

SOLUTION BRIEF

Scaling AI with Open Ethernet

Validated AMD Instinct™ MI350X AI Infrastructure with Supermicro and Arista Networks

SEPTEMBER 2026



Supermicro GPU Server AS-8126GS-TNMR

Table of Contents

Executive Summary	2
Common Implementation Issues	2
Necessary Design Objectives	3
Validated Reference Architecture from Supermicro, AMD, and Arista	3
Compute Platform	4
Networking Infrastructure	5
AI Networking Components	5
Architecture Design	6
Scale-Out Support	6
Operational Visibility and Telemetry	7
Multi-Tenant Design	7
Validation	8
Test Environment	8
Technical Validation Results	10
Business Impact	11
Deployment Scenarios	11
Scaling from Pilot to Production	11
Scalable Fabric Design	11
Conclusion	12
More Information	12

EXECUTIVE SUMMARY

Demand for AI infrastructure is exploding, as neocloud providers and enterprises expand AI deployments. Organizations are deploying copilots, RAG applications, AI assistants, and domain-specific models that require increasingly scalable compute and networking resources.

At the same time, growing GPU density and the shift towards large-scale inference workloads are driving higher east-west traffic across AI clusters. As user expectations for responsiveness continue to increase, reducing time to first token (TTFT) and maintaining consistent token throughput have become critical infrastructure objectives.

Supermicro, AMD, and Arista have validated an open Ethernet architecture for scaling AMD Instinct MI350X clusters. Testing demonstrated comparable AllReduce and AllGather performance while optimized congestion management and load balancing improved All-to-All bandwidth by up to 10x in the tested configuration.

- Supermicro, an AI Factory Infrastructure provider, delivers compute density, rack integration, DLC, deployment, and hardware lifecycle management;
- AMD provides EPYC™ CPUs, Instinct MI350X accelerators, and Pollara™ AI NICs driving computation and GPU-to-GPU communication;
- Arista provides RoCEv2-enabled spine-leaf fabric, congestion management, and telemetry for operating distributed AI workloads at scale.

Testing on an Arista AI Center of Excellence testbed validated the design:

- Improved all-to-all bandwidth with advanced load balancing (DLB/CLB) and congestion management (PFC/ECN)

RCCL Collective	Default	8×800G (CLB/DLB + PFC/ECN)	Improvement
All-to-All (2 nodes)	19.81 GB/s	89.85 GB/s	4.5x
All-to-All (4 nodes)	5.74 GB/s	59.12 GB/s	10x

Table: Collective Communication Bus Bandwidth (16 GB Messages)

- AllReduce and AllGather performance comparable to default configurations
- Built on open standards—Ethernet, BGP, RoCEv2, and AMD ROCm—avoiding proprietary lock-in

All-to-All is particularly relevant to distributed workloads where traffic patterns become less uniform, including architectures involving expert routing / MoE-style communication patterns. For enterprises, neoclouds, and AI service providers, this means faster time to value, better GPU utilization, lower risk, and a proven path from pilot to production scale, offering a credible, multi-vendor alternative to vertically integrated AI infrastructure.

COMMON IMPLEMENTATION ISSUES

As environments scale, organizations frequently encounter five major obstacles:

- 1. Network Bottlenecks**
Distributed AI workloads depend on continuous communication among GPUs. Collective communication operations such as AllReduce, AllGather, ReduceScatter, and All-to-All generate substantial east-west traffic. Network inefficiencies directly reduce GPU effectiveness.
- 2. GPU Underutilization**
Expensive GPU resources often spend significant time waiting for data movement operations to complete. Congestion, packet loss, retransmissions, and latency variation can reduce effective utilization.

3. Operational Complexity

AI infrastructure combines servers, accelerators, NICs, switches, storage, software frameworks, orchestration systems, and observability platforms. Managing these components at scale creates operational challenges.

4. Power Consumption

Power and cooling requirements have become major constraints. Traditional optical networking architectures contribute significantly to overall power consumption.

5. Multi-Tenancy Requirements

Many organizations require shared infrastructure supporting multiple business units, customers, or workloads. Resource isolation and predictable performance become essential design objectives.

Individually, each of these challenges can affect infrastructure efficiency. Collectively, they determine whether an AI initiative can successfully transition from a small proof-of-concept environment into a scalable production platform. As organizations increase GPU counts and inference capacity, the ability to address networking, utilization, operational, and power constraints through a unified architecture becomes increasingly important.

