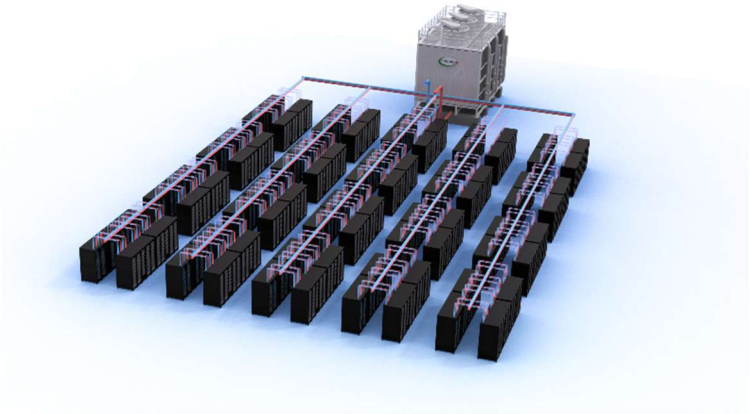
Validated Design

VALIDATED DESIGN FOR AI NETWORK CLUSTERS WITH SUPERMICRO, AMD, AND MICAS

CO-PACKAGED OPTICS WITH LOWER POWER, LOWER LATENCY, AND UP TO 2 KM OF REACH

TABLE OF CONTENTS

Executive Summary 1
Solution Overview 2
Solution Hypothesis & Validation 7
Conclusion 16
Contributors and References 17



Executive Summary

Artificial intelligence and machine learning (AI/ML) workloads are fundamentally reshaping the modern data center. Large-scale models, including generative AI and advanced recommendation systems, require thousands to tens of thousands of GPU accelerators operating as a single distributed system. At this scale, the network fabric becomes just as critical as compute, demanding predictable low latency, extreme bandwidth, and high operational efficiency.

One of the primary causes of low GPU utilization in large-scale AI training environments is network instability or link flapping, which can interrupt distributed communication and force portions of the training process to be recalculated. These interruptions reduce cluster efficiency and prolong overall training time.

To address this challenge, Supermicro, Broadcom, AMD, and Micas jointly deliver an open, standards-based AI infrastructure designed for the most demanding AI/ML workloads. This Co-Packaged Optics (CPO) architecture is particularly valuable for organizations seeking improved data center sustainability and extended connectivity. By integrating optical engines directly

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

with the switch ASIC, the CPO platform significantly reduces power consumption and thermal load while enabling high-bandwidth links reaching up to 2 kilometers that are essential to build a super spine architecture.

Compared with traditional pluggable optics architectures, this solution – built with Supermicro with AMD Instinct™ MI355X (Refer details to the following link: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/amd-instinct-mi355x-platform-brochure.pdf) , the Micas Tomahawk 5 CPO platform and AMD Pensando™ Pollara 400 AI NICs- delivers lower latency, improved link stability, reduced power consumption, and higher GPU utilization. These advantages translate into measurable improvements in AI training efficiency:

CATEGORY	KEY IMPROVEMENTS	BUSINESS IMPACT
CCL PERFORMANCE	• Accelerated collective communication library (CCL) completion • Reduced per-step training latency	Shortens end-to-end training cycles, enabling faster time-to-model and higher infrastructure productivity
RELIABILITY & STABILITY	• Proven stability across >1M cumulative port-hours • Zero link flaps observed • Consistent performance during thermal ramp-up/down	Eliminates re-compute overhead, maximizes GPU utilization, and ensures predictable large-scale training outcomes
ENERGY EFFICIENCY	• ~38% reduction in switch power consumption	Drives sustainable operations and significantly lowers total cost of ownership (TCO)
SCALE-OUT EFFICIENCY	• Near-linear scaling with only ~3% overhead vs. native throughput	Enables efficient expansion to large clusters without compromising performance or cost efficiency
OPEN ECOSYSTEM	• Support for Software for Open Networking in the Cloud (SONiC) • Integration with open-source ROCm stack • RDMA (remote direct access memory)-enabled high-performance data paths	Expands ecosystem flexibility, accelerates innovation, and reduces vendor lock-in

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Solution Overview

This section outlines a GenAI cluster solution that integrates AMD Instinct™ MI355X accelerators, Supermicro infrastructure, and networking components from Broadcom, Micas and AMD. Designed for both on-premises deployment and pre-integrated rack delivery, the architecture supports optimized AI workload execution with high-performance interconnects, flexible cooling options (liquid in Tomahawk 6 generation or air-cooled Tomahawk 5 generation), and dynamic resource partitioning for multi-tenant environments.

