AMD
together we advance_

5 REASONS WHY YOU SHOULD CHOOSE THE AMD PENSANDO™ VULCANO 800 AI NIC *TO ACCELERATE AI WORKLOADS*

AI NETWORKING SOLUTIONS

AMD Pensando™ Vulcano 800 AI NIC

AMD
together we advance_

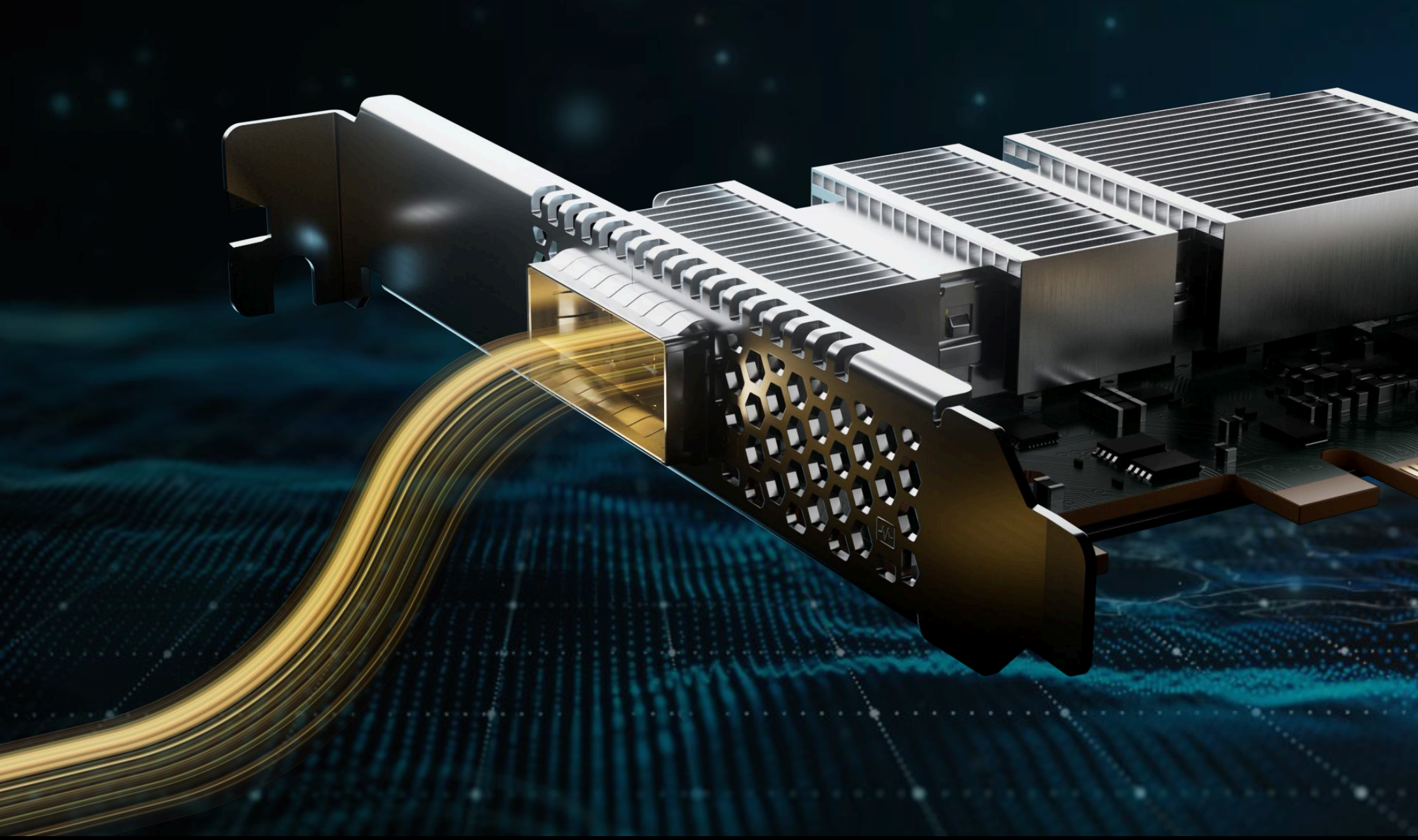
SUMMARY

Challenges with Existing Solutions.