NECESSARY DESIGN OBJECTIVES

Organizations expanding AI infrastructure generally require seven capabilities. These objectives directly inform the design decisions incorporated throughout the validated architecture.

1. Capacity

Infrastructure must support growing GPU populations without redesigning the entire architecture.

2. Performance

Low latency, high throughput, and efficient collective communication are essential.

3. Scalability

Architectures must scale from small evaluation clusters to production environments containing thousands of accelerators.

4. Operational Simplicity

Deployment, monitoring, troubleshooting, and lifecycle management should remain manageable as environments grow.

5. Sustainability

Power efficiency and cooling optimization increasingly influence purchasing decisions.

6. Open Architectures

Organizations seek standards-based solutions that avoid unnecessary vendor lock-in.

7. Supply Chain Diversity

Multi-vendor ecosystems provide greater flexibility and resilience than vertically integrated alternatives.

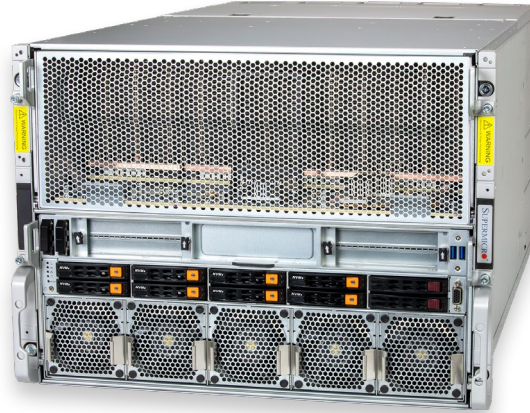
VALIDATED REFERENCE ARCHITECTURE FROM SUPERMICRO, AMD, AND ARISTA NETWORKS

The value of this solution does not come from any single component but from the integration of compute, networking, transport, observability, and operational automation into a repeatable design that can be deployed consistently across AI environments. Supermicro provides the rack-scale infrastructure foundation, AMD delivers the accelerator and networking technologies that drive AI computation and communication, and Arista provides the Ethernet fabric, telemetry, and operational visibility required to scale distributed AI workloads.

Together, these technologies create an architecture designed to help organizations expand AI capacity, improve infrastructure utilization, simplify operations, and move from pilot deployments to production-scale environments with greater confidence.

COMPUTE PLATFORM

Supermicro GPU Server AS-8126GS-TNMR



The solution uses Supermicro rack-scale infrastructure optimized for AI deployments. Supermicro provides integrated rack-level solutions supporting deployment from small clusters to large-scale AI environments. The architecture supports both air-cooled and liquid-cooled implementations while maintaining consistent operational models.

Each server includes:

- Dual AMD EPYC 9575F 64-Core processors
- Eight AMD Instinct MI350X GPU accelerator modules interconnected with AMD Infinity Fabric™ Link
- DDR5 ECC system memory across 24x96GB DDR5 DIMMs (2.3TB total)
- Pensando 400GbE, QSFP112, PCIe Gen5.0 x16, CPU-to-GPU connectivity for high-bandwidth accelerator access
- Eight front-accessible PCIe® 5.0 NVMe drive bays for high-performance local storage
- Direct-to-chip liquid cooling supporting CPUs up to 500 W and high-density GPU deployments
- Four redundant 6600 W Titanium (96% efficient) power supplies for resilient operation
- Integrated 10GbE management connectivity and Supermicro infrastructure management tools

For this solution, Supermicro also goes beyond the server hardware, providing validated rack-scale integration, power and cooling design, infrastructure packaging, and deployment consistency to help reduce implementation complexity. Organizations can deploy AI capacity using a repeatable operational model rather than being required to assemble individual infrastructure components independently.

AMD Instinct MI350X Accelerators

The AMD Instinct MI350X platform provides the computational power in the architecture.

Each accelerator includes:

- 288GB HBM3E memory (High Bandwidth Memory 3 Extended)
- Up to 8 TB/s memory bandwidth
- CDNA architecture (specialized GPU architecture from AMD)
- ROCm™ software ecosystem support (AMD Radeon Open Compute)

In an eight-GPU configuration, a single server provides more than 2.3 TB of aggregate GPU memory, supporting large AI models and memory-intensive workloads such as training and inference.

