Powering Next-Gen AI Clusters:
High-Performance Networking with AMD and Arista

Arista and AMD enjoy a deep partnership and have worked closely together for many years. When it comes to AI Clusters, we believe that it is the complete system that matters and that both network and compute are critical to delivering world-class performance and efficiency.

Similarly, both companies strongly believe in the benefits of an open ecosystem, where customers can select best-of-breed inter-operable components and combine them to create the solutions they require. The Ethernet ecosystem is a key enabler for this type of flexibility. The core of Ethernet is a set of interoperable standards within an ecosystem that is extremely large and highly competitive. For many decades now, that competition has driven rapid evolution and substantial innovation. When new core functionality is required from Ethernet, it is delivered through industry collaboration. Arista and AMD play leading roles in those collaborations. Both companies were founding members of the Ultra Ethernet Consortium (UEC) , initiators of the Open Compute Project (OCP) Ethernet for Scale-Up Networking (ESUN) workstream, and collaborators in the creation and validation of the OCP Multipath Reliable Connection (MRC) specification.

Arista and AMD have collaborated to deliver multiple large-scale production deployments across multiple generations of AMD GPUs. This white paper presents a number of production-validated designs developed through that collaboration.

Networking for AI Clusters

There are four distinct networks that are associated with AI Clusters: Front End, Scale Up, Scale Out, and Scale Across. This paper focuses on two of those.

Front End Network

All AI deployments have some form of Front End network. The Front End network connects all of the AI server CPUs together as well as providing access to a variety of other services. These services include high-capacity high-speed storage, non-AI compute servers, and access to external Cloud services and wide-area networking. Within an AI server, the Front End network attaches via dedicated CPU-attached NICs or DPUs.

This network is typically an extension to an existing network, as a Front End network is commonly deployed for other DC and Cloud use-cases. There are some differences in AI use-cases. It is typical to provision much larger per-server bandwidth (400G NICs are already common). The additional bandwidth is primarily to support high-bandwidth access to storage.

Scale Out AI Fabric

The Scale Out (or Back End) AI Fabric connects all the GPUs in a deployment together to form an AI/ML Cluster. AMD GPUs typically have dedicated AMD Pensando™ AI NICs and these NICs are interconnected to form the Scale Out network. This network is often the largest and the most complex to design. The performance of the Scale Out network is critical to the performance of large-scale AI workloads.

AMD Instinct™ MI3xx Series

AMD's MI3xx series includes several generations of GPUs that are available in multiple form factors. Additionally, AMD provides customers and OEMs with significant flexibility to configure AI servers in a variety of ways to meet their workload needs. The most common server configurations for enterprise and neocloud AI deployments leverage AMD's universal baseboard (UBB) module which provides eight MI3xx GPUs connected locally through an Infinity Fabric Scale Up network. While there are significant generational differences between the GPUs in this series, the external networking configuration is common. These servers typically provide eight AMD Pensando™ Pollara 400G AI NICs for Scale Out, and one or two Front End NICs (either additional AMD Pensando Pollara 400 AI NICs or AMD Pensando Salina 400 DPUs). Broadcom Thor 2 NICs are also an option from many AMD OEMs and we have deployed the designs that follow with both NIC options.

Validated Network Architectures

Arista has completed many deployments of AMD GPUs in collaboration with AMD and our customers and partners. These range in size from small enterprise through to large-scale neocloud and hyperscale. Our view is that there is no better qualification or validation process than a successful production deployment. One early public example was our work with AMD and Core42 to deploy the Maximus-01 cluster. Maximus-01 features more than 9,000 AMD MI300X GPUs and secured the 20th spot in the TOP500 list of the world's fastest supercomputers[1].

This section describes some common design patterns that we have validated in production. But given we have hundreds of AI customers with widely varying requirements, Arista typically works directly with customers to build design recommendations that are customized to their needs.

