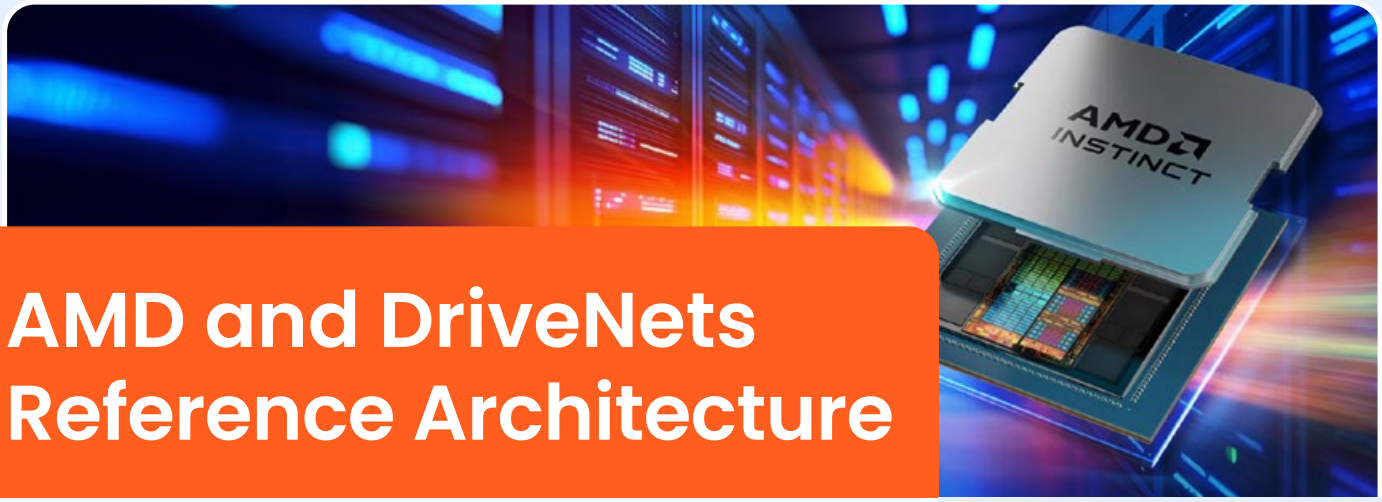
DRIVENETS



AMD and DriveNets Reference Architecture

System-Level Optimization Is the New AI Scaling Standard

Building high-performance AI clusters is no longer a matter of adding GPUs. Large-scale clusters require end-to-end system-level performance optimizations, that look into how well compute, NICs, networking, software, and workload optimizations work together.

The Reference Architecture published by **AMD** and **DriveNets** maximizes the performance and efficiency of AI clusters by looking at them end-to-end. It provides a validated full-stack blueprint for AI clusters that are based on **AMD Instinct MI355X GPUs**, **AMD Pensando Pollara 400 AI NIC**s, and **DriveNets AI Fabric.** It brings together AMD compute, DriveNets UEC-ready scale-out networking, including frontend and storage connectivity, ROCm and RCCL tuning, automation, benchmarking, and deployment services into one integrated architecture. This combined solution ensures rapid deployment and efficient end-to-end scaling.

Download the full reference architecture here.

AMD & DriveNets Joint Solution Architecture

Full-stack networking for AMD based clusters

DRIVENETS

ACO
AI Cluster Orchestrator

DNOS
Network Operating System

Switch
2500S

AMD

NICs
Pollara

Compute
MI355x

Reference architecture main elements:

AMD Instinct™ MI355X compute

High-density AI compute based on the AMD Instinct MI350 Series, designed for large-scale training, inference, and HPC workloads.

AMD Pensando Pollara 400 AI NIC

UEC-ready, programmable 400G AI NIC supporting intelligent packet spray, in-order delivery, selective retransmission, and path-aware congestion control.

DriveNets AI fabric

Extends lossless AI fabric connectivity across geographically separated data centers up to 100 km.

Scalable full-stack optimization:

Unifies GPU back-end, storage, and front-end traffic on a single Ethernet infrastructure. This eliminates separate physical networks.

Joint Architecture Key Benefits

Maximize GPU utilization

Optimized scale-out networking, RCCL tuning, and end-to-end fabric validation help reduce communication bottlenecks and keep GPUs productive.

Reduce operational risk

The architecture validates the compute, NIC, network, software, and workload layers together, helping teams avoid component-level tuning cycles.

Accelerate time to production

Validated architecture, deployment guidance, automated provisioning, benchmarking, and services to simplify cluster bring-up.

Build on Open Ethernet

Built on Ethernet with UEC-ready components, the architecture helps avoid proprietary lock-in while supporting large-scale AI infrastructure growth.

Highest performance AMD clusters

As part of the integrated reference architecture's build, planning, and delivery process, we conducted thorough testing and benchmarking to tune the end-to-end cluster configuration and evaluate performance across RDMA bisection, network resiliency, RCCL tests, LLM inference, and LLM training workloads

TRAINING

average of **2,019.89** Tokens Per GPU

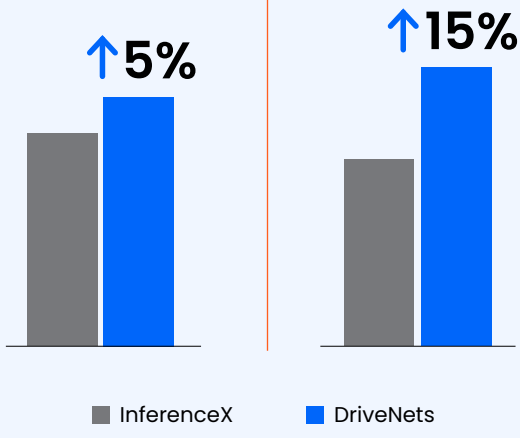
average of **796.97** TFLOP/s/GPU

Tested Grok-2 on a 64-GPU AMD MI355X cluster

INFERENCE

Output token throughput GPU ↑5%

Faster TTFT ↑15%



Tested DeepSeekR1 and Kimi 2.5 on a single node AMD MI355X cluster

DRIVENETS