MANGOBOOST

ISFPGA 2026

MangoBoost

Democratizing FPGA-Accelerated Infrastructure for the AI Era: The MangoBoost Perspective

Eriko Nurvitadhi, PhD, MBA
Chief Product Officer & Co-Founder

Website www.mangoboost.io
Contact contact@mangoboost.io

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

MANGOBOOST

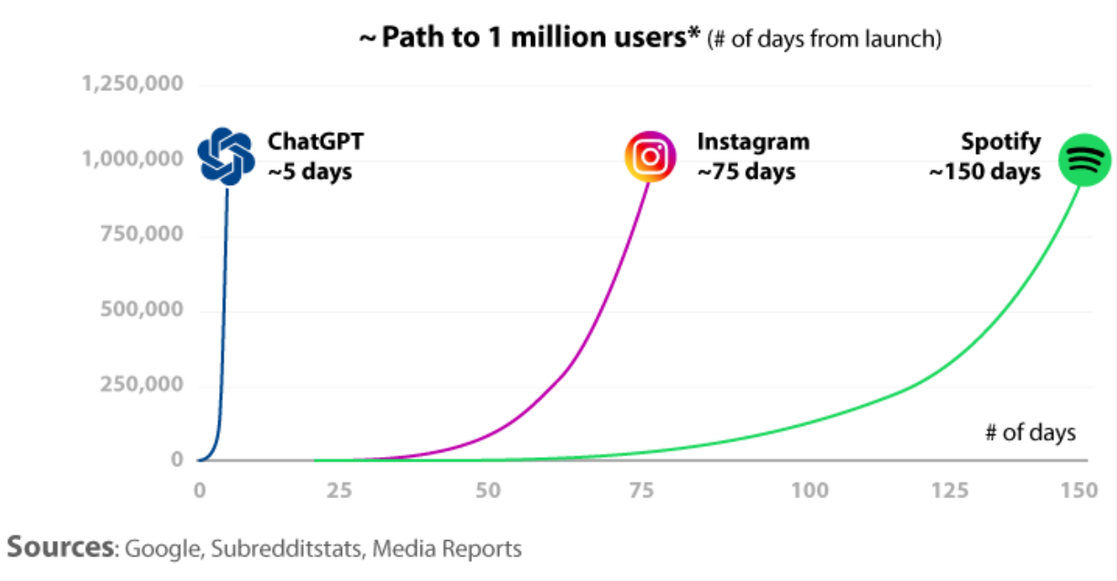
PART I.

AI Era and Key Bottlenecks

We are in the AI Era

MANGOBOOST

ChatGPT: The first “Killer App” for AI



Sources: Google, Subredditstats, Media Reports

Leads to Generative AI Boom

ChatGPT

deepseek

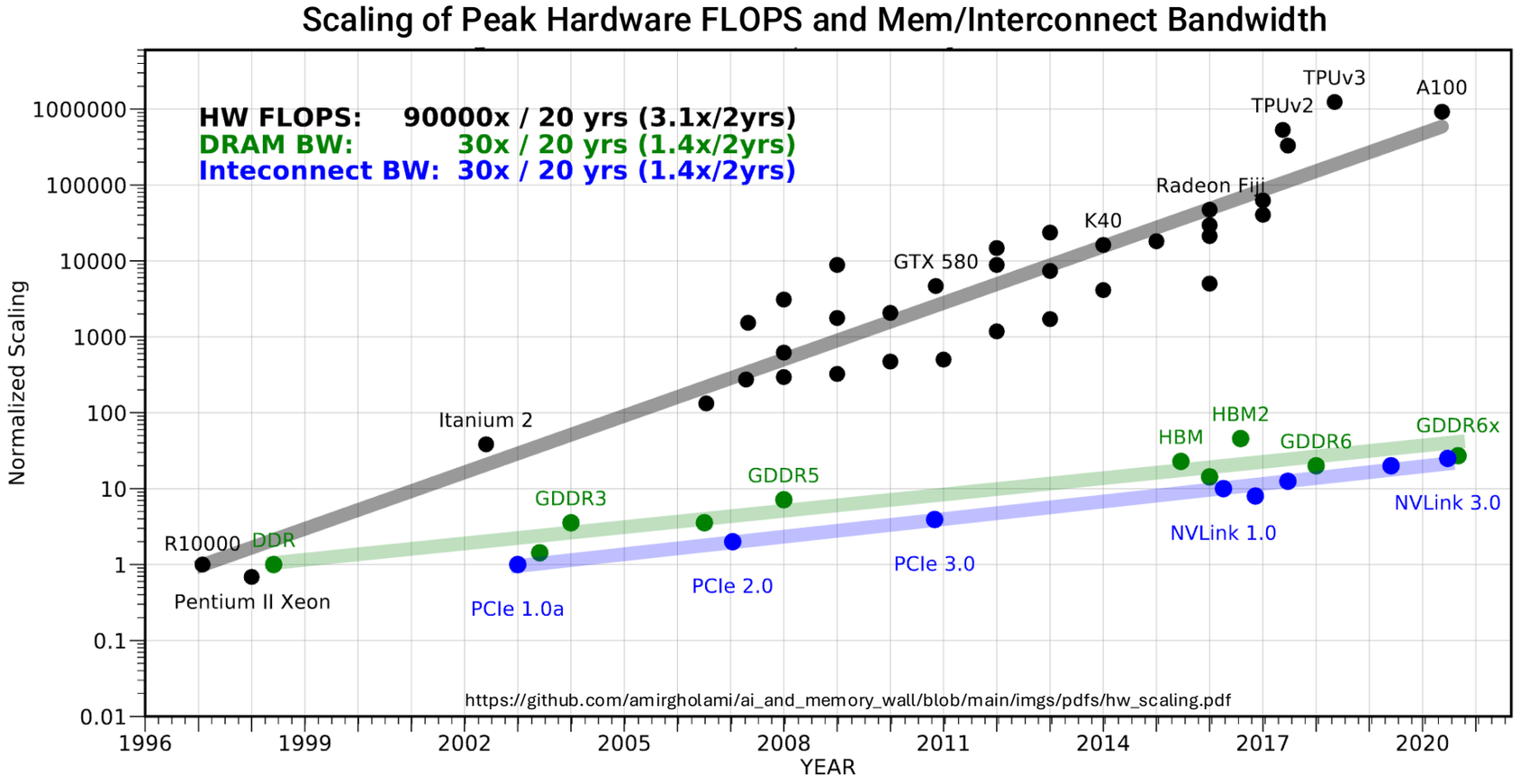
Gemini

LLAMA 3

Grok

Claude

Challenge: Need More Compute?



90,000x HW compute growth in past 20 years
However, mem & interconnect bandwidth grow only by ***30x***

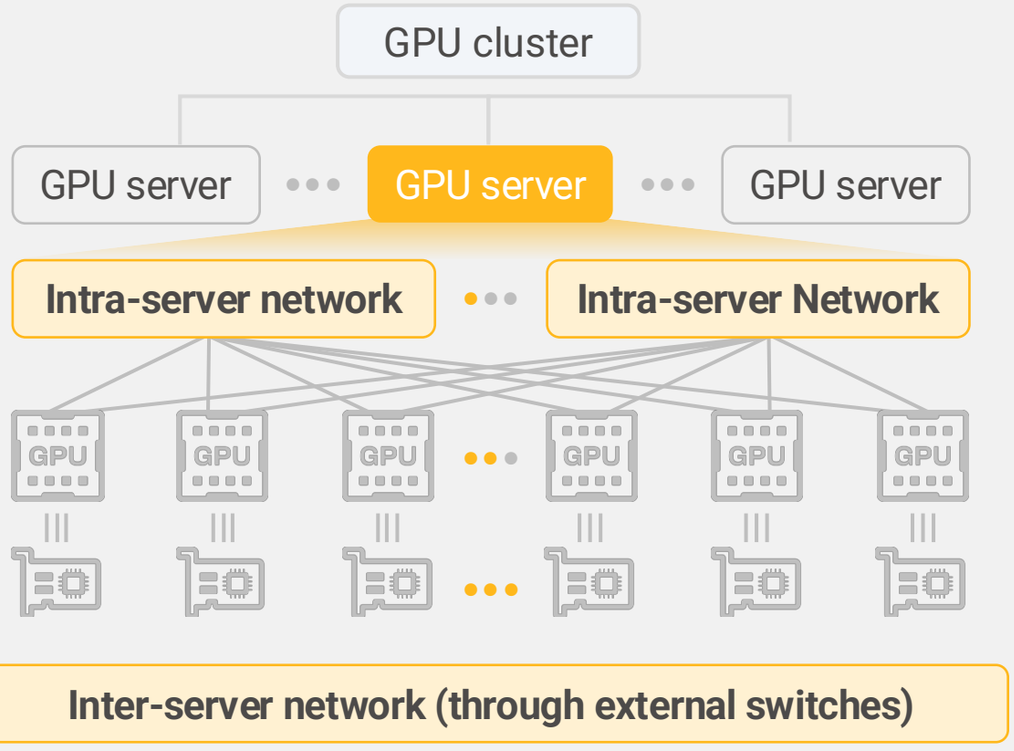
The Real Problem: Big AI needs AI Infrastructure

MANGOBOOST

Data movement: across GPUs, storage, network

"DeepSeek's modular, distributed AI training will likely drive demand for efficient networking solutions"

- Source : Forbes



Growing needs for Connectivity

"AI applications are ***bottlenecked by … communication across/to AI accelerators***, rather than compute."

- Amir Gholami, International Computer Science Institute, UC Berkeley

AI Infra Scalability Needed for (Pre)Training & Inference

MANGOBOOST

Inference beyond a single GPU (e.g., Disaggregated P/D)

MoneyWatch

What is DeepSeek, and why is it causing Nvidia and other stocks to slump?