Sub-components At A Glance

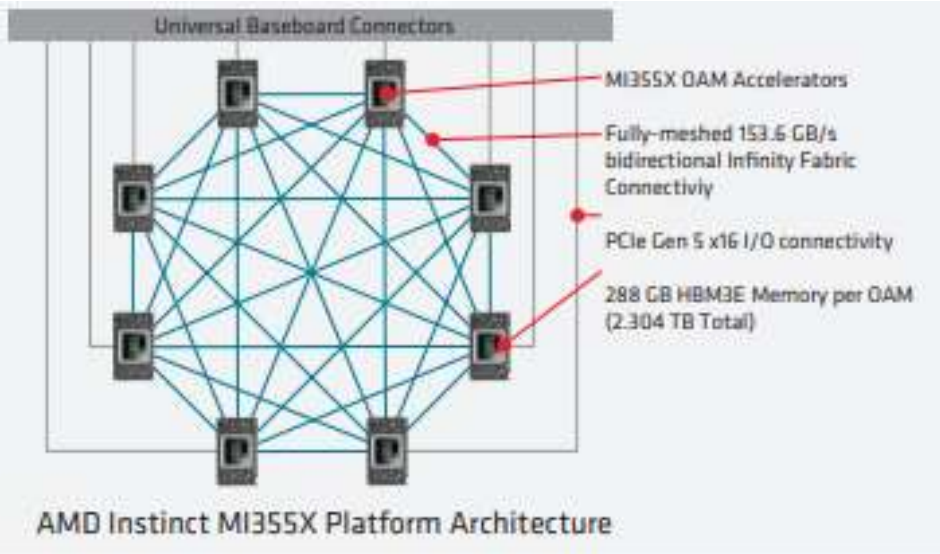
Supermicro DP AMD 4U Liquid-Cooled System with AMD Instinct™ MI355X 8-GPU: • Dual processor(s) AMD EPYC™ 9005/9004 Series Processors up to 500W with Liquid Cooling • Supports up to 24 DIMMs slots, 6400 MT/s 6TB DDR5 in 1DPC 	Micas Co-Packaged Optics (CPO) Switch: • 51.2 Tbps of full-duplex switching capacity • 128 ports of 400G FR4 connectivity • Designed for AI and high-performance computing 	Micas M2-W6940-64OC 800G Switch: • A Tomahawk 5–based system offering • 64 ports of 800GbE in a 2RU form factor, • Designed for AI/ML clusters requiring high-performance lossless RoCEv2 	AMD Pensando™ Pollara 400 AI NIC: • Up to 400 Gbps Ethernet connectivity • First Ultra Ethernet Consortium (UEC)-ready RDMA capable AI NIC • Fully hardware and software programmable • Designed to reduce network bottlenecks through intelligent load balancing and congestion management • Optimized for GPU-to-GPU communication in distributed AI training and inference

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Individual Products Details

Supermicro AMD Instinct™ MI355X Platform

The Supermicro AMD Instinct™ MI355X platform is designed to accelerate large-scale AI training and inference by combining high compute throughput with extremely large on-package memory. Each MI355X GPU includes 288 GB of HBM3E memory and up to 8 TB/s of memory bandwidth, allowing AI clusters to process massive datasets and large language models without frequent memory transfers or bottlenecks. Deployed in an 8-GPU OAM platform with over 2.3 TB of total GPU memory, the system enables dense AI nodes that can train or serve complex generative AI models more efficiently while scaling across high-speed Ethernet fabrics. Built on AMD's CDNA architecture and supported by the open ROCm™ software ecosystem, MI355X provides an open, high-performance alternative for enterprises and cloud providers building next-generation AI infrastructure for workloads such as large language models, recommendation systems, and scientific computing.



VIDEO DECODERS AND VIRTUALIZATION	
DECODERS[1]	32 groups for HEVC/H.265, AVC/H.264, V1, or AV1
JPEG/MJPEG CODEC	320 cores, 10 cores per group
GPU PHYSICAL PARTITIONS	SR-IOV, up to 64 partitions
MEMORY PARTITIONS	1 or 4 per module

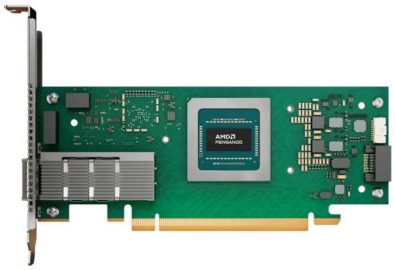
AI PEAK THEORETICAL PERFORMANCE		W/SPARSITY
FP16 VECTOR (TFLOPS)	1,258.4	N/A
FP16 MATRIX (PFLOPS)	20.1328	40.2656
BFLOAT16 MATRIX (PFLOPS)	20.1328	40.2656
INT8 MATRIX (POPS)	40.2656	80.5312
MXFP8 (PFLOPS)	40.2656	N/A
OCP-FP8 (PFLOPS)	40.5312	80.5312
MXFP6 (PFLOPS)	80.5304	N/A
MXFP4 (PFLOPS)	80.5304	N/A

HPC PEAK THEORETICAL PERFORMANCE	
FP64 VECTOR (TFLOPS)	628.8
FP32 VECTOR (PFLOPS)	1.3
FP64 MATRIX (TFLOPS)	628.8
FP32 MATRIX (PFLOPS)	1.3

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

AMD Pensando™ Pollara 400 AI NIC

The AMD Pensando™ Pollara 400 AI NIC is used to accelerate AI workloads and improve performance and efficiency in large-scale AI clusters by accelerating GPU-to-GPU communication at speeds up to 400 Gbps. Built for high-bandwidth, communication-intensive AI workloads, the AI NIC enhances Ethernet-based distributed training by improving traffic efficiency, reliability, and cluster uptime for collective communication. The first of its kind using UEC-ready RDMA, the AI NIC offers intelligent load balancing, path-aware congestion control, and fast failure recovery to help AI clusters scale efficiently across hundreds or thousands of nodes—helping improve GPU utilization and delivering more consistent training performance for workloads such as large language models, recommendation systems, and other data-intensive AI applications.