The latest AI workloads have exposed a fundamental gap in how infrastructure is built. As clusters scale to support the increasing demands and complexity of model training, distributed AI, and agentic AI workloads, the deciding factor is no longer raw compute alone, but how the entire system works together across compute, memory, and storage through a connected network. For example, in AI training, infrastructure must keep every node fed, resilient, cost-efficient, operationally consistent, and capable of scaling across often sprawling environments. The AMD Pensando™ Vulcano 800 AI NIC makes this possible.

AMD Pensando™ Vulcano 800 AI NIC: Purpose-Built for Scale-Out AI.

The AMD Pensando™ Vulcano 800 AI NIC helps organizations accelerate and scale the most demanding workloads on open, standards-based Ethernet rather than proprietary interconnects. Built on the proven, fully hardware- and software-programmable P4 architecture, at 800 Gbps speeds, the AI NIC is the only one on the market to offer up to 2.4 Tbps of scale-out bandwidth per GPU, delivering up to 13% faster AI job completion times[1]. The AI NIC gives organizations the performance, availability, cost efficiency, resilience, and transport flexibility needed for next-generation AI infrastructure. Here's how.



AMD
together we advance_

WHY CHOOSE THE AMD PENSANDO™ VULCANO 800 AI NIC

(1)

LEADERSHIP PERFORMANCE WITH UP TO 13% FASTER AI TRAINING TIMES[1]

Keep GPUs Continuously Fed.
As the only NIC to support up to 2.4 Tbps of scale-out bandwidth per GPU, the AMD Pensando™ Vulcano 800 AI NIC is designed to keep GPUs continuously fed, helping reduce network delays that slow AI workloads. Get leadership Ethernet performance for GPU-to-GPU communication at any scale.

(2)

HIGHLY AVAILABLE NETWORKS

Minimize Downtime. Maximize Availability.
The AMD Pensando™ Vulcano 800 AI NIC is engineered for production AI workloads with the high availability that can keep training and agentic tasks operating through failures.

(3)

COST-EFFICIENT AI NETWORKING WITH UP TO 33% LOWER SWITCHING COSTS[2]

Don't Just Scale AI. Control the Cost.
High-throughput connectivity cuts the number of switches, cables, and optics your AI fabric needs. Against competitors, the AMD Pensando™ Vulcano 800 AI NIC can lower fabric capital spending, so you scale on cost-efficient, standards-based Ethernet.

(4)

OPERATIONAL EXCELLENCE AT SCALE

Simplify and Run Resilient AI Networks.
The AMD Pensando™ Vulcano 800 AI NIC helps minimize downtime, validate cluster health, and provide advanced telemetry. Maintain infrastructure production readiness through live monitoring and hitless and graceful upgrades that roll out with virtually no disruption.

(5)

ADAPTABLE NETWORK FABRICS THAT FRONTIER AI DEMANDS

One Programmable NIC. Any Transport AI Needs.
Built on a fully programmable P4 engine, the AMD Pensando™ Vulcano 800 AI NIC supports a wide range of transports, from UEC-ready RDMA and Multipath Reliable Connection (MRC), to custom transports built for specific workloads. Tune and upgrade these transports as standards evolve, so the same NIC can continually be adapted to new AI networking demands as fast as software evolves.

TECHNICAL DEEP DIVE

(1) LEADERSHIP PERFORMANCE

- Deliver up to 2.4 Tbps of scale-out bandwidth per GPU with three NICs, keeping accelerators fed across training and inference.
- Shrink the all-reduce and all-to-all communication that sits exposed on the training critical path, so added bandwidth can turn directly into shorter step times instead of idle accelerators.
- Spread traffic across hundreds of simultaneous paths with adaptive multipath load balancing, designed to keep every link busy and hold throughput at peak.
- Run RCCL collective operations directly on the NIC to help cut all-reduce latency and reduce GPU idle time at scale.

(2) HIGHLY AVAILABLE NETWORKS

- Detect path degradation and redirect traffic to healthy links with hardware-native fast failover and automatic rerouting.
- Recover from packet loss with selective retransmission that resends only the lost packets, unlike the go-back-N behavior of RoCE environments.
- Contain failures within a network plane before any disruption spreads across the cluster, helping keep long-running training jobs on track.
- Reduce in-cast congestion and head-of-line blocking with hardware-based congestion control that needs no complex switch-side configuration.