[1] *https://aicloud.core42.ai/news/core42-maximus-01-cluster-secures-top-20-ranking-on-the-global-top500-supercomputers-list*

Front End

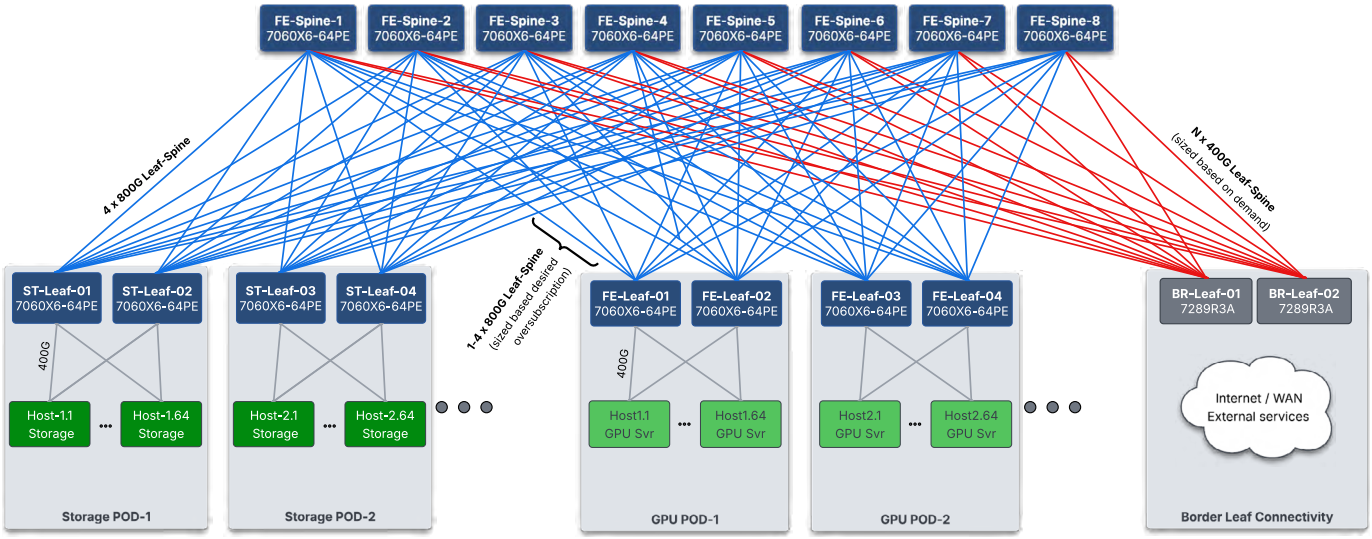

Figure 1: Typical validated medium-scale Front End network

As mentioned earlier, Front End networks in AI clusters are an extension and continuation of existing enterprise and service provider network designs. There are two common differences compared to existing leaf-spine datacenter designs. The first is that there is typically a high-speed storage subnetwork, which is typically provisioned 1:1 (without oversubscription) in order to meet the needs of AI workloads. The second is that the AI servers themselves often have 400G or Nx400G connectivity with dedicated leaf pairs provided to meet that need. The links to the AI servers may be modestly over-subscribed depending on the precise configuration and expected workload. Figure 1 above shows a typical validated Front End design for a multi-thousand GPU AMD MI3xx deployment. Storage connectivity is provisioned 1:1, so there is no oversubscription towards storage. Oversubscription for the GPU server connectivity can be adjusted to meet the needs of the workload. Internet/WAN connectivity is customized to the specific deployment requirements.

Scale Out

In an MI3xx deployment, the Scale Out network provides the connectivity between the GPUs. Each GPU is typically connected to a single 400G NIC, so the network must support 400G per GPU.

Single Tier Architecture with AI Spine

For deployments that are not planned to grow beyond 144 AI Servers (1,152 MI3xx GPUs), a single-tier network can be a very simple solution. With the 7800R4 AI Spine Switch, a single modular, internally redundant AI Spine can incrementally grow from a few GPUs up to over 1K GPUs. The 7800R4 provides high-scale, incremental growth, lossless transport through hierarchical, hybrid packet buffering with Virtual Output Queues (VOQs), and an overprovisioned cellular internal fabric that provides perfect load balancing and 100% bandwidth. This configuration is production validated with multiple customers.