Micas Networks Tomahawk 5 Bailly CPO

Co-Packaged Optics (CPO) fundamentally addresses these limitations by integrating optical engines directly with the switch ASIC. This approach shortens electrical paths, reduces power consumption, and improves signal integrity, enabling more efficient high-bandwidth networking for AI clusters. This document evaluates the Micas Tomahawk 5 Bailly CPO switch platform in a Supermicro-validated environment using AMD EPYC™ CPUs, AMD Pensando™ Pollara 400 AI NICs, and RoCEv2 networking workloads relevant to large-scale AI training clusters. The evaluation combines several validation methodologies, including Physical-layer RDMA throughput testing, following Supermicro Solution & Integration Center validation methodology; ROCm Collective Communication Library (RCCL) performance testing, following CAST validation methodology; lower latency and reliability characterization of the networking platform.

Feature	Traditional Switch (Discrete Optics)	CPO (Co-Packaged Optics)	
Power Consumption	Higher power due to long electrical trace between ASIC and optical module, typical DSP power ~12 W (per 800 G port).	Integrated optics next to ASIC dramatically shortens electrical distance; DSP power reduced to ~5–6 W.	60% ~ power savings per port.
Latency	Longer electrical path and signal conversion delay — latency around 90–100 ns.	Optical I/O directly adjacent to ASIC; latency drops to 9–10 ns.	10× ~ lower latency, ideal for AI and HPC workloads.
Distance	Distance determined by external pluggable modules. Optics modules can reach ~2 km and require high-power DSP optics (~15–17 W per port) with longer electrical trace between ASIC and optics.	Optical engines integrated next to the ASIC dramatically shorten the electrical path. Supports FR4 optics up to ~2 km reach with improved signal integrity and lower optical power consumption.	Maintains 2 km data center interconnect reach while significantly reducing optical power, signal loss, and system complexity, enabling more efficient large-scale AI fabric deployment.
Signal Integrity (reduction of link flap)	Electrical-to-optical conversion across PCB causes loss and reflections.	Minimal electrical channel loss thanks to in-package connection.	Higher signal fidelity and better BER margin.
Thermal Efficiency	Discrete optics spread heat unevenly; cooling inefficiency.	Shared thermal design between ASIC and optics.	Better overall cooling and energy efficiency.
Form Factor / Density	Limited by front-panel optical module size (QSFP-DD, OSFP).	Optics integrated around ASIC; front panel only fiber breakout.	Enables smaller chassis and higher port density.
Application Suitability	Suitable for general networking.	Optimized for AI clusters, GPU interconnects, and data-center fabrics demanding ultra-low latency.	Critical enabler for next-gen AI networks.

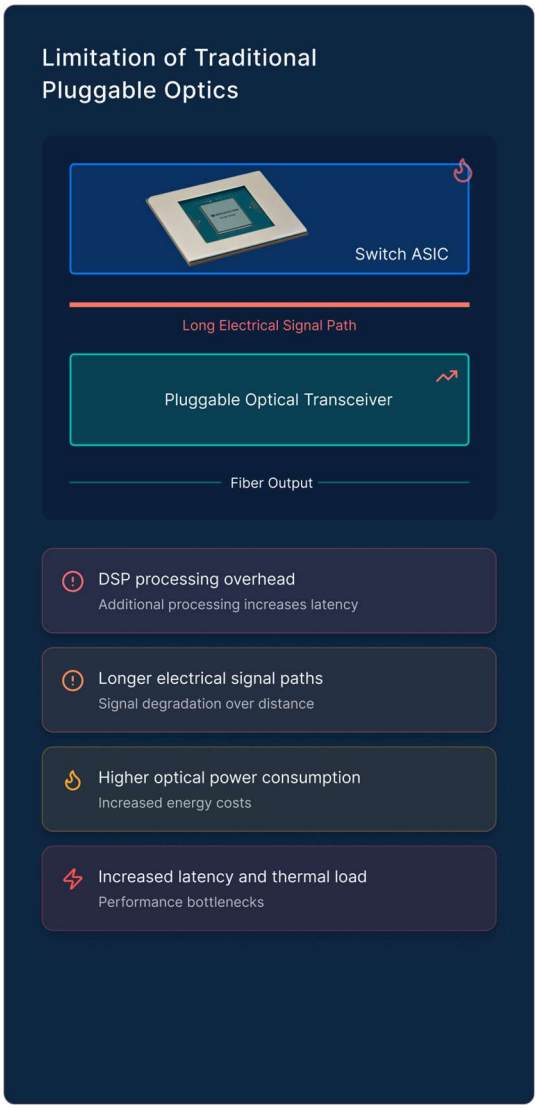
© 2026 Copyright Super Micro Computer, Inc. All rights reserved

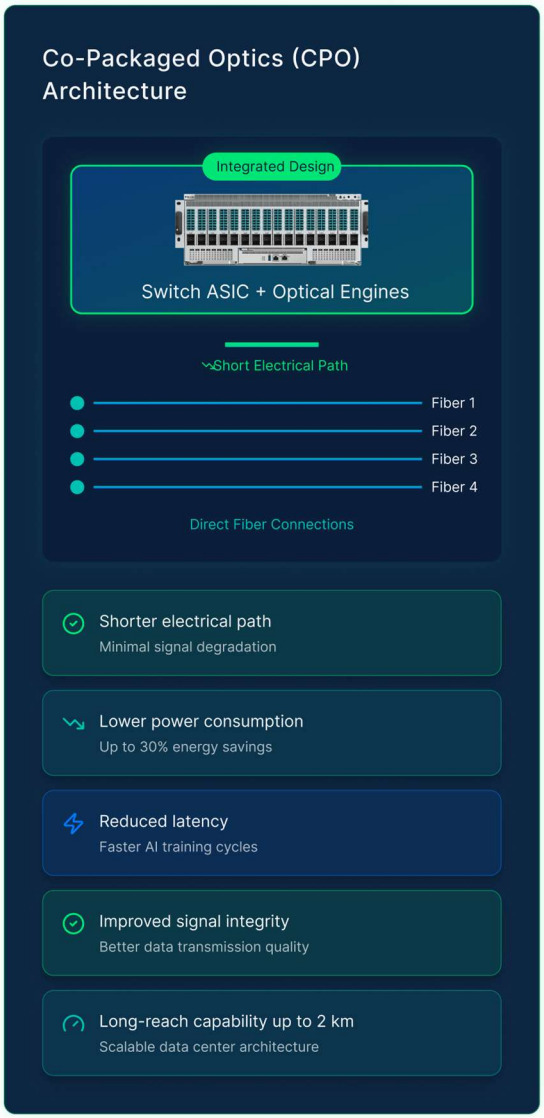
Testing has shown that CPO technology can reduce optics power consumption by up to 60% compared to traditional pluggable optics, while delivering superior link reliability. This validated design provides both business and technical leaders with a clear path forward—from a two-node evaluation cluster to multi-rack, multi-thousand GPU production deployments. The architecture demonstrates how the joint solution delivers:

- Performance at scale: Predictable, lossless networking optimized for distributed AI training.
- Efficiency: Lower power consumption and reduced thermal requirements through next-generation CPO technology.
- Openness: Support for open-source SONiC networking software and flexible deployment models.

In addition to performance validation, architecture also demonstrates how CPO networking with FR-class optics supporting up to 2 km reach can enable long-distance AI interconnects across large data center environments. This capability allows the platform to be positioned as a Super Spine solution connecting multiple AI halls or data center buildings, enabling scalable deployment of large-scale AI clusters.

- Excessive optical power consumption and cooling overhead up
- Increased packet latency caused by DSP processing and longer electrical signal paths
- Link instability under thermal cycling
- Higher operational risk at hyperscale deployment sizes





© 2026 Copyright Super Micro Computer, Inc. All rights reserved

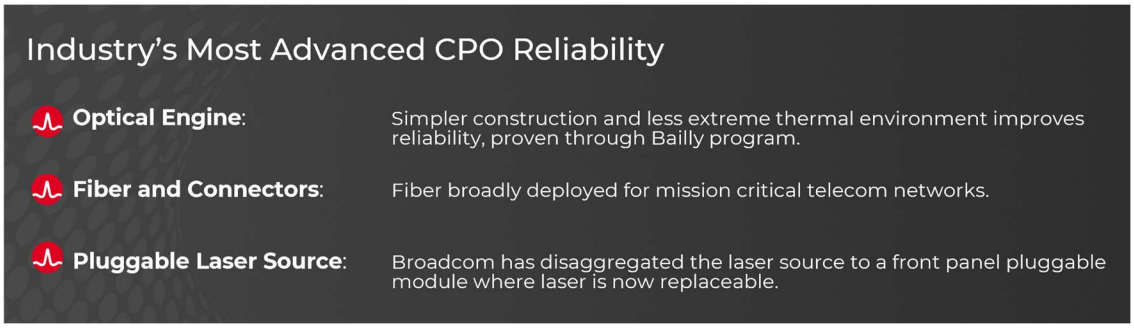
Solution Hypothesis & Validation

The Supermicro Rack Scale Solutions provide fully integrated, tested, and optimized rack-level infrastructure for AI, HPC, and cloud workloads, enabling organizations to deploy scalable clusters from individual nodes to massive Super POD systems. Designed for high-density computing environments, these solutions support the latest processors and accelerator platforms while integrating high-performance networking, advanced system management, and flexible rack-scale architecture using open standards such as Redfish. Supermicro's in-house liquid cooling technologies, including Coolant Distribution Units (CDUs), support up to 250 kW per rack to improve power efficiency and enable dense AI deployments. Delivered as turnkey, plug-and-play systems with integrated power, cooling, and networking, Supermicro Rack Scale Solutions allow enterprises and cloud providers to rapidly deploy energy-efficient infrastructure while maintaining the flexibility to customize configurations through Supermicro's Building Block architecture to meet specific performance and scale requirements.

The Micas Co-Packaged Optics (CPO) switch, M2-W6940-128X1-FR4 with 128 400G ports, integrates silicon photonics optical engines directly with the switch ASIC, reducing the electrical trace distance between the switching silicon and the optical interface. The Micas CPO switch integrates Broadcom Tomahawk 5 Bailly (51.2 Tb/s) with co-packaged optical engines and remote laser modules (RLMs). Key architectural advantages include power efficiency, eliminating DSP-based pluggable optics, shortens electrical channels dramatically, reduces optical power per 400G/800G port; latency reduction, no retimer/DSP buffering, fewer pipeline stages between MAC and optical engine, measured lower packet latency vs pluggable; reliability, removes module cages, connectors, and thermal heats.