(3) COST-EFFICIENT AI NETWORKING

- Scale on cost-efficient, standards-based Ethernet instead of proprietary interconnect fabrics, opening the door to competitive multi-vendor switching.
- Lower total fabric capital cost by up to 33%2 at 32,000-GPU scale on standards-based Ethernet, without giving up performance.

(4) OPERATIONAL EXCELLENCE AT SCALE

- Achieve end-to-end visibility into the AI fabric with real-time telemetry across tens of thousands of NIC endpoints at once.
- Capture per-flow statistics and path utilization through P4-powered programmable observability.
- Simplify switch-side configuration by running load balancing and congestion management on the NIC, helping lower the operational burden at scale.
- Bring network and GPU insight together through integration with AMD ROCm™ software for simpler lifecycle management across the cluster.

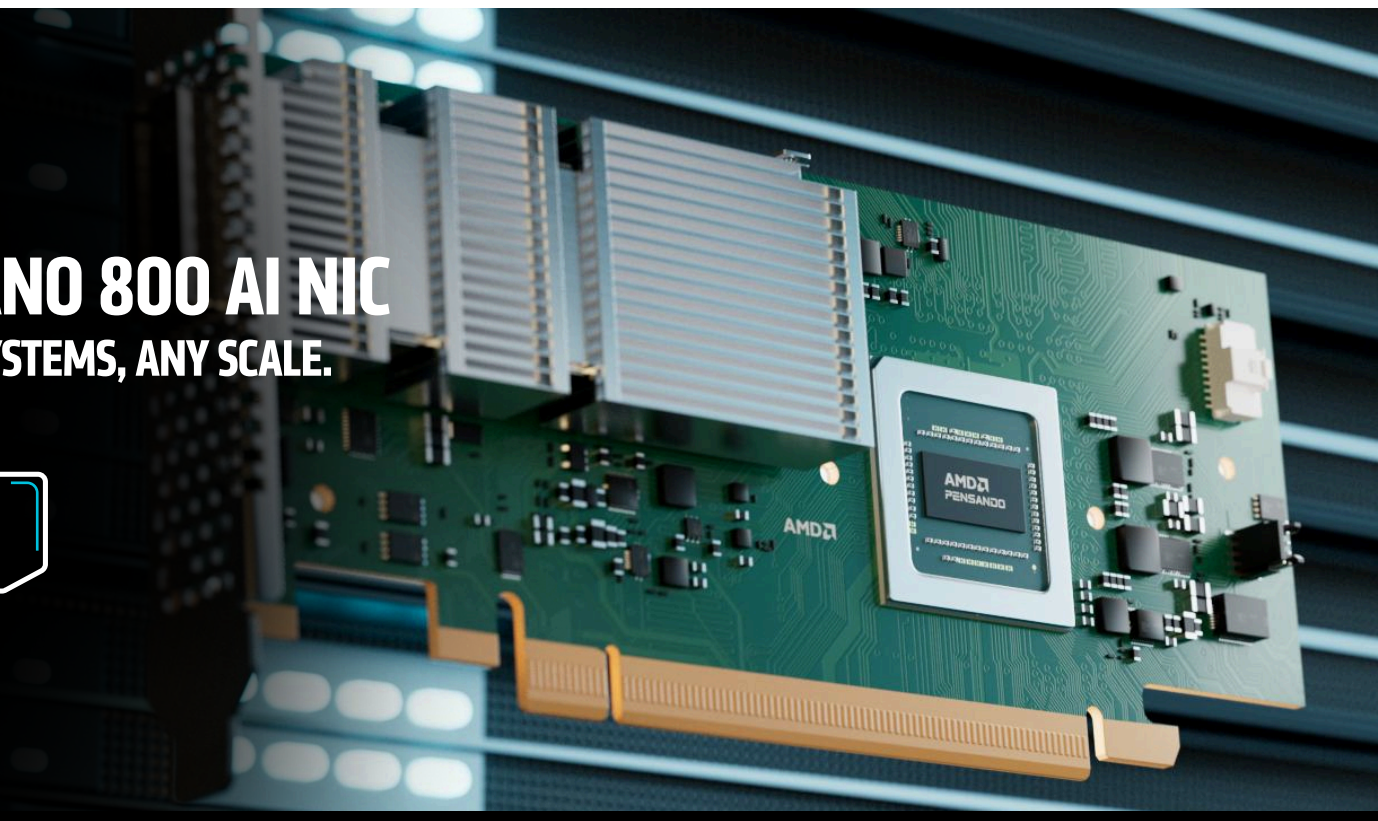
(5) ADAPTABLE NETWORK FABRICS THAT FRONTIER AI DEMANDS

- Deploy across any standards-conformant Ethernet switch with full UEC 1.0 compliance, building multi-plane topologies that minimize proprietary hardware dependencies.
- Choose any transport protocol on the fully programmable P4 engine, from UEC-ready RDMA and MRC (Multipath Reliable Connection) to custom transports built for specific workloads, with MRC spreading traffic across hundreds of paths to help remove fabric hot spots.
- Control congestion with NSCC, the AMD hardware-based algorithm now defined in the UEC 1.0 specification, helping reduce in-cast congestion without proprietary switch dependencies.
- Tune and upgrade transports in software as standards evolve, so one programmable NIC adapts to evolving AI demands faster than the pace at which new hardware is developed.

AMD PENSANDO™ VULCANO 800 AI NIC

ACCELERATE AI NETWORKING. OPEN SYSTEMS, ANY SCALE.

Learn more at AMD.COM/Networking



FOOTNOTES:

1. MI400-018 :Based on AMD Engineering silicon modeling and AMD synthetic benchmark simulation, using the MOE_4p5T hi_sparsity_ea9 benchmark test to project time-to-solution speed up in days using the FP8 training datatype on a simulated LLM system modeled with 8000 AMD Instinct MI455X GPUs and (3) versus (2) AMD Pensando Vulcano NICs. Results reflect