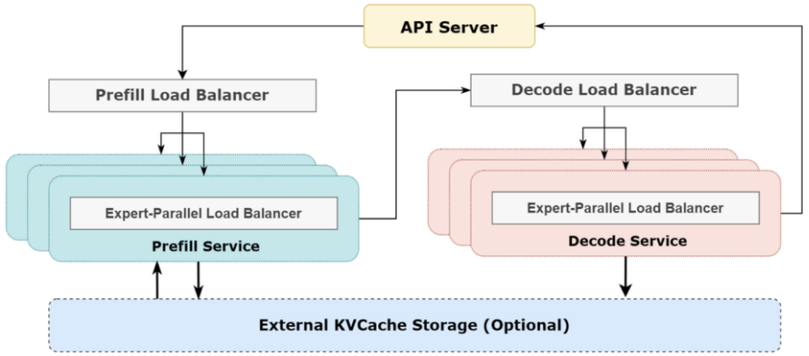
By Aimee Picchi

Updated on: January 28, 2025 / 12:29 PM EST / CBS News

Preview Code Blame Raw

Day 6: One More Thing, DeepSeek-V3/R1 Inference System Overview

Diagram of DeepSeek's Online Inference System



https://github.com/deepseek-ai/open-infra-index/blob/main/202502OpenSourceWeek/day_6_one_more_thing_deepseekV3R1_inference_system_overview.md

Training Continues to Need More GPUs

Meta is using more than 100,000 Nvidia H100 AI GPUs to train Llama-4 — Mark Zuckerberg says that Llama 4 is being trained on a cluster "bigger than anything that I've seen"

News By Jowi Morales published October 31, 2024

Llama 4 slated to have new modalities, stronger reasoning, and faster performance

Elon Musk says xAI is targeting 50 million 'H100 equivalent' AI GPUs in five years — 230k GPUs, including 30k GB200s already reportedly operational for training Grok

News Analysis By Anton Shilov published July 23, 2025

But how many nuclear power plants will it require?

....

MANGOBOOST

PART II.

FPGA opportunities in AI

First, Can FPGA Beat GPU in AI Compute?

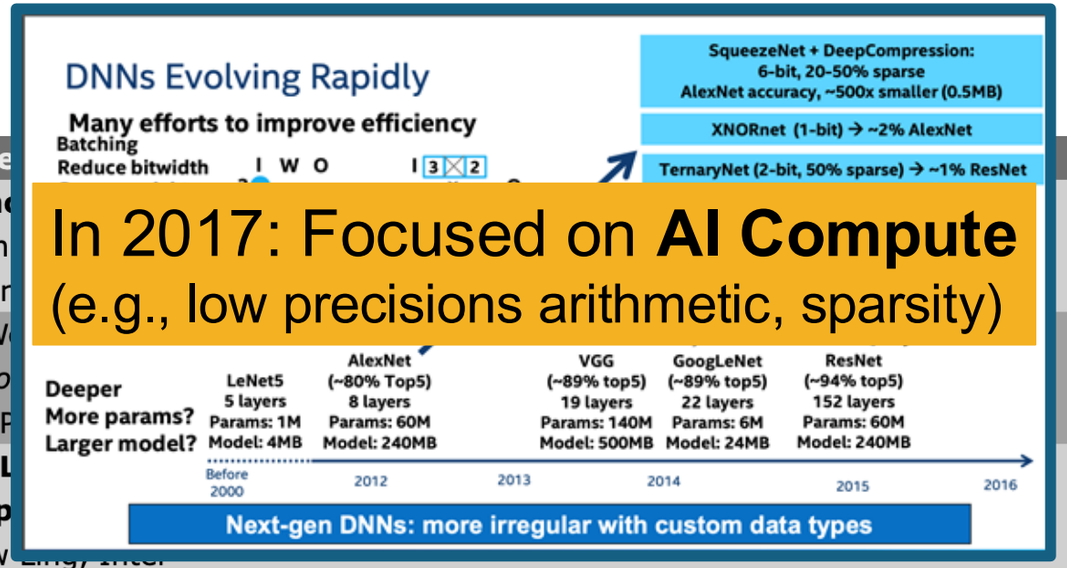
MANGOBOOST

FPGA 2017: Program

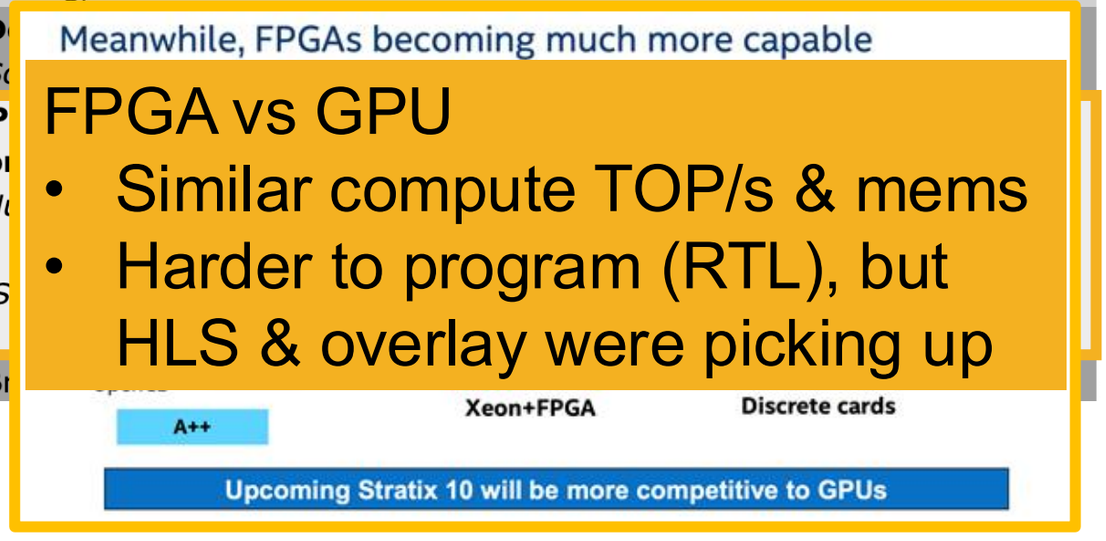
Wednesday February 22 (All Technical Sessions in San Carlos 2-4)	
Int'l Workshop on Overlay Architectures for FPGAs (OLAF)	
Chair: Hayden So, The University of Hong Kong	
Co-Chair: John Wawryznek, UC Berkeley	
9:00 - 9:10	Welcome and Opening Remarks *John Wawrzynek (UC Berkeley)*
9:10 - 12:00	Paper Presentations (http://olaf.eecs.berkeley.edu/program)
12:00 - 1:30	**Lunch (Ferrantes Room, 10th Floor)**
Afternoon Special Session: The Role of FPGAs in Machine Learning	
Chair: Andrew Ling, Intel	
1:30 - 2:30	**Deep Learning -- Tutorial and Recent Trends** [slides] *Song Han (Stanford and DeePhi)*
2:30 - 3:00	**Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks?** [slides] *Eriko Nurvitadhi, Ganesh Venkatesh, Jaewoong Sim, Debbie Marr, Randy Huang, Jason Gee Hock Ong, Yeong Tat Liew, Srivatsan Krishnan, Duncan Moss, Suchit Subhaschandra, Guy Boudoukh* Intel
3:00 - 3:30	Break

First, Can FPGA Beat GPU in AI Compute?

MANGOBOOST



In 2017: Focused on **AI Compute**
(e.g., low precisions arithmetic, sparsity)



FPGA vs GPU
- Similar compute TOP/s & mems
- Harder to program (RTL), but HLS & overlay were picking up

2026: Built-in AI HW compute everywhere!

GPUs (TensorCores), CPUs (Intel AMX), FPGAs (AIE, AITB), AI ASICs

GPU has highest compute. Why?
Business vs technical: Can you pay for big silicon @ latest process tech?

Low-precision is already common in GPUs (FP8, FP4, ..)
Not much more opportunity for further lower precisions on FPGA

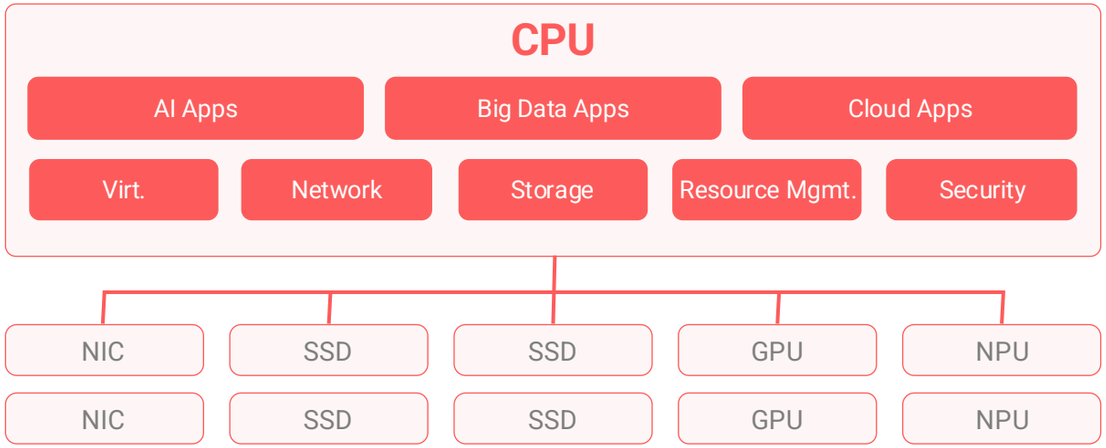
Sparsity support also common
(structured only, but decent enough)

