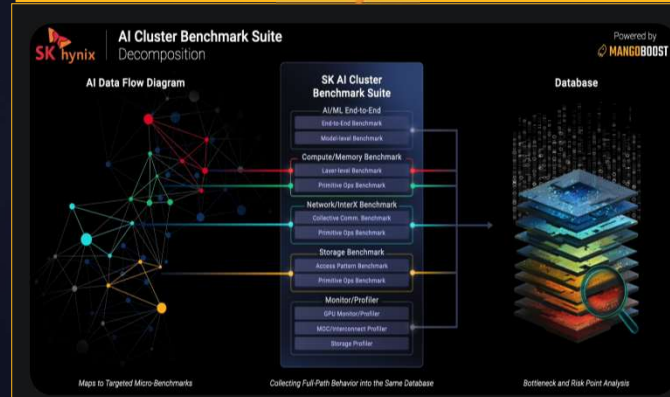
141,521 Tokens/s

LLAMA 2-70B (OFFLINE)

151,843 Tokens/s

Submission by Dell and MangoBoost

SubmissionID: 6.0-0031



SUPERMICRO GPU SERVERS WITH MANGOBOOST LLMBOOST™ AI ENTERPRISE SOFTWARE: UNLOCKING THE FULL POTENTIAL OF AMD INSTINCT™ GPUS

Combination Delivers a Cost-Effective, Easy-to-Use, High-Performance, Flexible,
and Scalable Multiple-Node GenAI Server Solution



Dell PowerEdge XE9680 Server and AMD Instinct™ MI300X GPUs Deliver Superior MLPerf™ Training v5.0 Performance

July 2nd, 2023 | Read Time: 7 minutes | MS Mangreet Sokhi | NY Frank Han

Overview

MangoBoost's LLMBoost™ AI Enterprise Software

MangoBoost's LLMBoost™ software complements Dell's PowerEdge XE9680 server with AMD Instinct GPUs with a turn-key full-stack AI MLops platform. Unlike the open-source frameworks that are typically used in MLPerf submissions, Mango LLMBoost provides the following benefits:

**LLMBoost™ AI
Enterprise Software
(Patent Pending)**
Auto-tuning, scaling, etc.

• Turn-key deployment - Mango LLMBoost is packaged as pre-built Docker images. An intuitive CLI helps developers launch LLM workloads quickly on any GPU.



Management Team & Major Investors



CEO

Jangwoo Kim

- CMU Ph.D.
- SNU Professor
- Hall of Fame ISCA&MICRO



CAIO

Eriko Nurvitadhi

- CMU Ph.D.
- Intel Principal Engineer



CTO

Dongup Kwon

- SNU Ph.D.
- Microsoft Ph.D. Fellow



CFO & COO

Junki Park

- SNU Ph.D.
- Samsung Ventures
- Samsung Economic Research Inst.



CCO

Chanik Park

- SNU Ph.D.
- Samsung Semiconductor Executive VP



VP of Business Development

Tom Spencer

- 30+ Year Industry Veteran
- Intel/Altera · Xilinx · Marvell · Broadcom



Major Investors

SAMSUNG

Carnegie Mellon University
Investment Office



Bon Angels

AION
MANAGEMENT



kt



DSC investment

STONEBRIDGE

Helios
PRIVATE EQUITY

KB Investment

About MangoBoost



Where AI Infrastructure 2.0 Begins — Built for the Agentic AI Era

MANGOBOOST INC. (Delaware C-corp)

Founded in 2022

Bellevue (HQ), San Jose, Toronto, Seoul, Tokyo, Indonesia



50+ Patents

10+ Years of R&D



120+ Engineers

Top-notch Engineers with PhDs



\$100M+ Raised

20+ Global Investors