Why it Matters

For infrastructure operators, high memory capacity and bandwidth help reduce model partitioning requirements and support larger AI models within fewer servers. This can improve overall infrastructure efficiency while simplifying deployment architectures for both training and inference workloads.

NETWORKING INFRASTRUCTURE

Arista Ethernet Fabric

The architecture uses a routed Ethernet spine-leaf fabric based on BGP. The Arista AI Center of Excellence (COE) testbed demonstrates a scalable two-tier design using layer-3 connectivity to GPU hosts and BGP-based reachability across the fabric.

The architecture leverages:

- Arista Switches – DCS-7060X6-64PE
- 64 ports of 800 GbE OSFP in 2RU
- Support for 2x400 GbE and 800 GbE optics for connectivity to the GPUs
- Support for low-power and highly reliable Linear Pluggable Optics (LPO)
- ECMP Forwarding
- Advanced load-balancing technologies (DLB, CLB)
- Congestion-aware transport mechanisms

800G Networking

Modern AI clusters require substantial bandwidth between servers and fabric switches. The design supports 400G and 800G Ethernet connectivity, enabling non-blocking communication across scale-out environments.

AI NETWORKING COMPONENTS

RDMA and RoCEv2

The architecture uses RDMA over Converged Ethernet version 2 (RoCEv2), enabling direct memory access across Ethernet fabrics while minimizing CPU overhead.

RoCEv2 provides:

- Low-latency communication
- High throughput
- Efficient collective operations
- Scalable GPU-to-GPU communication

AMD Pollara AI NIC

The AMD Pensando Pollara 400 AI NIC provides up to 400 Gbps Ethernet connectivity and supports advanced congestion management, helping optimize GPU-to-GPU communication across distributed AI fabrics.

Capabilities include:

- Ultra Ethernet Consortium-ready RDMA
- Intelligent load balancing
- Congestion management
- Fast failure recovery
- GPU-to-GPU communication optimization

Why it Matters

As AI environments scale, communication efficiency becomes increasingly important to overall infrastructure performance. By helping optimize GPU-to-GPU communication and reducing the effects of congestion, the Pollara AI NIC contributes to higher effective accelerator utilization and more predictable application behavior.

ARCHITECTURE DESIGN

Overview

The validated reference architecture combines Supermicro AI servers, AMD Instinct GPUs, AMD Pollara networking adapters, and an Arista Ethernet fabric into a unified platform designed to support both AI training and large-scale inference workloads. Within each server, GPUs process AI workloads and exchange data through high-speed internal interconnects. As workloads expand across multiple servers, Pollara adapters provide efficient GPU-to-GPU communication over a RoCEv2-enabled Arista fabric. Dynamic load balancing, congestion management, and telemetry visibility work together to maintain consistent performance as traffic volumes increase. The result is an architecture designed to scale AI capacity while preserving operational visibility and deployment simplicity.

Arista CloudVision as a Service (CVaaS) is a cloud-hosted network operations platform that provides centralized management, telemetry collection, analytics, and automation for Arista-based environments. In AI fabrics, CVaaS enables operators to monitor congestion indicators such as PFC events, ECN markings, interface utilization, and flow behavior, helping validate performance and troubleshoot issues as clusters scale.

Architecture Detailed

- CVaaS telemetry and AI Agent capabilities provide end-to-end visibility across both network and compute infrastructure.
- Flows in this design vary based on the collective in use. Some traffic was locally switched on the same leaf switch, and some traversed the spines to reach the second pair of servers.
- Leaf switches were connected to spine switches with 8x800G links.

Scale-Up Support

Within each server, GPUs communicate through high-bandwidth internal interconnects and PCIe infrastructure designed to minimize latency and maximize locality.

The architecture emphasizes:

- Direct GPU connectivity
- High-bandwidth memory
- PCIe optimization
- NUMA-aware placement

These characteristics reduce communication overhead before traffic reaches the external network.

SCALE-OUT SUPPORT

Spine-Leaf Fabric

The AI COE environment uses a spine-leaf architecture with multiple uplinks between leaves and spines, creating a scalable non-blocking fabric.

Lossless Ethernet

Lossless Ethernet is achieved through coordinated use of:

- Priority Flow Control (PFC)
- Explicit Congestion Notification (ECN)
- Queue tuning
- Data Center Quantized Congestion Notification (DCQCN)

Congestion Management

Traditional ECMP alone often produces uneven traffic distribution across AI fabrics.