AI software: FPGA is still lacking ☹

FPGA's Key Opportunity: Data Processing Units (DPUs) for AI Infra

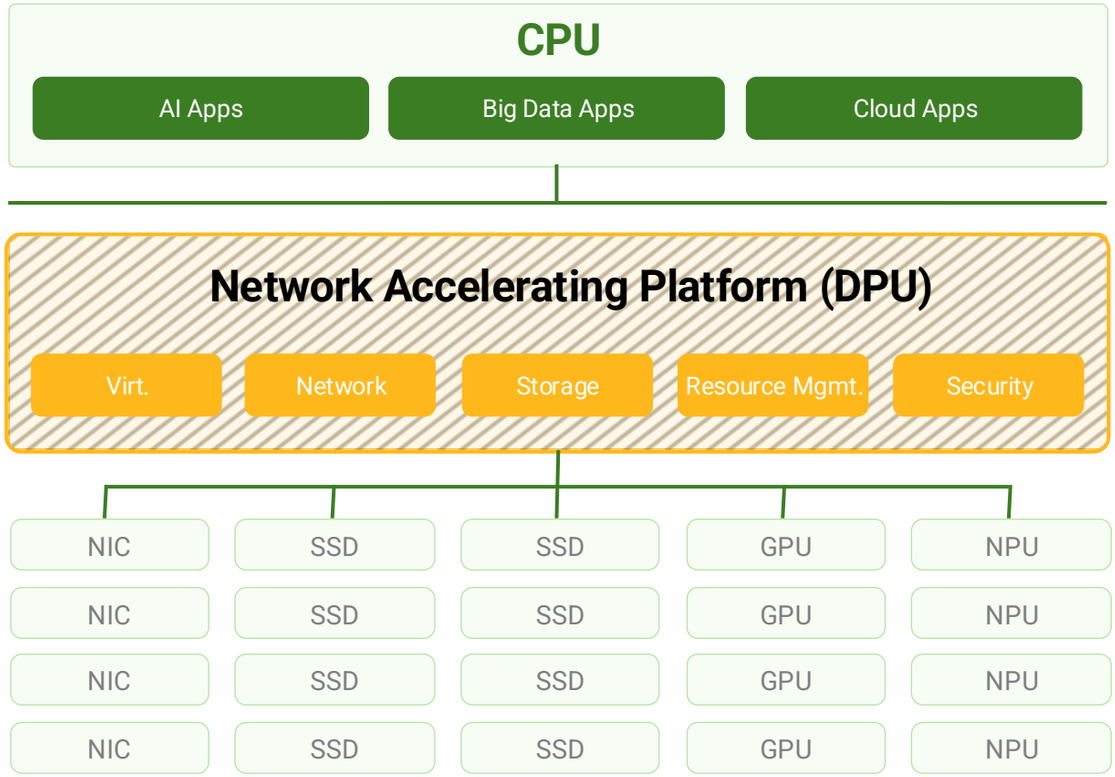
MANGOBOOST

Traditional Data Center:
CPU runs heavy data center I/O tasks



CPUs are located **too far from** data devices
CPUs are **fully overloaded** with data features
CPUs **cannot accelerate** data features

DPU-Centric Data Center:
DPU offloads & accelerates data center I/O tasks

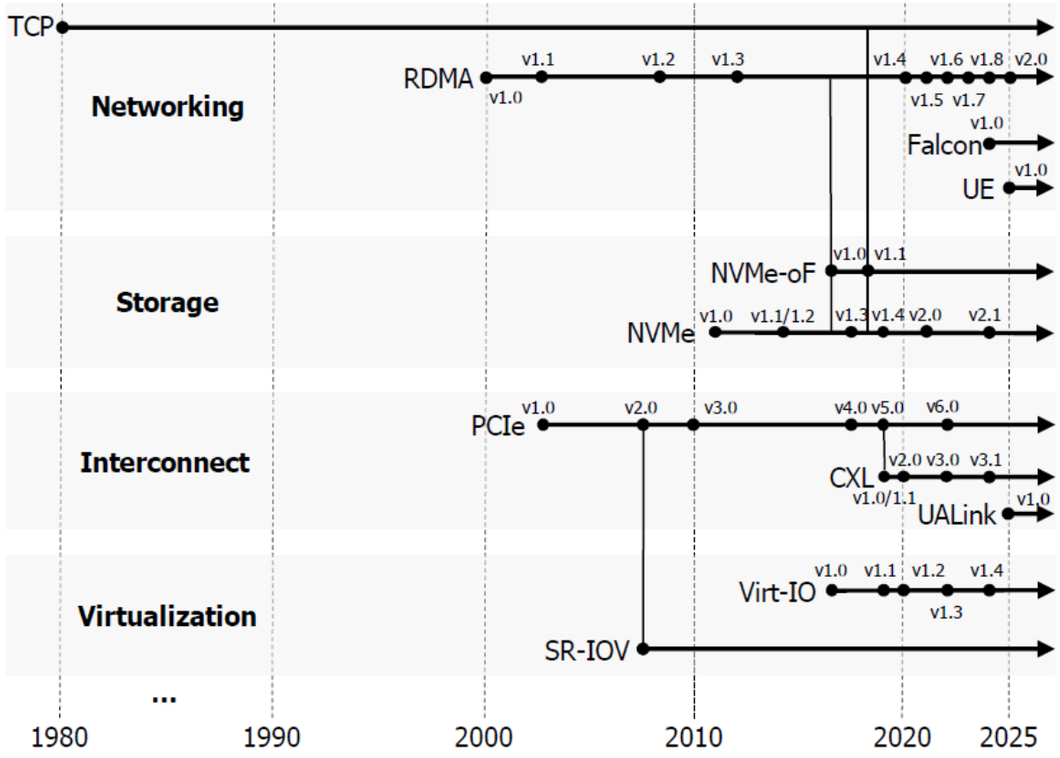


DPUs **boost data center performance and scalability**

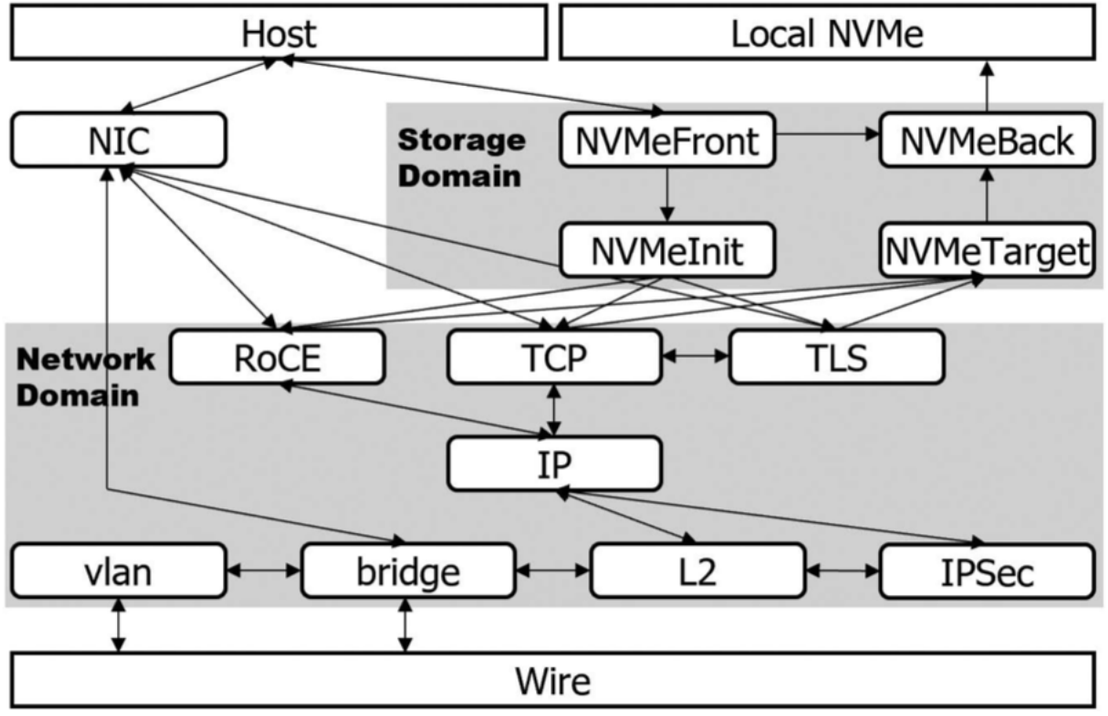
AI Infra Needs: Flexible Data Processing @ Line Rate

MANGOBOOST

Infra Protocols & Standards Evolve



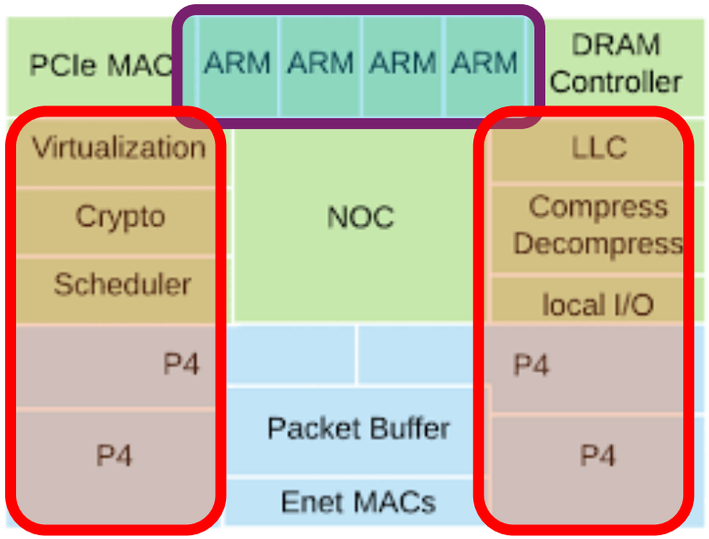
Infra use variety compositions of functions



Related Work: Existing DPUs

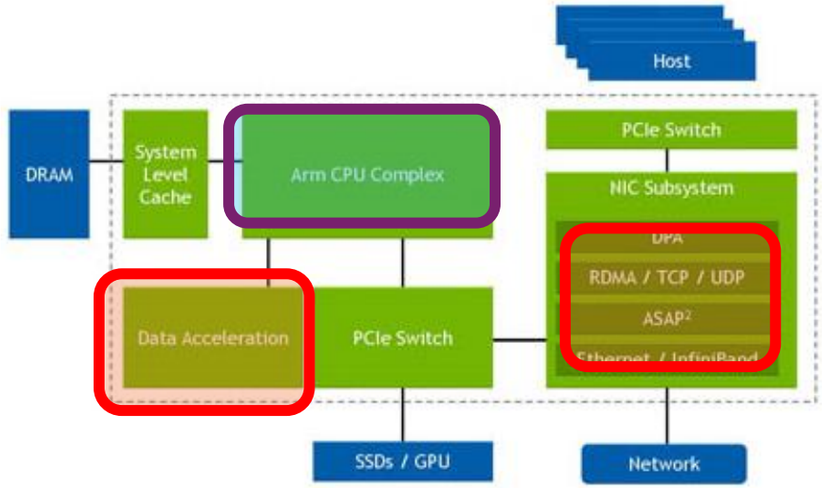
MANGOBOOST

AMD Pensando [HC32]