Hypothesis

The reliability profile of Co-packaged Optics (CPO) technology demonstrates significant advantages over traditional architectures, positioning it as a robust solution for next-generation systems. The optical engine features a simpler construction and operates in a less extreme thermal environment, a benefit validated through programs like Bailly. Furthermore, CPO leverages mature fiber and connector technologies, which have a proven track record in mission-critical deployments. Crucially, the disaggregation of the laser source into a front-panel pluggable module introduces enhanced serviceability and uptime by allowing laser replacement. These factors collectively contribute to the CPO architecture's superior reliability, making it an ideal choice for demanding, high-availability data center applications. Further for details on reliability, please refer to the following link: https://www.broadcom.com/company/news/product-releases/63616

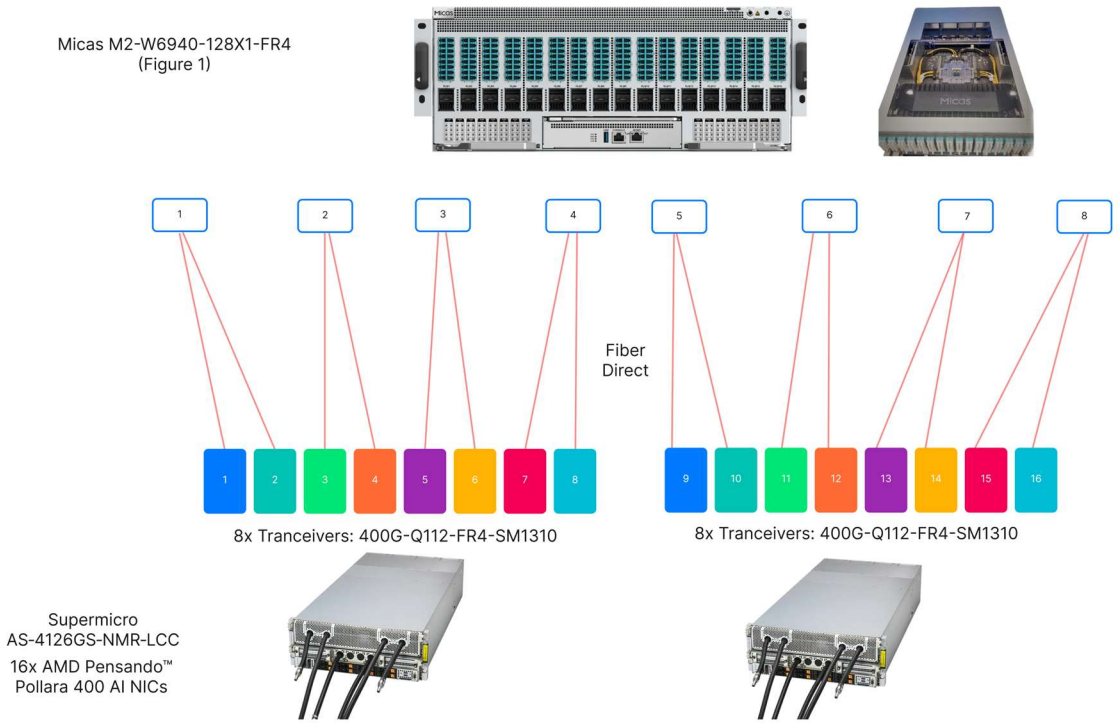


© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Test Topology and Methodology

400G CPO Switch Validated Design Topology

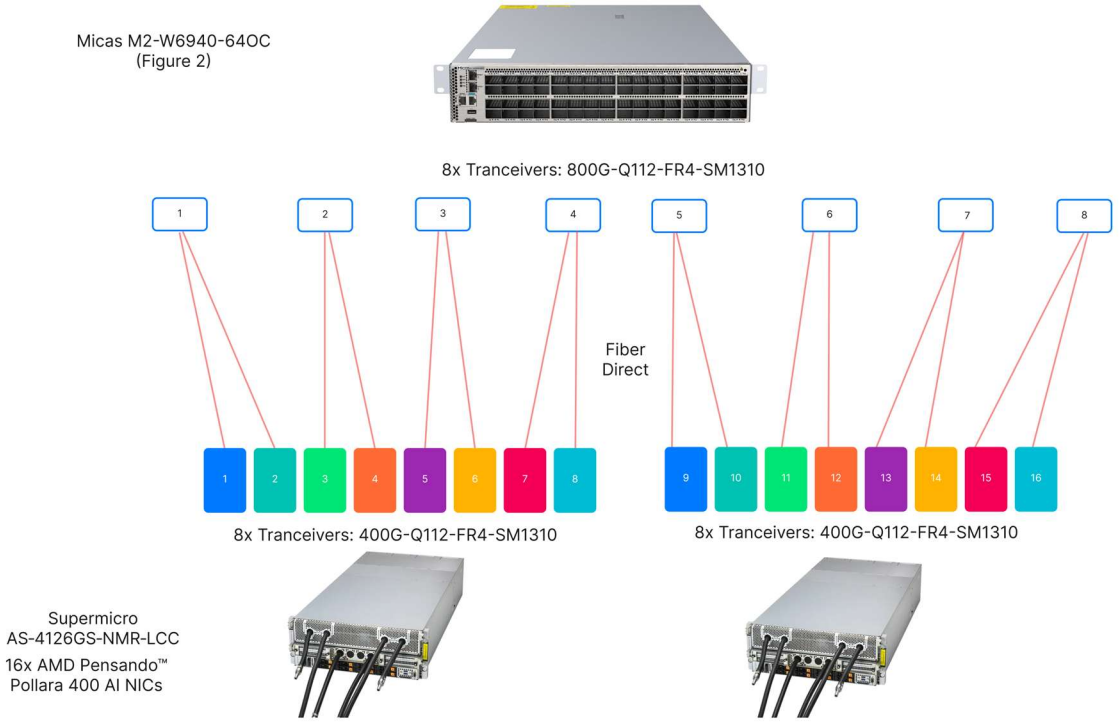
Figure 1 below illustrates a RDMA throughput topology with direct optical connection between the Micas M2-W6940-128X1-FR4 CPO switch and two Supermicro AS- 4126GS-NMR-LCC GPU servers using AMD Pensando™ Pollara 400 AI NICs. Unlike conventional switches in Figure 8, the CPO switch does not require pluggable transceivers on the switch side because the optical engines are co-packaged with the Tomahawk 5 ASIC. Instead, fiber cables connect directly from the switch's optical ports to external 400G FR4 transceivers located on the server side. Each server uses 8× 400G-Q112-FR4-SM1310 transceivers, which connect to 8 AMD Pensando™ Pollara 400 AI NIC ports, for a total of 16 links from the switch across the two servers. This architecture demonstrates a low-latency AI fabric design, as eliminating switch-side pluggable optics removes DSP/retimer stages and reduces electrical path length, resulting in lower latency and improved signal integrity for high-performance AI cluster networking.