The architecture addresses this using:

- Dynamic Load Balancing (DLB)
- Cluster Load Balancing (CLB)
- Congestion-aware routing
- Queue management

Testing of this reference architecture demonstrated substantially improved uplink utilization consistency when DLB and CLB were enabled (see below). More balanced traffic distribution helps prevent localized congestion and improves the efficiency of distributed AI communication patterns. As environments grow, these capabilities help maintain predictable infrastructure performance without requiring operators to continually redesign network architectures.

OPERATIONAL VISIBILITY AND TELEMETRY

Observability becomes increasingly important as AI clusters scale.

The architecture incorporates:

Arista CloudVision-as-a-Service (CVaaS) Telemetry

Arista telemetry tools provide centralized visibility into:

- Throughput
- ECN activity
- PFC activity
- Drop counters
- LANZ (latency analyzer) metrics

AI Agent

The Arista AI Agent extends visibility directly into attached hosts, enabling infrastructure teams to correlate network and compute behavior. This visibility becomes essential for identifying performance bottlenecks before cluster expansion.

Operational visibility is particularly important when scaling from pilot deployments to production environments. By correlating network telemetry with host-level behavior, operators can identify bottlenecks, validate performance, accelerate troubleshooting, and make more informed capacity planning decisions.

MULTI-TENANT DESIGN

The architecture supports shared infrastructure through logical segmentation and workload isolation.

Benefits include:

- Resource partitioning
- Predictable performance
- Infrastructure sharing
- Operational efficiency

This enables enterprises, service providers, and neocloud operators to support multiple workloads on common infrastructure.

VALIDATION

The reference architecture was validated using an Arista AI Center of Excellence (COE) testbed and complementary infrastructure validation environments.

The methodology focused on:

- Network performance
- GPU communication efficiency
- Congestion behavior
- Load balancing effectiveness
- Visibility and Reliability
- Power efficiency

TEST ENVIRONMENT

To assess the behavior of a modern Ethernet-based AI infrastructure under realistic AI training and inference traffic patterns, the validation system was built with the following elements.

Compute Infrastructure

- AMD Instinct GPUs
 - 32 GPUs across four servers
- Supermicro AI server platforms
 - AS -8126GS-TNMR
- AMD Pollara AI networking adapters
 - Attached at 400Gbps to the leaf switches using 800G to 2x400G breakout cables
- NIC Configuration
 - dscp 24 -> tc3 for data traffic
 - dscp 48 -> tc6 for rdma-ack ecn-cnp
 - pfc and dcqcn are also enabled

Network Infrastructure

- Arista Ethernet network architecture
 - 6 Arista DCS-7060X6-64PE
- Layer-3 routed fabric using BGP
- RoCEv2 transport
- ECMP forwarding
- Dynamic Load Balancing (DLB)
- Cluster Load Balancing (CLB)