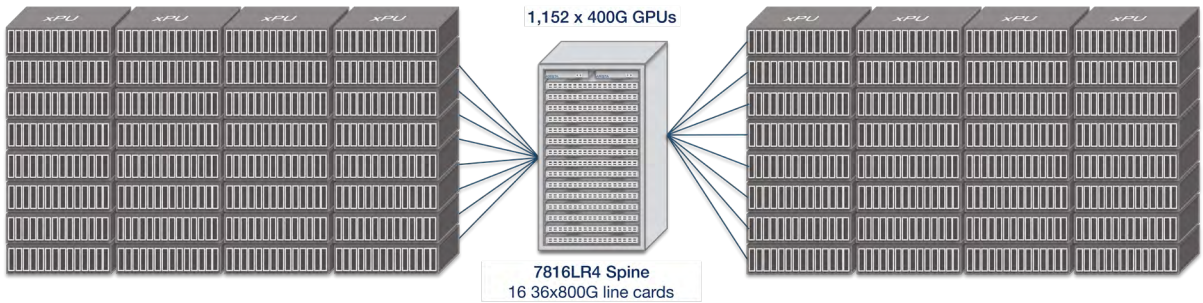

Figure 2: Single tier network supporting up to 1,152 MI3xx GPUs

Two Tier Architectures

For larger deployments, Arista has validated multiple two-tier designs in production with AMD GPUs. These designs are typically scalable and designed around PODs of GPU capacity. These PODs are typically 64 servers/512 GPUs. Depending on the needs of the customer, these PODs can be configured as Rail-aligned or as Rail-unaligned, sometimes known as "Tree". Rail-aligned is more common and, in this case, the NICs in each server fan out across eight leaf switches. Tree is also supported for customers who prefer a top-of-rack configuration for their DC build. Figure 3 below shows both types of POD.