© 2026 Copyright Super Micro Computer, Inc. All rights reserved

800G Convention Switch Validated Design Topology

Figure 2 illustrates a networking configuration using the Micas M2-W6940-64OC switch, where connectivity to two Supermicro AS-4126GS-NMR-LCC GPU servers equipped with AMD Pensando™ Pollara 400 AI NICs is implemented through pluggable optical transceivers on the switch side. In this architecture, the switch requires 400G-Q112-FR4-SM1310 transceivers, and each 800G switch port is broken out into two 400G links using DAC breakout cables. These breakout connections feed 16 total AMD Pensando™ Pollara 400 AI NIC ports across the two servers, with 8 AMD Pensando™ Pollara 400 AI NICs per server, each populated with corresponding 400G optical transceivers. Because this design relies on traditional pluggable optics on the switch, the data path includes additional DSP/retimer stages and longer electrical channels, which increases processing overhead compared to CPO architecture. As a result, this configuration does not provide the same low-latency advantage, since the presence of switch-side transceivers introduces additional latency and power consumption typical of conventional pluggable-optic switch designs used in AI cluster networking.



© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Validated Design Equipment and Configuration

Item Name	Model Name	Qty
System	AS-4126GS-NMR-LCC	1
Motherboard	H14DSG-OD	1
Processor	AMD EPYC 9575F 64-Core Processor	2
Memory	MTC40F2047S1RC64BB1	24
GPUs	AMD Instinct MI355X	8
Disk	Micron_7450_MTFD	1
Disk	Micron_7450_MTFDKCC3T8TFR	8
NIC cards	AOM-DP805-IO	1
NIC cards	AOC-S400G-B1C	2
NIC cards	POLLARA-1Q400P	8
Power Supply	PWS-6K61G-2R	4
Fans	NA	3

Performance Test

To evaluate AI collective behavior beyond raw RDMA bandwidth, the test extends the topology using RCCL workloads:
In large AI fabrics, the most challenging condition for collective communication is multi-hop, multi-switch traversal, where congestion, path imbalance, and latency amplification typically cause noticeable bandwidth degradation. In traditional RoCE deployments, performance often drops sharply as traffic crosses from a single-switch domain to a leaf–spine–leaf topology.

In this evaluation, the test fabric starts with 1 DUT (single switch, same rail group). In the future, we will scale it to 3 DUT (leaf–spine–leaf, three-hop non-blocking fabric).

BusBW (Bus Bandwidth) is the effective communication bandwidth observed by GPUs during collective operations such as AllReduce, AlltoAll, AllGather, and ReduceScatter. Unlike raw link speed, BusBW captures the end-to-end performance seen by applications, accounting for protocol overhead, multi-hop latency, congestion, and synchronization effects. It reflects how efficiently data is exchanged across GPUs during real AI training workloads. As a result, BusBW is a more meaningful metric than port throughput or link bandwidth, because it directly correlates with collective completion time and overall training efficiency. Higher BusBW indicates better scaling behavior and higher effective GPU utilization.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

In conventional designs, multi-switch fabrics suffer from:

- ECMP path collisions caused by a small number of large “elephant” flows
- Micro-burst congestion at spine links
- Increased tail latency, which delays collective completion
- Over-reliance on PFC, leading to pauses and head-of-line blocking

These effects typically compound with hop count, making 3-hop or 4-hop fabrics significantly slower than single-switch baselines. CPO-based fabric may improve the BusBW, driven by lower per-hop latency and reduced jitter during collective operations.

Topology:

- 1 x CPO switch
- Supermicro Server (AMD EPYC) x 2
- GPU: AMD Instinct™ MI355X x 16
- NIC: AMD Pensando™ Pollara 400 AI NIC x 16
- ECN + PFC enabled with default parameters
- AMD GPUs using RCCL

Workloads

- AllReduce (Ring)
- AlltoAll
- AllGather
- ReduceScatter
- Message sizes: 64 MB – 512 MB

MlPerf Training Benchmark

An MLPerf training benchmark was performed on the following 1 CPO switch, 2 nodes setup. In the future, we plan to run the benchmark testing on a 3-node cluster connected through a 2-leaf, 1-spine non-blocking fabric, comparing a traditional pluggable-optics switch against the Micas Tomahawk 5 Bailly CPO switch. Both setups used identical Supermicro servers, AMD GPUs, AI NICs, software stacks, and MLPerf configurations. CPO-based fabric may reduce end-to-end training time, driven by lower per-hop latency and reduced jitter during collective operations. (NOTE: This remains a hypothesis that requires testing to validate or invalidate).