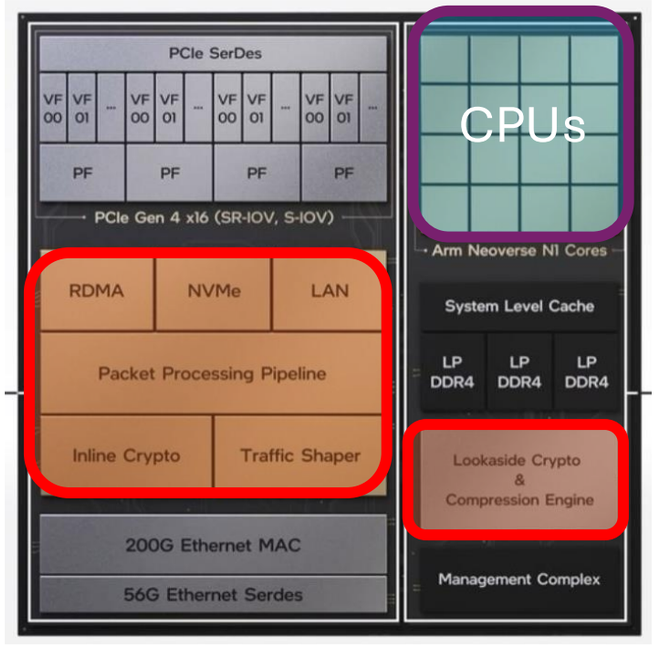
https://hc32.hotchips.org/assets/program/conference/day2/HotChips2020_Networking_Pensando_v3.pdf

Nvidia Bluefield-3 [HC33]



https://www.servethehome.com/wp-content/uploads/2021/08/HC33-NVIDIA-BlueField-3-DPU-System-Architecture.jpg

Intel Mt. Evans [HC33]



https://www.servethehome.com/wp-content/uploads/2021/08/Intel-Architecture-Day-2021-IPU-Mount-Evans-ASIC-Block-Diagram.jpg

In-DPU CPUs

Hardware Accelerators

At high-level, these DPUs contains CPUs + HW Accelerators

Existing DPUs: Challenges

MANGOBOOST

In-DPU CPUs

Challenge 1: In-DPU CPU cores are flexible & general, BUT slow! (vs. host high-end server CPUs)

Hardware Accelerators

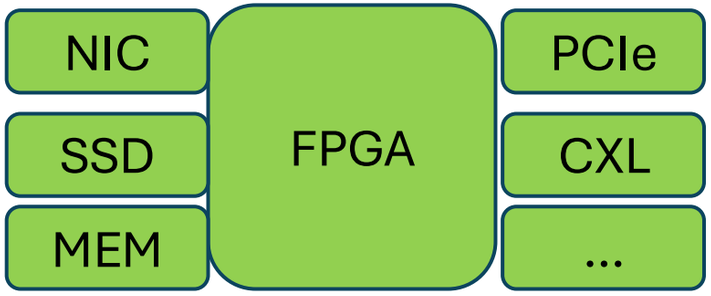
Challenge 2: HW Accelerators can be efficient, but difficult to cover myriad of evolving AI infra functions

Can we do better?

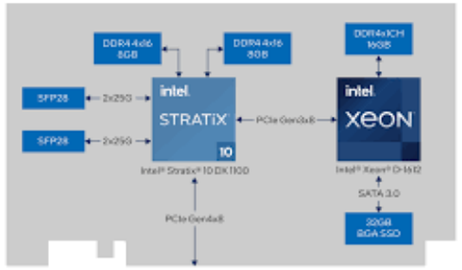
How About FPGA-Based DPUs?

MANGOBOOST

FPGA has **rich I/Os**, to **move data** from many parts of system



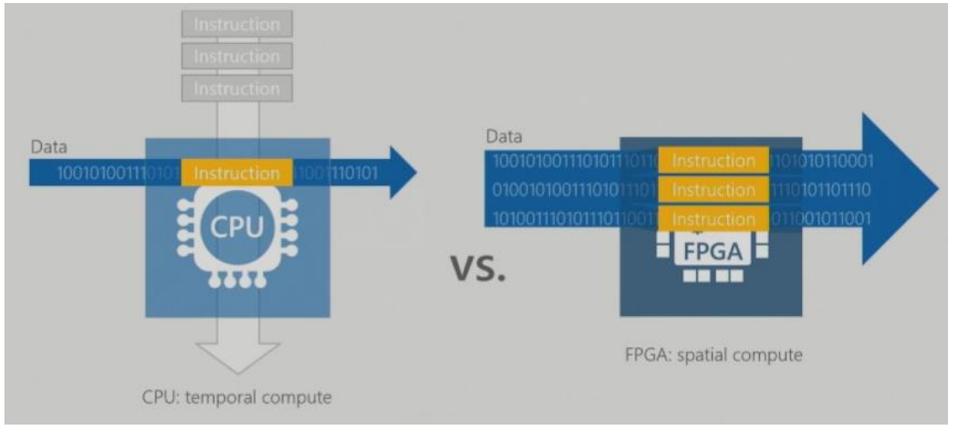
FPGA has **integrated CPUs** to allow full SW stack to run



E.g., Intel IPU
(FPGA + Xeon)

https://www.intel.com/content/www/us/en/products/details/network-io/ipu/c5000x-pl-platform.html

FPGA spatial fabric is great for **processing data "on the move"**

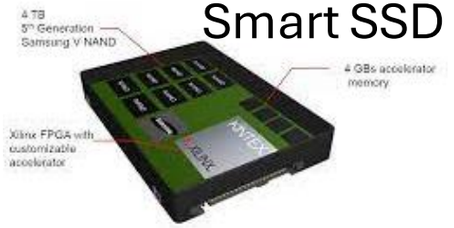


Source: https://www.youtube.com/watch?v=v_4Ap1bjwgs

Many commercial FPGA platforms available, for server systems

Smart NIC

Smart SSD

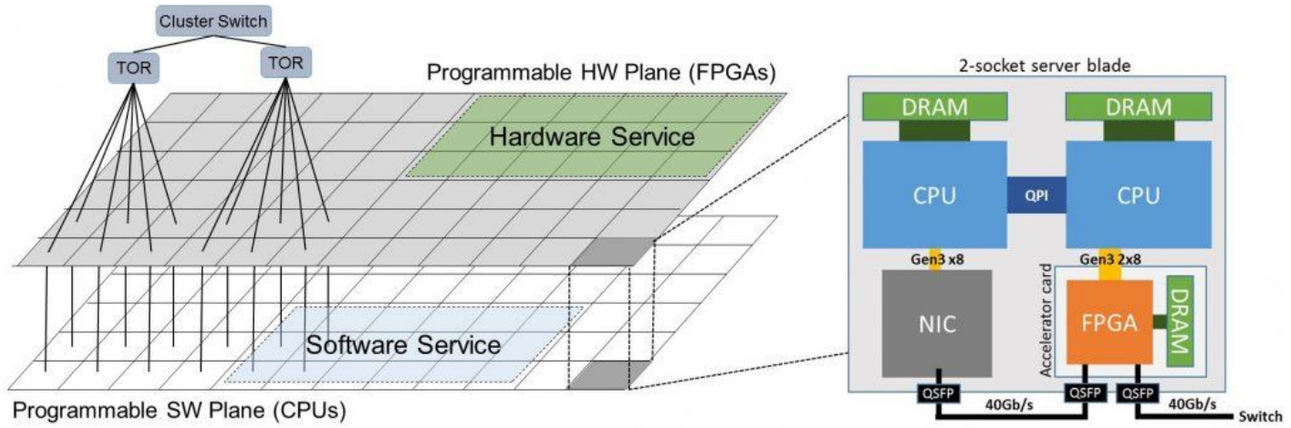


...

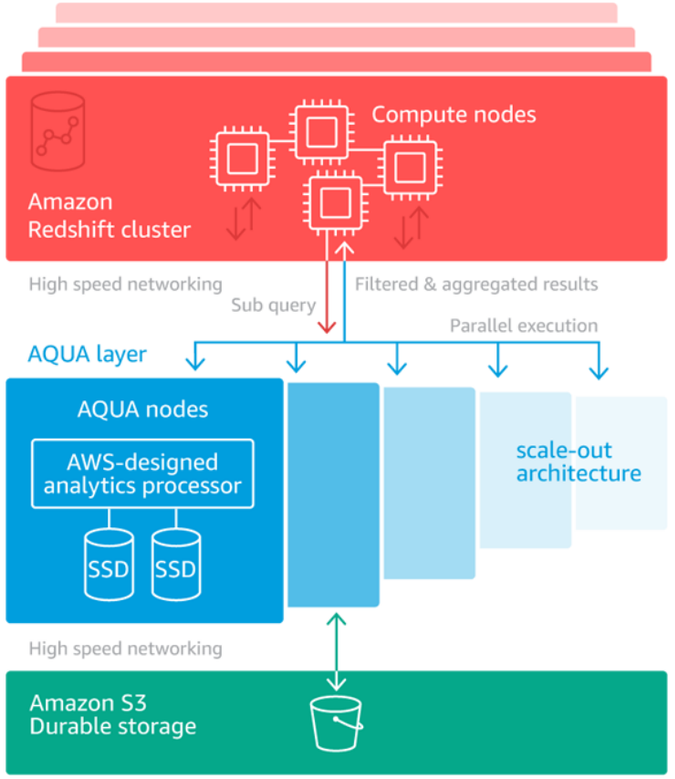
FPGAs are Already Proven @ Cloud Scale

MANGOBOOST

Ex1: As SmartNIC at Microsoft



Ex 2: as storage processor at Amazon



But, these are **hyperscaler specific solution**

And, they require **large development effort**

FPGA-Based DPUs: Challenges

MANGOBOOST

Need FPGA hardware expertise

- Verilog RTL
- Timing closure
- FPGA tool
- …

Need system expertise

- Network protocols
- Hypervisors
- PCIe
- drivers
- Security
- …

Demands interdisciplinary team of experts
(FPGA, systems, architecture, software, etc)

MANGOBOOST

PART III.