analysis of pure training compute (decode layers only) configuration evaluated with a global batch size of 4096 and a sequence length of 4K, using SwiGLU activation. Results exclude evaluation and checkpointing overhead. FlashAttention v3 with matmul–softmax overlap is assumed. AllReduce and All2All communication costs are fully accounted for and not hidden via tiled compute–communication overlap. Gradient synchronization and FSDP weight prefetch are evaluated across varying levels of overlap to assess scale-out sensitivity. Results assume ideal and not fully optimized real-world behavior and may vary when actual product(s) are released in market.
2. PEN-022: AMD comparison and pricing as of May 18, 2026, for network fabric costs to support 32,000 GPUs. Comparison of a Vulcano-based NIC (VULCANO-CUSTOM-2.4T) deployed as part of a Helios rackscale system with a network based on 1.6T Tomahawk 6 switching with 200G SerDes versus using a competitor 800G NIC with 800G Tomahawk 6 switching with 100G SerDes. Both fabrics were fat-tree topologies built on Tomahawk 5 800G switching platforms, with NIC costs considered comparable. The Vulcano-based design is estimated to deliver up to 33% savings in network switching costs by enabling a more cost-effective architecture with fewer switching platforms, more bandwidth per port on the network, and reduced transceiver cables/optics.

   TH6-100G Serdes Fat-Tree (Competition):

   Switching (TH6C BCM78914 - 128x800G):

   - Leaf Units 1,000
   - Spine Units 500
   - Total Switches 1,500
   - TH6 100G Unit Price $79,587
   - Total Switching Cost $60M

   Cables/Optics:

   - NIC Transceivers (800G-DR8) 64,000 @ $500
   - Leaf/Spine Transceivers (800G-DR8) 192,000 @ $500
   - MPO Cables 256,000 @ $89
   - Total Optics/Cables $64M

   Total Fabric Cost TH6-100G (Switches + Cables/Optics): $124M

   TH6-200G Serdes Fat-Tree (Vulcano-Custom2.4T / AMD Solution):

   Switching (TH6P BCM78910 - 64x1.6T):

   - Leaf Units 1,000
   - Spine Units 500
   - Total Switches 1,500
   - TH6 200G Unit Price $66,336
   - Total Switching Cost $50M

   Cables/Optics:

   - NIC Transceivers (1.6T-DR8) 32,000 @ $900 (50% fewer vs. competition)
   - Leaf/Spine Transceivers (1.6T-DR8) 64,000 @ $900