Fabric Topology

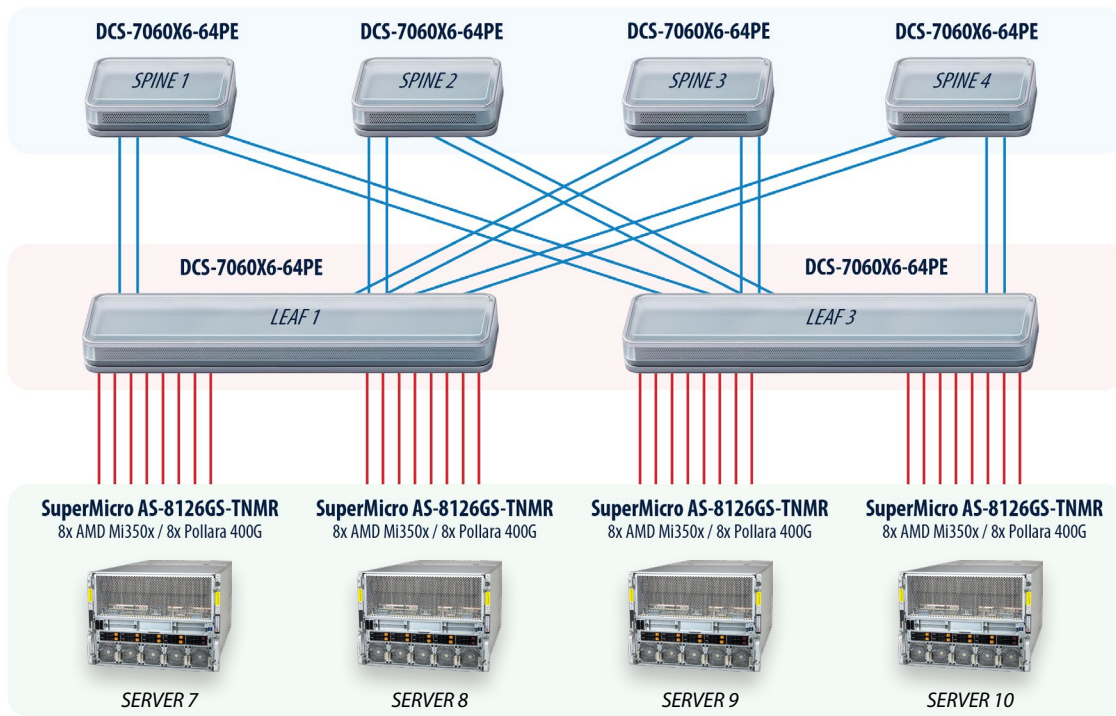


Figure 1: AMD InstinctMI350X GPU Cluster – Spine-Leaf Network Topology

The validation environment utilized a non-blocking spine-leaf topology representative of modern AI scale-out architectures. Multiple uplinks between the leaf and spine layers enabled load-balanced east-west traffic flows while supporting RoCEv2 communication between distributed GPU resources.

Oversubscription was kept at 1:1 with a total fabric bandwidth of 12.8T.

Optics and Connectivity

The objective was to compare latency, reliability, power consumption, and operational behavior under sustained AI workload conditions.

This architecture employed 800GBASE-2FR4 with 800G links for leaf-to-spine, A-O800-2Q400-5M with 800G to 2x400G breakout for the leafs to the servers.

Software Environment

GPU communication testing was performed using AMD's software ecosystem and collective communication libraries designed for distributed AI workloads.

The software stack included:

- ROCm platform
- ROCm Collective Communication Library (RCCL)
- RoCEv2 networking
- AI workload validation frameworks

Workloads

Validation focused on traffic patterns commonly observed in large-scale AI environments and fundamental to distributed model training and multi-node inference architectures. The workloads included:

Collective Communication Operations

- AllReduce
- AllGather

Traffic generation was designed to stress fabric utilization, congestion-management mechanisms, and load-balancing capabilities while measuring their impact on communication efficiency and overall infrastructure performance.

TECHNICAL VALIDATION RESULTS

RCCL Benchmark Results

To evaluate the impact of network configuration on distributed GPU communication, the reference architecture measured RCCL bus bandwidth using 16 GB message sizes across AllGather, AllReduce, and All-to-All collective operations in both two-node and four-node configurations. Testing compared a default network configuration using hash-based load balancing and default QoS against an optimized configuration using Cluster Load Balancing (CLB), Dynamic Load Balancing (DLB), and PFC/ECN congestion-management policies.

The benchmark measurements show that the optimized network configuration maintained AllGather and AllReduce bus bandwidth comparable to the default configuration while increasing measured All-to-All bus bandwidth.

Why It Matters

Distributed AI workloads rely heavily on collective communication among GPUs. These benchmark measurements show that the optimized network configuration improved measured All-to-All bus bandwidth while maintaining AllGather and AllReduce bus bandwidth comparable to the default configuration. Although the benchmark does not isolate the contribution of individual networking features, it shows the combined effect of CLB/DLB load balancing together with PFC/ECN congestion-management policies on RCCL.