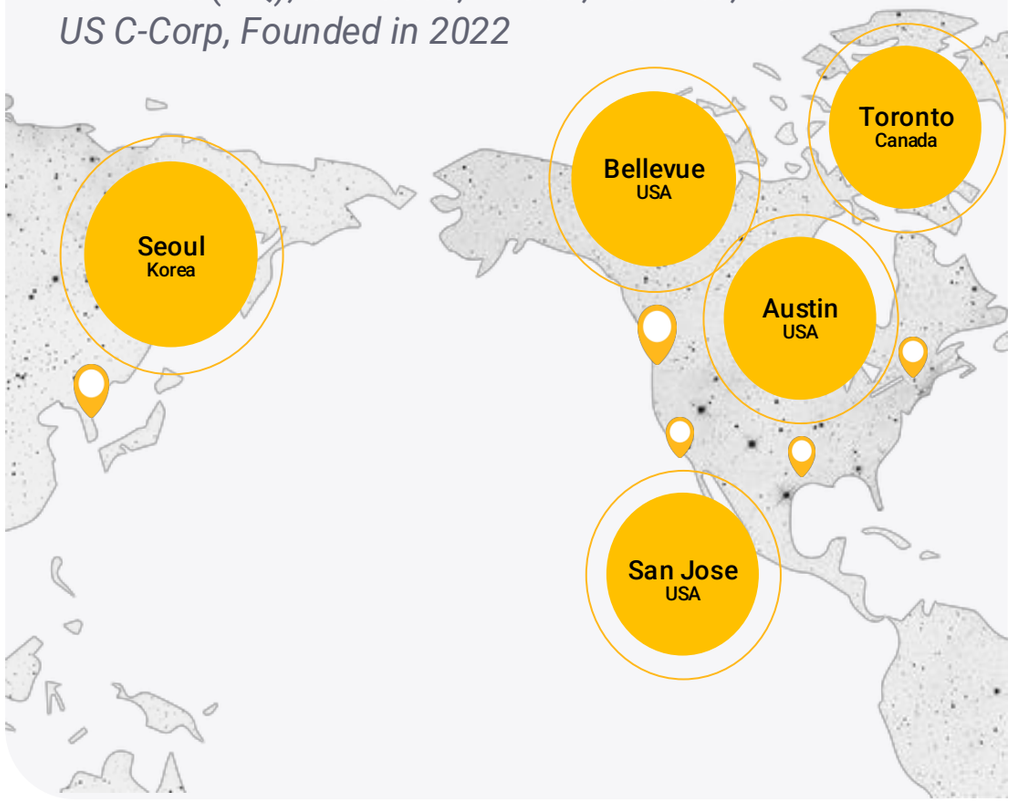
Introducing MangoBoost

Company Profile

MANGOBOOST

Committed to Supporting Customers Worldwide

Bellevue (HQ), San Jose, Austin, Toronto, Seoul
US C-Corp, Founded in 2022



40+ Patents

10+ Years of R&D

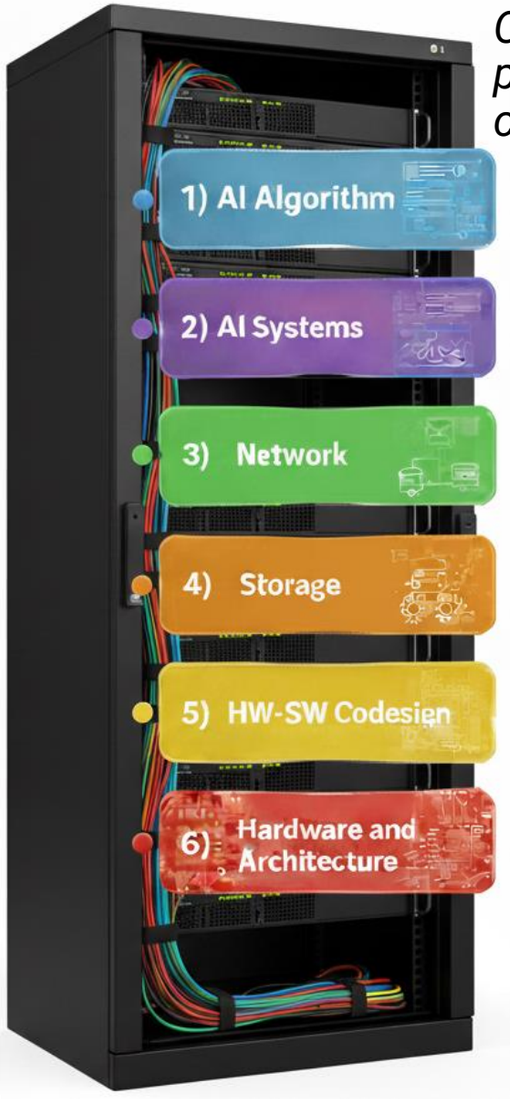
95+ Engineers

64% PhDs

$65M+ Raised

10+ Global Investors

Expertise Across AI Stack



Our engineers published at top conferences

1) AI Algorithm	CVPR, ICLR
2) AI Systems	OSDI, EuroSys, SoCC
3) Network	SIGCOMM, OSDI
4) Storage	ASPLOS, FAST, EuroSys
5) HW-SW Codesign	ASPLOS, OSDI, EuroSys, ATC
6) Hardware and Architecture	ISCA, MICRO, ASPLOS, FPGA, VLSI

Leadership & Investors

MANGOBOOST

CEO

Jangwoo Kim

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

CPO

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

FELLOW

James Hoe

- MIT Ph.D.
- CMU Professor
- IEEE Fellow

IMM Investment Corp.

Carnegie Mellon University Investment Office

서울대학교 SEOUL NATIONAL UNIVERSITY

Shinhan Financial Group

DSC investment

KDB

KB Investment

Premier Partners

STONEBRIDGE

Must Asset Management Make Sure Money

Research to Products: 4-year Company + 10-year Univ Research

2012~2021

Academic

- 10+ years of R&D

Papers

- MICRO 2015 • ISCA 2018 • HPCA 2019 • MICRO 2019 • OSDI 2020 • MICRO 2020 • ATC 2021

Patents:7

2022~2023

MangoBoost

- Founded
- Prototype Sales
- End-to-End MB-DPU Demo

Papers

- ISCA 2023

Patents: +18

2024

MangoBoost

- Official Product Release
  - ✓ NVMe/TCP Initiator and Target
  - ✓ RDMA
  - ✓ TCP Offload Engine
- PoC with Major Customers

Patents: +11

2025~2026

MangoBoost

- LLMBoost Release
- Custom Card
  - ✓ 400G Low Profile Card
  - ✓ 400G Standard Profile Card
- DPU Server Solutions (GPU, Storage)

Patents: +3 (with more in pipeline)

PART IV.

MangoBoost Approach

MangoBoost offers FPGA-based Customizable DPUs

MANGOBOOST

MANGOBOOST PATENTED TECHNOLOGIES, PROTECTED BY 40+ PATENTS

Extensive Full-HW Accelerated DPU IP Portfolio

AI
- Large-scale DNN Training
- Pre/Post-processing Accel.
- MPI Acceleration

Security
- Root of Trust
- Crypto Acceleration
- Network Security

Network
- P4 Support
- TCP Full Acceleration
- GPU Over RDMA (RoCEv2)

Disaggregation
- NVMe over Fabric
- NVMe/TCP
- GPU over Fabric

Virtualization
- SR-IOV, ATS/ATC
- VirtIO Acceleration
- vDPA Support

Storage
- Data Deduplication
- In-line Compression
- LVM/RAID

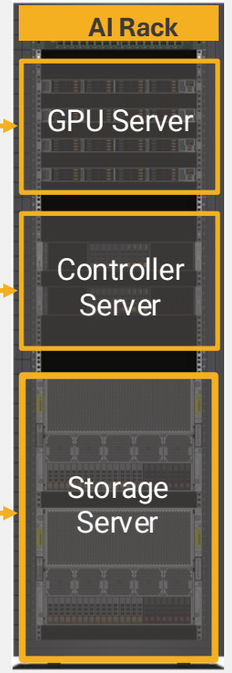
DPU Production Agile Dev Flow

GPUBoost DPU

StorageBoost DPU (client)

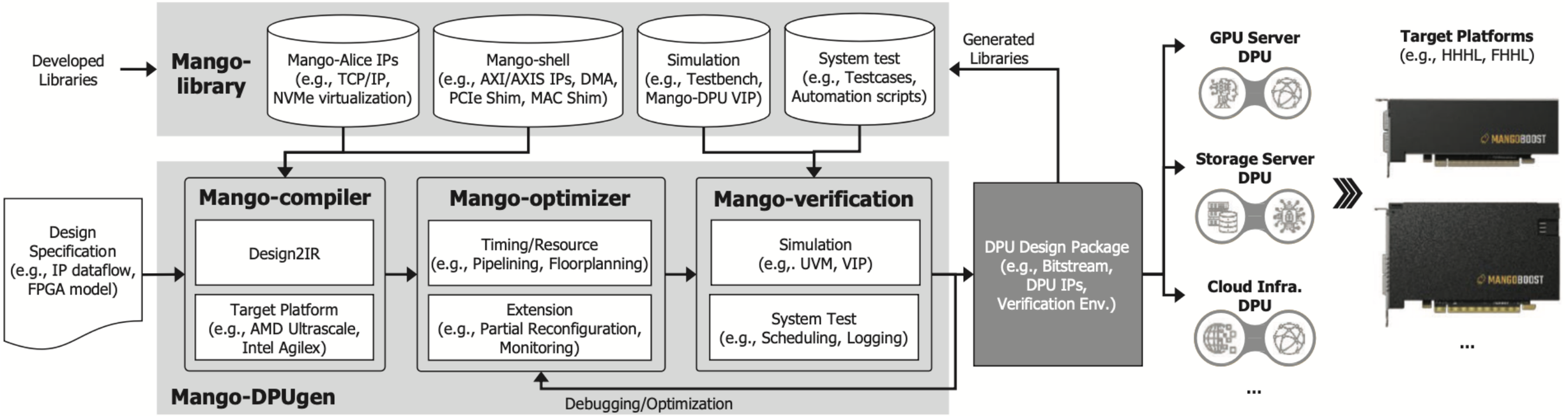
StorageBoost DPU (target)