Topology:

- 1x CPO switch
- Supermicro Server (AMD EPYC) x 2
- GPU: AMD Instinct MI355X x 16
- NIC: AMD Pensando™ Pollara 400 AI NIC x 16
- ECN + PFC enabled with default parameters
- AMD GPUs using RCCL

Workloads: Two Node MLPerf 5.1 Llama2-70B Inferencing

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Test Results

Power Consumption

The power measurement results show a significant reduction when using the Tomahawk 5 Bailly CPO configuration compared to the Micas conventional Tomahawk 5 M2-W6940-64OC with eight 800Gx FR4 pluggable transceivers. The Tomahawk 5 CPO is configured equivalent to eight 800G 2x FR4 transceiver. The Tomahawk 5 CPO system consumed 465.87 W, while the pluggable Tomahawk 5 configuration required 754.85 W, resulting in a power reduction of 288.98 W, or approximately 38.28% savings. This result highlights the efficiency advantage of CPO architecture by eliminating DSP-based pluggable optics and reducing electrical interface losses, enabling lower overall power consumption while maintaining the same switching ASIC capability.

TH5-CPO vs TH5 (Eight 800G 2x FR4 Transceivers)

TH5-CPO Power (W)	TH5 Power (W)	Delta	Power Saving
465.87	754.85	288.98	38.28%

Power consumption comparison showing 38.28% efficiency improvement with CPO architecture

Throughput and Latency Performance

Tomahawk 5 vs Tomahawk 5 CPO Latency Comparison

Platform	Pol-1 (µs)	Pol-2 (µs)	Pol-3 (µs)	Pol-4 (µs)	Pol-5 (µs)	Pol-6 (µs)	Pol-7 (µs)	Pol-8 (µs)	Average (ns)
TOMAHAWK 5	4.99	5.04	5.05	4.94	4.95	5.04	5.04	4.98	5003.75
TOMAHAWK 5-CPO	4.89	4.96	4.95	4.88	4.90	4.99	4.97	4.90	4930.00
Difference	0.10	0.08	0.10	0.06	0.05	0.05	0.07	0.08	73.75

The Tomahawk 5 CPO architecture demonstrates a measurable latency improvement of 73.75 ns across all Pollara NIC ports.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Latency measurements using the ib_send_lat benchmark from the Perftest suite show that introducing a CPO-based switch reduces end-to-end message latency by approximately 73 ns (about 1.5%) in a two-node, single-switch topology. This reduction reflects the shorter electrical path and fewer high-speed electrical interfaces in the CPO architecture compared with traditional DSP-based pluggable optics. Although the absolute latency improvement is modest in this small topology, it demonstrates that CPO can reduce per-hop communication delays. Such improvements can become increasingly beneficial in large-scale GPU clusters where collective communication operations traverse multiple switches and links.

CPO shows identical ib_send_bw performance compared to the non-CPO configuration because both systems use the same switch ASIC, resulting in the same bandwidth capability and data-path processing.

CPO vs Non-CPO Throughput Comparison

#bytes	Non CPO	CPO
65,536	392.03	392.04
65,536	392.04	392.03
65,536	392.02	392.05
65,536	392.04	392.03
65,536	392.06	392.02
65,536	392.02	392.04
65,536	392.01	392.02
65,536	392.02	392.03
Average	392.03	392.03

Performance measurements using 65,536-byte packets demonstrate equivalent throughput between CPO and traditional pluggable optics architectures.

Reliability and Thermal Stability

Reliability testing aligns with Broadcom CPO qualification data:

- >1M cumulative port-hours
- Zero link flaps
- Stable operation during temperature ramp-up/down

For Supermicro deployments, this eliminates:

- Optical-related link resets
- Training job failures due to fabric instability
- Operational risk during thermal excursions

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

RCCL Performance Testing Result

- Topology: 1x CPO switch
- Supermicro Server (AMD EPYC) x 2
- GPU: AMD Instinct™ MI355X x 16
- AMD Pensando™ Pollara 400 AI NIC x 16
- ECN + PFC enabled with default parameters
- AMD GPUs using RCCL

Workload:

- AllReduce

The test results show that the CPO switch increases the measured RCCL AllReduce bus bandwidth by approximately 0.7–0.9%. This improvement indicates that the reduced electrical path length and lower end-to-end link latency enabled by the CPO architecture can slightly enhance collective communication efficiency.

Although the topology is small, the results demonstrate that lower-latency optical interconnects can improve RCCL bus bandwidth during AllReduce operations, suggesting that CPO-based fabrics can provide incremental performance benefits for scale-out GPU clusters.

Two-Node RCCL All Reduce Performance Comparison

Two-Node Non-CPO RCCL All Reduce

Run#	out-of-place busbw (GB/s)	in-place busbw (GB/s)
1	363.28	363.86
2	364.97	364.11
3	366.70	366.62
Average	**364.98**	**364.86**

Two-Node CPO RCCL All Reduce

Run#	out-of-place busbw (GB/s)	in-place busbw (GB/s)
1	368.75	368.78
2	368.78	366.82
3	366.97	367.02
Average	**368.17**	**367.54**

Metric	out-of-place	in-place
CPO Improvement	**0.9%**	**0.7%**

MLPerf testing result

Topology:

- 1 CPO switch
- Supermicro Server (AMD EPYC) x 2
- GPU: AMD Instinct™ MI355X x 16
- NIC: AMD Pensando™ Pollara 400 AI NIC x 16
- ECN + PFC enabled with default parameters
- AMD GPUs using RCCL

Workloads

- Two Node MLPerf 5.1 Llama2-70B Inferencing
- Test 1: 2 nodes running independently without ethernet connection
- Test 2: 2 nodes with 1 CPO switch

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

AI Inference Performance Comparison (CPO vs Non-CPO Networking)

Test Scenario	Node 1 (tokens per sec)	Node 2 (tokens per sec)	Total
Test 1: 2 Nodes without switch	88,709	89,241	177,950
Test 2: 2 Nodes with CPO switch	172,635	172,618	171,956
Difference	2.99%	3.00%	3.37%

Based on the test results, introducing a CPO-based switch into the AI fabric results in only a minimal overhead when scaling out the cluster – only 3.37% overhead. This small overhead demonstrates that CPO integration does not materially affect application-level performance while enabling significantly improved horizontal scalability. By leveraging high-radix switching and high-bandwidth optical interconnects, the CPO architecture allows AI clusters to scale out efficiently across a larger number of nodes while maintaining near-native training and inference throughput.