Table: Collective Communication Bus Bandwidth (16 GB Messages)

RCCL Collective	Default	(CLB/DLB + PFC/ECN)
AllGather (2 nodes)	383.19 GB/s	382.04 GB/s
AllGather (4 nodes)	382.94 GB/s	383.48 GB/s
AllReduce (2 nodes)	382.89 GB/s	384.16 GB/s
AllReduce (4 nodes)	384.17 GB/s	383.84 GB/s
All-to-All (2 nodes)	19.81 GB/s	89.85 GB/s
All-to-All (4 nodes)	5.74 GB/s	59.12 GB/s

BUSINESS IMPACT

- **Faster Time to Value**
This validated reference architecture reduces implementation uncertainty by providing a tested blueprint that organizations can deploy with greater confidence.
- **Improved Investment Utilization**
Efficient GPU communication, congestion management, and balanced traffic distribution help maximize the value of accelerator investments by reducing idle time and communication delays.
- **Reduced Investment Risk**
Validation across compute, networking, optics, and software components helps reduce operational surprises during production deployment.
- **Greater Operational Visibility**
Integrated telemetry and visibility tools provide insight into both network and host behavior, helping operators manage increasingly complex AI environments.
- **Reduced Vendor Lock-In**
The architecture leverages open technologies including Ethernet, BGP, RoCEv2, and ROCm, providing flexibility compared to proprietary alternatives.

DEPLOYMENT SCENARIOS

- **Neoclouds:**
 - Rapid expansion of GPU capacity while maintaining operational consistency.
- **AI ISVs:**
 - Scalable infrastructure for hosted inference services.
- **Enterprise AI:**
 - Internal AI platforms supporting copilots, RAG systems, and domain-specific applications.
- **Large-Scale Training and HPC:**
 - Distributed workloads requiring efficient collective communication and predictable performance.

SCALING FROM PILOT TO PRODUCTION

One of the primary objectives of the validated architecture is to provide a repeatable path from initial AI deployments to large-scale production environments. Rather than treating each expansion phase as a separate infrastructure project, organizations can extend the architecture incrementally while maintaining consistent operational models, networking behavior, and management processes.

SCALABLE FABRIC DESIGN

Additional GPU servers can be introduced by adding leaf switches and compute resources while maintaining consistent latency and bandwidth characteristics. As clusters grow, additional spine capacity can be added to preserve non-blocking communication and support increasing east-west traffic demands.

Efficient Traffic Distribution

The architecture combines ECMP forwarding with Dynamic Load Balancing (DLB) and Cluster Load Balancing (CLB) to improve traffic distribution across available network paths. These capabilities become increasingly important as AI workloads scale and collective communication traffic grows.

Operational and Infrastructure Growth

The observability framework described earlier provides operators with the visibility required to monitor utilization, identify bottlenecks, and plan future expansion. Power and cooling requirements also increase with cluster size.

CONCLUSION

AI infrastructure has evolved into a capacity-building challenge that extends far beyond accelerator performance. Organizations must deploy scalable compute resources, high-performance networking, operational visibility, congestion management, and efficient power infrastructure while maintaining the flexibility to grow from pilot deployments to production-scale environments.

The validated architecture presented here combines Supermicro AI infrastructure, AMD Instinct GPUs, AMD Pollara networking technology, and Arista Ethernet fabric capabilities into a scalable AI platform designed for both training and inference. We demonstrate that open Ethernet architectures can provide the performance, reliability, visibility, and efficiency required to support modern AI environments while maintaining deployment flexibility and avoiding unnecessary vendor lock-in.

Most importantly, the architecture provides a repeatable path from pilot environments to production-scale AI capacity expansion. Rather than treating AI infrastructure as a collection of individual hardware components, it presents a validated operational blueprint that organizations can use to deploy, scale, and manage AI infrastructure with greater confidence and lower risk.

MORE INFORMATION

Supermicro: <https://www.supermicro.com/en/accelerators/amd>

AMD Instinct: <https://www.amd.com/en/products/accelerators/instinct.html>

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit: www.supermicro.com

AMD

More than 50 years driving innovation in high-performance computing. AMD Instinct accelerators power AI workloads for leading cloud providers, enterprises, and sovereign AI initiatives worldwide. The AMD commitment to open standards and open ecosystem ensures enterprises can deploy AI with control, flexibility, and long-term vendor independence. Visit www.amd.com

ARISTA

Arista Networks is an industry leader in data-driven, client to cloud networking for large data center/AI, campus and routing environments. Arista's award-winning platforms deliver availability, agility, automation, analytics and security through an advanced network operating stack. It is a worldwide leader in cloud and AI high-performance networking. Visit: Arista.com



©Super Micro Computer, Inc. Specifications subject to change without notice. All other brands and names are property of their respective owners. All logos, brand names, campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part, without express written permission by Supermicro Corporate Marketing.