LLMBoost SW



OPTIMAL AI INFRASTRUCTURE

MangoBoost DPU Production Flow Detailed @ MICRO'25 Paper

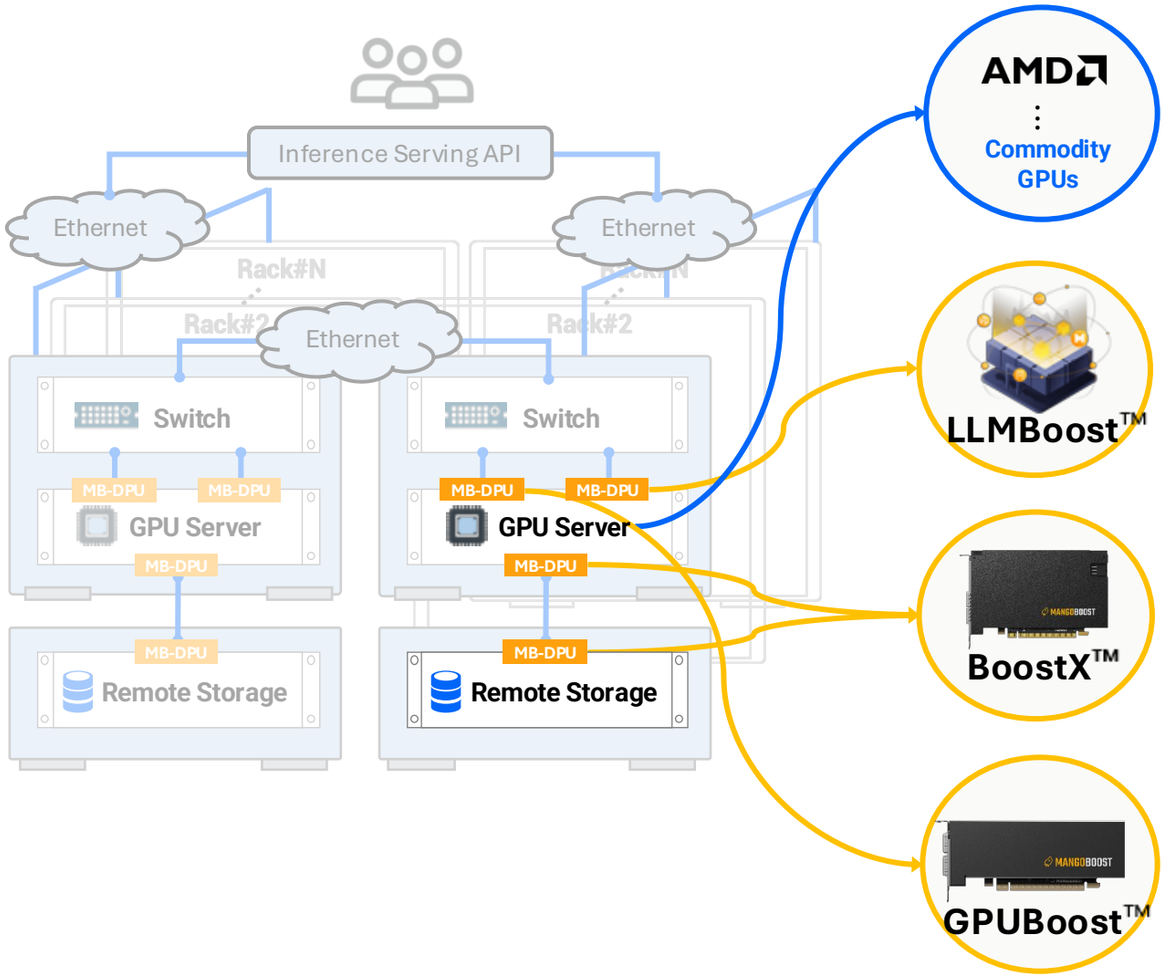


More detail available in the paper: https://ieeexplore.ieee.org/document/11236493

End-to-end development framework to productize DPU solutions on FPGA

Beyond DPUs: MangoBoost Offers System-Level Solutions

MangoBoost Powered AI Infrastructure



Benefits

Easy-to-Deploy Standard-Based Solutions @ Excellent Performance & Scalability

Software

- Containerized & ready to deploy
- Supports many GPUs & open models
- Scales easily across nodes

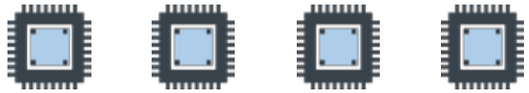
Connectivity on Standard Ethernet

GPUBoost™

BoostXt™

- Software-defined firmware
- High-efficient connectivity
- Based on standard protocols

GPU Hardware



Diverse GPU Options from major vendors (AMD, Nvidia)

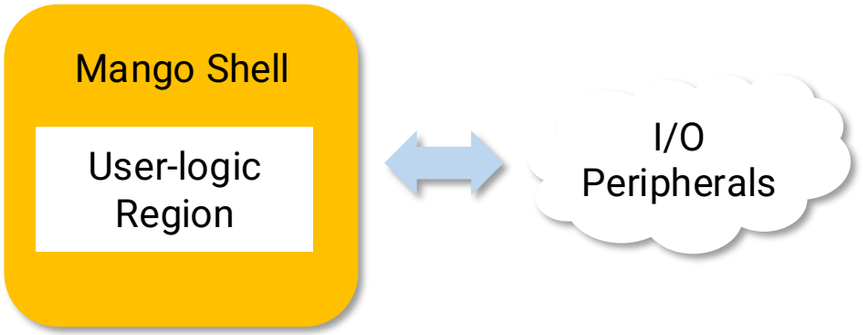
Significantly Lower TCO

Product 1: Mango BoostX SmartNIC

Mango BoostX FPGA Card

- FPGA Chip: AMD Versal (VP1202)
  - 2M System Logic Cells
- Embedded ARM CPU: 2x ARM Cortex-A72
- Power: <75W
- Form factor: Half-height Half-length
- Network Port: 1x QSFP-DD
- PCIe: Gen5 x8x8 or Gen4 x16

Mango Shell

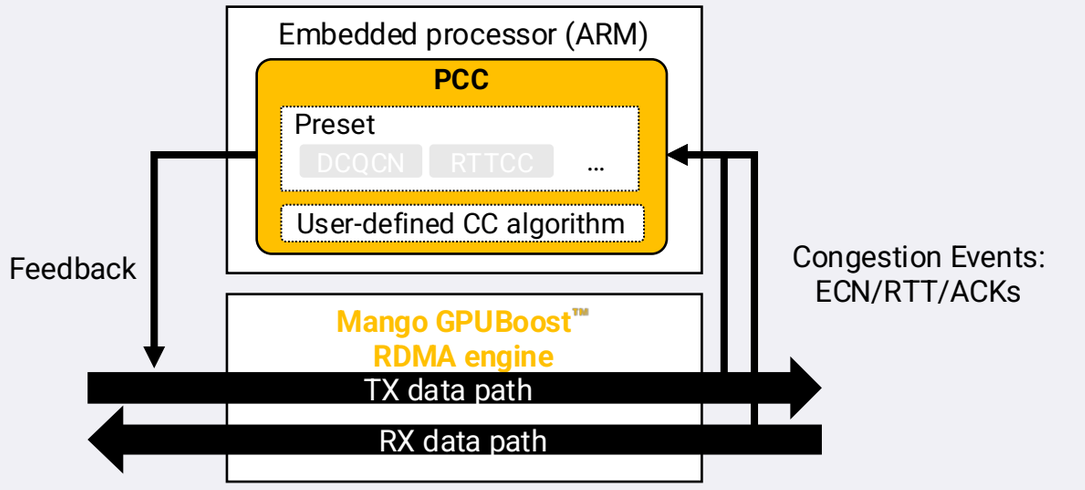


- 2x200Gbps BaseNIC with NIC offloading features
- Provide an abstraction layer for I/O peripherals
- User-logic region intercepts data traffic between PCIe , ARM cores, and Ethernet
  - **70-80% available resource** for user-logic
- Validated with Ubuntu 24.04/22.04

Product 2: Mango GPUBoost™ - For AI Networking Across GPUs

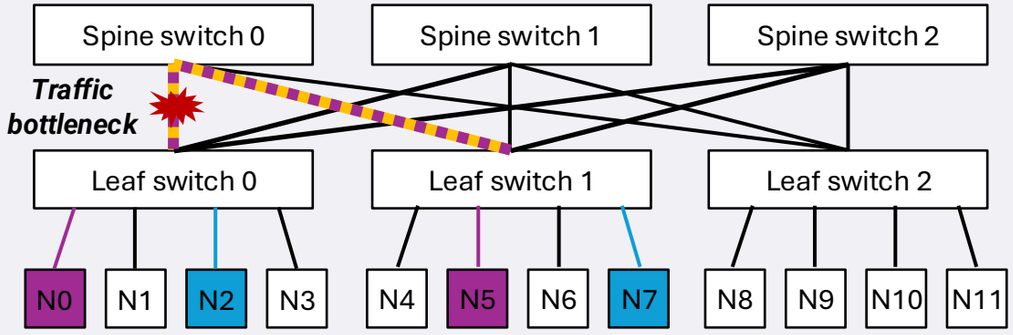
- **Mango GPUBoost™ RDMA** supports **RoCEv2** with **Ultra-Ethernet ready features** tailored to AI
  - Programmable Congestion Control (PCC) / Packet Spraying for load balancing
- **Mango GPUBoost™ RDMA** works with **no dependencies on specific network switches**
  - Support UEC-ready features at the endpoint

Programmable Congestion Control



Mango GPUBoost™ RDMA provides Programmable Congestion Control (PCC), enabling flexible and customizable CC tailored to diverse deployment environments.

Packet Spraying

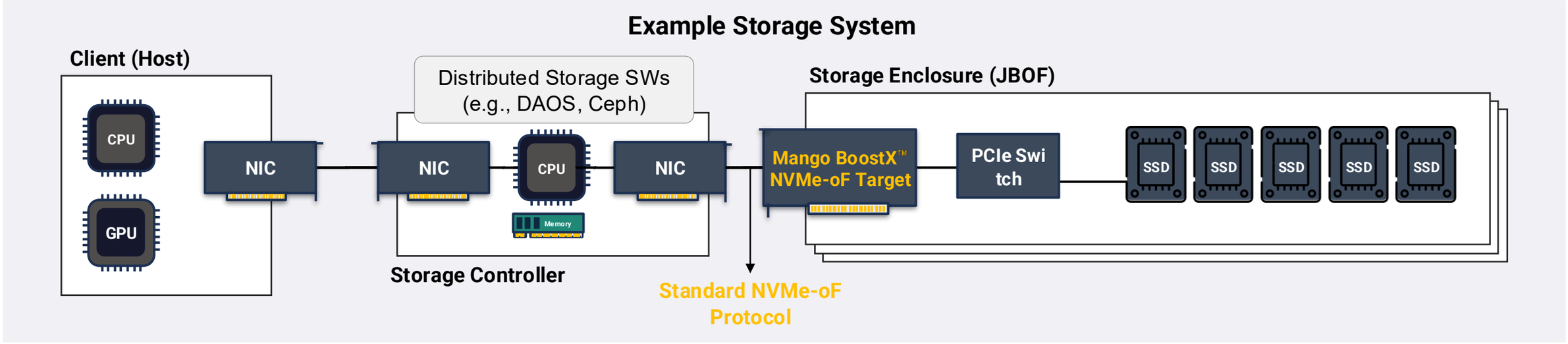


RoCEv2 typically maps a single RDMA flow to a single path, which can lead to "hot spots" in the network fabric.
Mango GPUBoost™ RDMA implements Packet Spraying, which distributes packets across multiple paths by dynamically varying the UDP source port.