Value

The Micas Tomahawk 5 Bailly CPO switch, validated in a Supermicro AMD GPU + AMD Pensando™ Pollara 400 AI NIC environment, delivers clear and measurable advantages:

- Lower power and cooling cost
- Lower latency for AI collectives
- Exceptional link reliability under thermal stress
- Seamless compatibility with existing RoCEv2 and CCL software stacks

This evaluation confirms that CPO is production-ready for Supermicro AI platforms, providing immediate TCO, performance, and reliability benefits without architectural risk.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Side-by-Side Comparison: Pluggable Optics vs CPO (TOMAHAWK 5)

Architectural Comparison

Dimension	Pluggable Optics	CPO (Micas TOMAHAWK 5)
Optical location	Front-panel modules	Co-packaged with ASIC
Electrical reach	Long SerDes traces, connectors	Ultra-short die-to-optic
DSP requirement	Required (DSP) / Partial (LPO)	Not required
Per-400G optical power	~16 W (DSP) / ~7.5 W (LPO)	~5.5 W
Switch system power	Highest	Lowest
Cooling demand	High, front-panel hotspots	Lower, evenly distributed
Port density scaling	Limited by thermal budget	Scales cleanly to 51.2T
Failure points	Modules, cages, connectors	Reduced (no pluggables)
Link stability	Sensitive to temperature	Stable under thermal cycling
Software changes	None	None

Performance & Operations Comparison

Category	Pluggable-Based Fabric	CPO-Based Fabric
Port-to-port latency	Higher (DSP buffering)	5–8% lower
Latency jitter	Higher under load	Lower, more deterministic
Link flap risk	Increases with temperature	Demonstrated zero flaps
Maintenance	Module replacements	Fewer field failures
Cluster TCO	Higher	Significantly lower

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Conclusion

Our solution test validates the hypothesis that "By eliminating network-induced inefficiencies, we convert infrastructure from a bottleneck into a scaling advantage—delivering faster time-to-model, lower TCO, and production-ready reliability**".**

To Recap:

Hypothesis
Network instability and inefficient collective communication are primary bottlenecks to GPU utilization in large-scale AI clusters. Improving CCL efficiency and fabric reliability will:

- Increase effective GPU utilization
- Reduce training step time
- Minimize costly re-compute cycles
- An open, optimized stack (network + software) can deliver both performance and ecosystem flexibility

Validation Approach (Solution Testing)
Deployed AI clusters with optimized fabric (CPO-based Ethernet + RDMA + CCL tuning)
Executed distributed training workloads at scale
Measured across key dimensions:

- CCL completion latency
- End-to-end training step time
- GPU utilization under sustained load
- Network stability (link flaps, thermal conditions)
- Power consumption and efficiency

Results & Outcomes:

- Faster Training: Reduced step time driven by improved CCL performance
- Higher Utilization: Eliminated re-compute from network instability
- Proven Reliability: >1M port-hours with zero link flaps
- Energy Savings: ~38% lower switch power consumption
- Scalable Architecture: ~3% overhead at cluster scale
- Open Innovation: Enabled via SONiC, ROCm, and RDMA ecosystem

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Contributors

This validated design reflects a close and trusted collaboration across the AI infrastructure ecosystem. Micas Networks led networking architecture, open networking integration, and deployment validation, drawing on real-world hyperscale and enterprise AI cluster experience. Broadcom provided foundational silicon, CPO technology leadership, and deep technical guidance on Tomahawk-class switching and Ethernet-based AI fabrics. Supermicro contributed system design expertise, platform integration, performance benchmarking, power measurement, and rack-level deployment insights, ensuring alignment with production-ready AI infrastructure requirements. AMD supported GPU-centric workload considerations, performance characteristics, and AI compute architecture alignment.

We would like to recognize and thank the following individuals for their technical leadership, architectural input, and hands-on collaboration throughout the development of this validated design: from Micas Networks, Tim Zhou and Kris Tsao; from Supermicro, Max Chung, Ziming Zhu, Reeann Zhang, Linda Yang, Dhanashree Karlekar, and Allen Huang; and from AMD, Jacqueline Nguyen and Jessi McKain. Their combined expertise and partnership were instrumental in delivering a scalable, power-efficient, and production-validated AI networking validated solution.

References

- Supermicro MI355X Platform: https://www.supermicro.com/en/products/system/gpu/4u/as%20-4126gs-nmr-lcc
- Micas Networks Industry First 51.2T Co-Packaged Optics Switch: https://micasnetworks.com/company/news/30202
- AMD Instinct™ MI355 Platform: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/amd-instinct-miI355x-platform-brochure.pdf
- Broadcom CPO Press Release: https://www.broadcom.com/company/news/product-releases/61946
- AMD Pensando™ Pollara 400 AI NIC https://www.amd.com/en/products/network-interface-cards/pensando.html

© 2026 Copyright Super Micro Computer, Inc. All rights reserved