Provide Scalable GPU connectivity, compatible with Various GPU and Network Switch Vendors

Product 3: Mango BoostX™ - NVMe-oF Target – For Network Storage

- **Mango BoostX™** enables **cost-effective headless JBOF (use card CPU, no host CPU)** storage enclosures
- **Mango BoostX™** provides **200Gbps performance in a low-power (<75W), and small form-factor (HHHL)**
  - Fully accelerate NVMe-oF Datapath in the specialized HW DPU
- **Mango BoostX™** is compatible **with the standard protocols (i.e., RoCEv2, NVMe-OF)**



Provide high-performance and cost-effective Headless NVMe-oF Target solution

Product 4: Mango LLMBoost™ AI Enterprise SW

AI Software Stack

Application APIs
(Python, REST, OpenAI, WebSocket, etc.)

AI Microservices

Deployment
(Inference Serving)

Data Analytics
(RAG, etc.)

Fine-tuning
(LoRA, etc.)

Training

Infrastructure Management
(Kubernetes, admin UI, monitoring, etc.)

Multi-Node Scaling & Optimized Kernels
(collective communication, auto-scaling, etc.)

Auto Tuning

Automatically optimizes configuration, scaling, scheduling, and parallelization

Easy-to-use, performant, scalable AI SW stack for GPU systems
→ Enhanced with DPU co-designed features (telemetry, custom congestion ctrl, etc)

MANGOBOOST

PART V.

MangoBoost Journey to Bring FPGAs into AI Infra

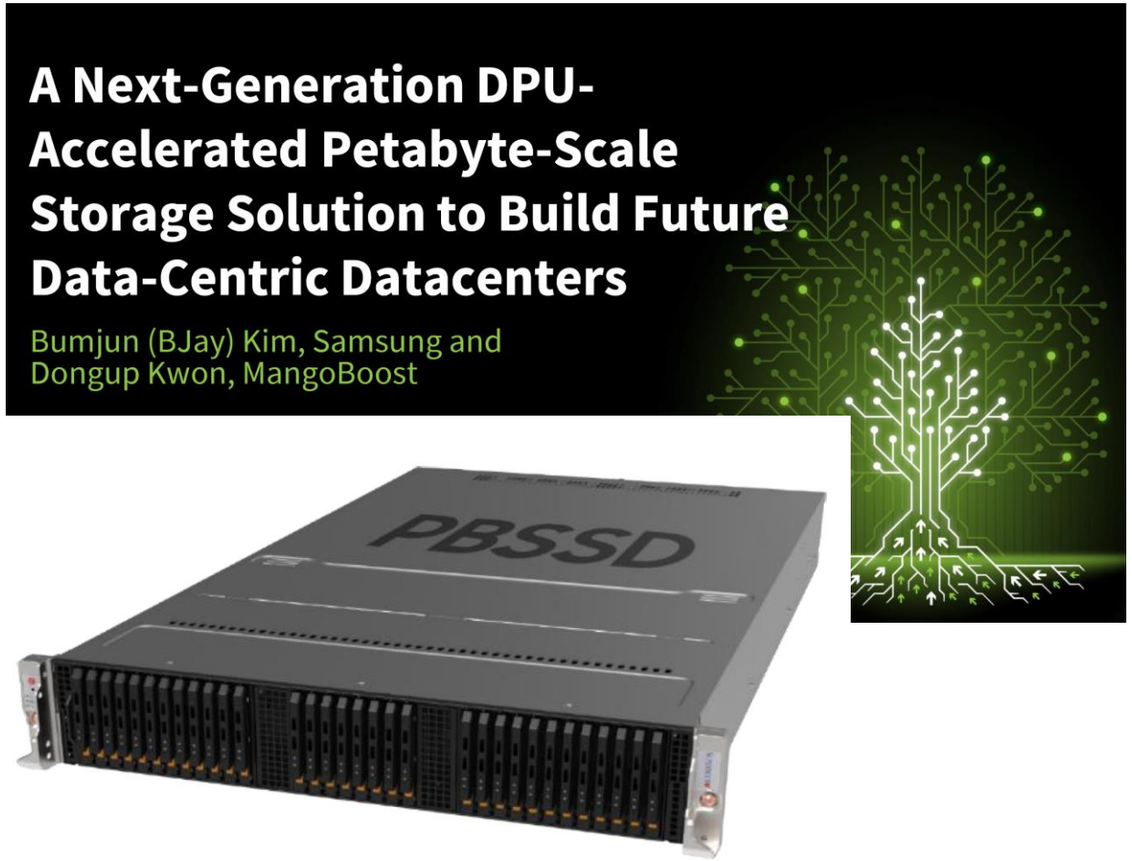
Various FPGA Card Options Available for MangoBoost DPUs

AMD

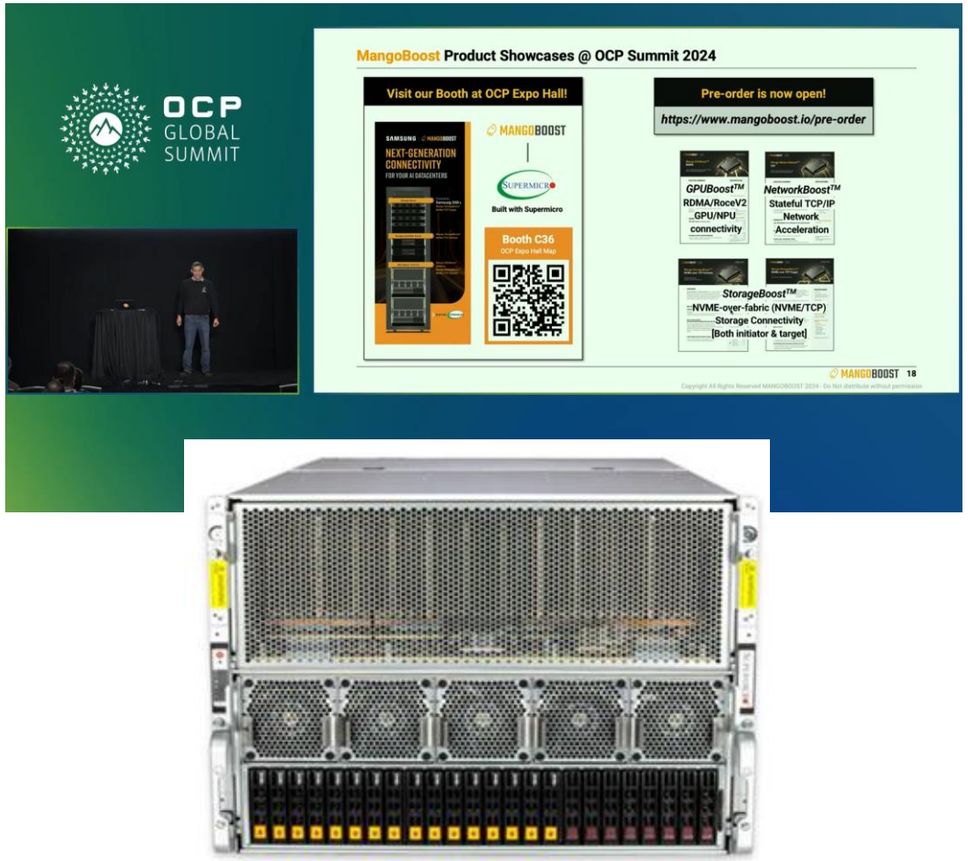
altera

MANGOBOOST

MangoBoost FPGA DPUs Showcased with Partners



Samsung PBSSD for Storage Connectivity



SuperMicro GPU Server for GPU Connectivity

MangoBoost DPU-Accelerated Server Solutions

GPU Servers

High-end Server
- CPU: AMD EPYC 9004/9005
- GPU: 8x AMD MI355X GPUs
- DPU/SmartNIC: upto 8x BoostX
- Form factor: Rack-mounted, 8U

Mid-range Server
- CPU: AMD EPYC 9004/9005
- GPU: upto 8x AMD AI Pro R9700S, 4x Nvidia RTX Pro 6000
- DPU/SmartNIC: upto 2x BoostX
- Form factor: Rack-mounted, 2U

Mid-range Workstation
- CPU: AMD Ryzen Threadripper
- GPU: upto 4x AMD AI Pro R9700, 2x Nvidia RTX 5090
- DPU/SmartNIC: upto 2x BoostX
- Size: Workstation

Storage Server

Controller Node
- CPU: AMD EPYC™ 9000 series processor
- SSD: upto 26x PCIe Gen5 NVMe (U.2)
- DPU/SmartNIC: upto 3x BoostX
- Form factor: Rack-mounted, 2U

Enclosure Node (JBOF)
- CPU: None
- SSD: upto 26x PCIe Gen5 NVMe (U.2) DPU/SmartNIC: upto 2x BoostX
- Form factor: Rack-mounted, 2U

MangoBoost GPU/Storage Servers will be available soon !

MANGOBOOST

PART VI.

Results on Industry-Standard AI Benchmarks: MLPerf

Inference @ MLPerf Inference v5.1

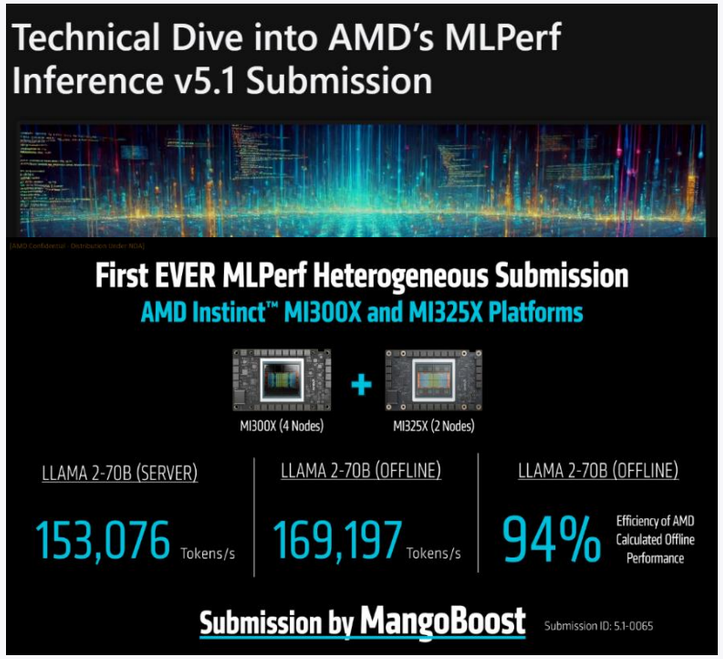
First Heterogeneous Multi-Node Results
[6 nodes, MI300X + MI325X]

- 35% performance vs. next-best result
- 3.9X cost-efficiency vs. vLLM
- 94% Multi-node Performance Scaling

First Multi-Node Results with Latest AMD GPU
[8 nodes MI355X]

- 1st third-party submission on AMD MI355X
- Highest performance in the open division

AMD's Official Tech Blog Post



Meena Arunachalam

Fellow, AI Performance Design Engineering, AMD

"We are thrilled with our MLPerf Inference 5.1 co-submission with MangoBoost. **Together, we set a new performance record for Llama2-70B using AMD Instinct™ MI355X GPUs**, scaling to 4-node and 8-node configurations. MangoBoost also **delivered the first heterogeneous multi-node submission showcasing the combined power of AMD Instinct MI300X and MI325X GPUs**, …. With nine high-performance MLPerf inference submissions, **MangoBoost's LLMBoost GenAI, powered by AMD ROCm™, provides seamless scalability and easy deployment for enterprise AI workloads**."

Training @ MLPerf Training v5.0 on Multi-Node AMD MI300X

MANGOBOOST

Highlights

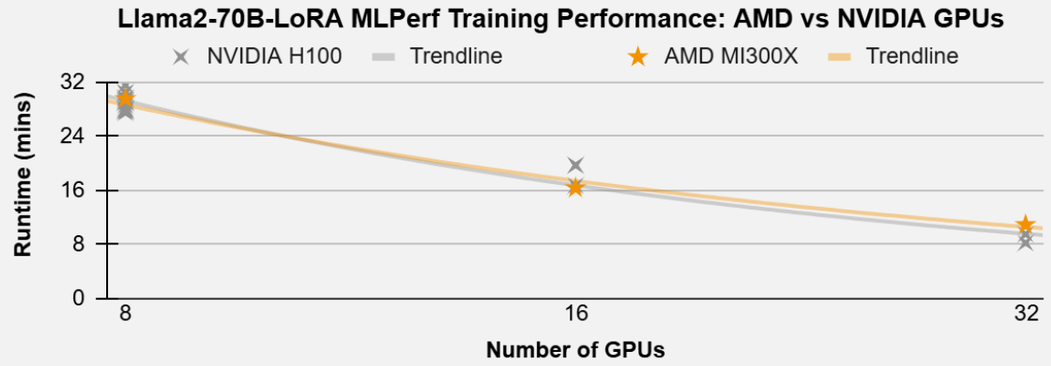
- First-ever MLPerf multi-node training on AMD Instinct MI300X GPUs
- Fastest fine-tuning of LLaMa2-70B-LoRA on AMD hardware
- Seamless orchestration across 32 GPUs

David Kanter
Founder, Head of MLPerf, MLCommons

"I'm excited to see MangoBoost's first MLPerf Training results … **Their results underscore that a well-optimized software stack is critical to fully harness the capabilities of modern AI accelerators.**"

Meena Arunachalam
Fellow, AI Performance Design Engineering, AMD

"We congratulate MangoBoost on their MLPerf 5.0 training results on AMD GPUs and are **excited to continue our collaboration with them to unleash the full power of AMD GPUs.**"



LLMBoost™ with MI300X delivers competitive multi-node performance, powered by our co-designed software and hardware stack.

Related Articles

Nvidia's Blackwell Conquers Largest LLM Training Benchmark…

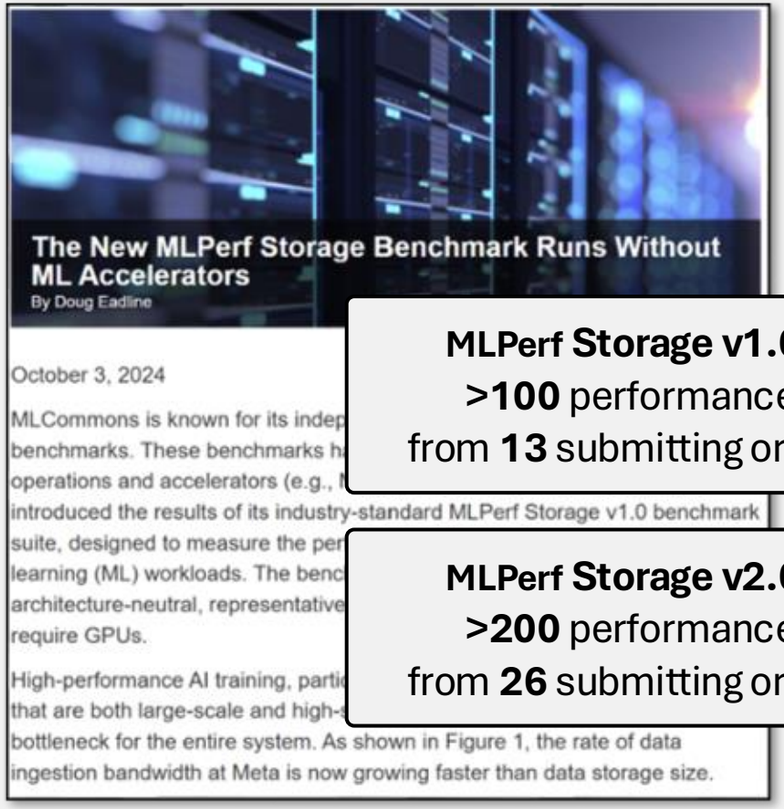
Making AI work is increasingly a matter of the network, latest …

MLPerf Shows AMD Catching Up with Nvidia's Older H200 GPU…

Storage @ MLPerf Storage (v1.0 & v2.0)

MANGOBOOST

MLPerf Storage:
Industry Standard Storage Benchmark for AI


The New MLPerf Storage Benchmark Runs Without ML Accelerators

By Doug Eadline

October 3, 2024

MLCommons is known for its indep
benchmarks. These benchmarks h
operations and accelerators (e.g., t
introduced the results of its industry-standard MLPerf Storage v1.0 benchmark
suite, designed to measure the per
learning (ML) workloads. The benc
architecture-neutral, representative
require GPUs.

High-performance AI training, parti
that are both large-scale and high-
bottleneck for the entire system. As shown in Figure 1, the rate of data
ingestion bandwidth at Meta is now growing faster than data storage size.

MLPerf Storage v1.0 (2024):
>100 performance results
from **13** submitting organizations

MLPerf Storage v2.0 (2025):
>200 performance results
from **26** submitting organizations

Key Highlights

World 1st DPU-accelerated system Results for MLPerf Storage

Best-performing Ethernet-based systems in MLPerf Storage

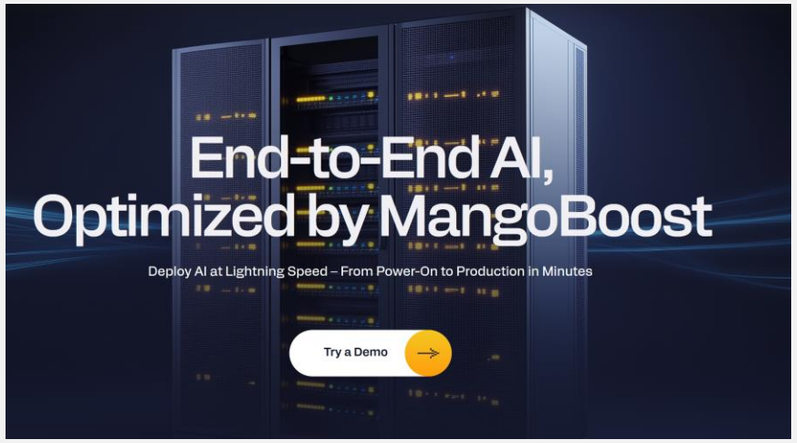
- *[v1.0]* ***1.2x-4.7x Speedup*** *vs ALL other single-host submitter*
- *[v1.0]* ***1.4x-7.8x Speedup*** *vs ALL other multi-host submitter*
- *[v2.0]* ***1.5x Speedup*** *over MangoBoost v1.0 results*

Demo for MangoBoost AI Solutions

MANGOBOOST

Experience MangoBoost AI

www.mangoboost.io/ai-virtual-demo



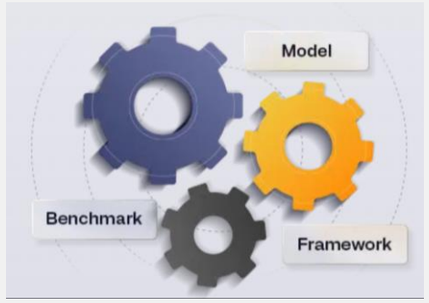

```
$ llmboost serve --model_path google/gemma-2-2b-it

[INFO] Initializing LLMBoost server...
[INFO] Loading model from: google/gemma-2-2b-it
[INFO] Model loaded successfully in 3.42s
[INFO] Starting inference server...
[INFO] LLMBoost server is running.
[INFO] Model path: google/gemma-2-2b-it
```

Competitor Comparison Metrics

Start your trial request instantly

Capex & Opex Saving Breakdown

Our engineer will configure your workload environment



Optimal Deployment Sizing

24-hour access to your AI server testbed

MANGOBOOST

PART VII.

Closing Remarks

Lessons from Commercializing FPGA Research into Products

MANGOBOOST

FPGA's flexibility indeed enable fast time-to-solution (vs. ASIC)

- multiple FPGA-based solutions (network, storage) ready within first year
- on multiple FPGA cards, speeds, form-factors quickly enabled via partnerships

For ease-of-use, may need to "hide" the FPGA from the end users

- comply with standards (e.g., NVMEoF, RoCEv2) with known SW already
- general end users do not need to be exposed to FPGAs

From research prototype to products need a lot of effort

- verification, compliance, maintenance, support,...

It’s exciting time for FPGA research @ AI Infra!

MANGOBOOST

Many research topics, some examples below

FPGA role for emerging new AI system approaches

- e.g., multi-node disaggregated P/D, KV-cache storage systems, etc

How to leverage FPGA as DPU in end-to-end AI systems

- e.g., SW/HW co-design with DPUs, DPU-enhanced servers

DPU extensions / enhancements

- e.g., new functions, protocols, near/in-network compute, etc

MANGOBOOST

End of Document

Website www.mangoboost.io
Contact contact@mangoboost.io