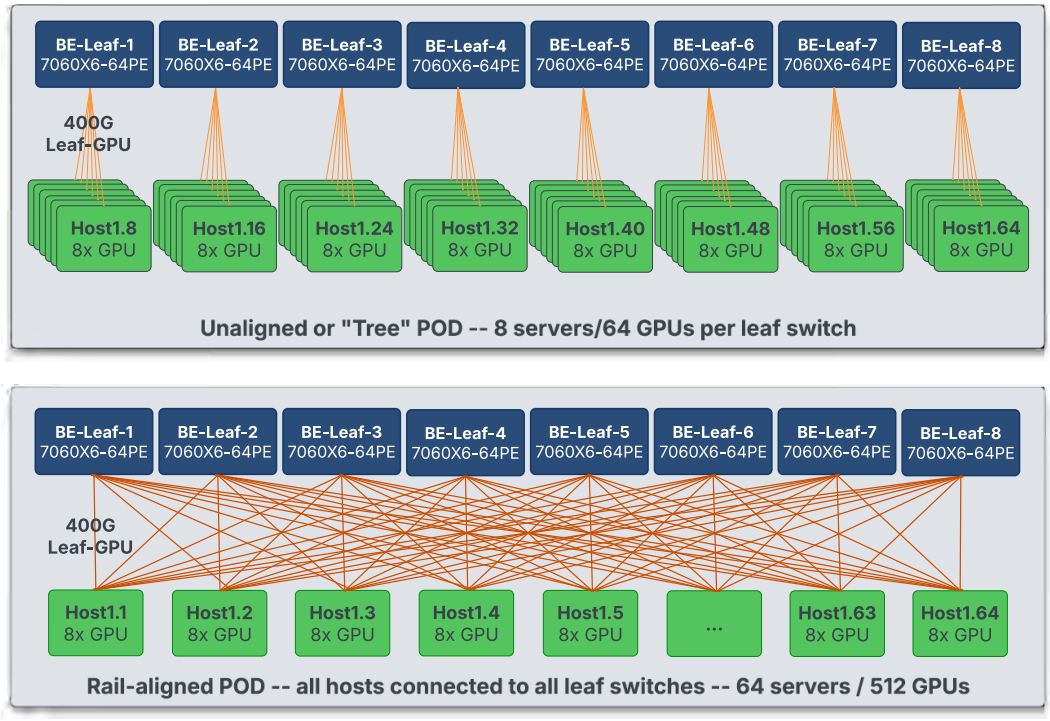

Figure 3: Common POD configurations: Unaligned/Tree and Rail-aligned.

These are leaf-spine architectures, typically provisioned at 1:1, or overprovisioned to provide headroom. The leaf devices are Arista 7060X6 AI Leaf switches. The spine devices utilised depend on the scale that is required as well as the degree of incremental scaling that is required. For deployments that do not exceed 8K GPUs, the 7060X6 may be used as the spine device also. Figure 4 below shows an example medium-size 7060X6 two-tier design that supports eight PODs (4,096 GPUs) and can be scaled up to support sixteen PODs (8,192 GPUs) by moving to a wider spine.

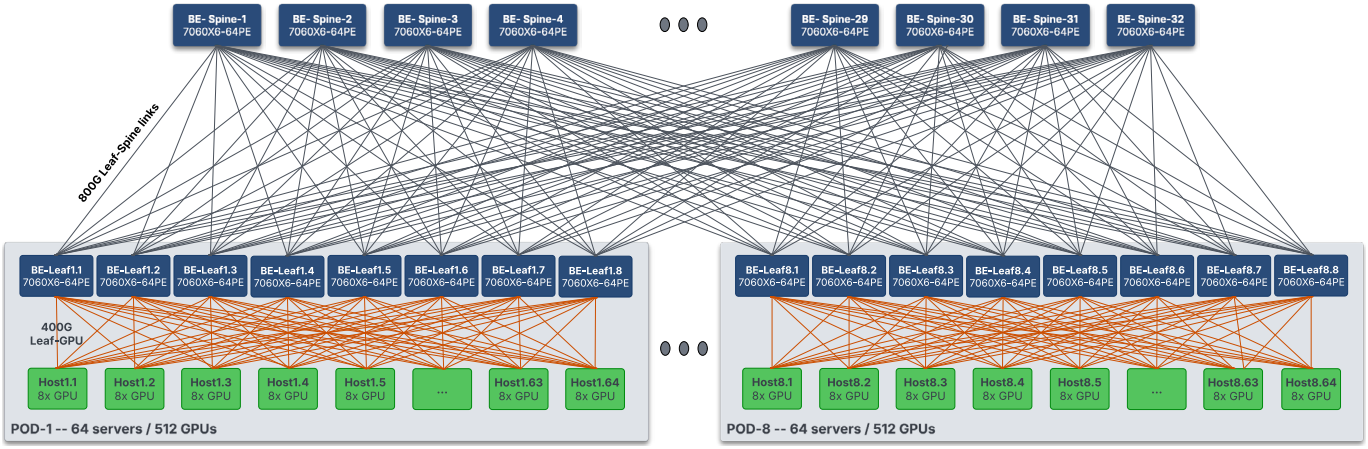

Figure 4: Two-tier with 16-way 7060X6 spine. Supports up to 4,096 GPUs. Shown with Rail-aligned PODs.

For larger deployments, the 7800R4 AI Spine is used and can support deployments up to 73K GPUs. Figure 5 below shows an example that uses a 16-wide 7812R4 spine, which can scale up to 27 PODs (13,824 MI3xx GPUs). Finally, Figure 6 below shows a smaller design that utilizes a 16-wide 7060X6 spine to support four PODs (2,048 GPUs) in an unaligned/tree configuration. All of these design approaches have been delivered and validated in production deployments.

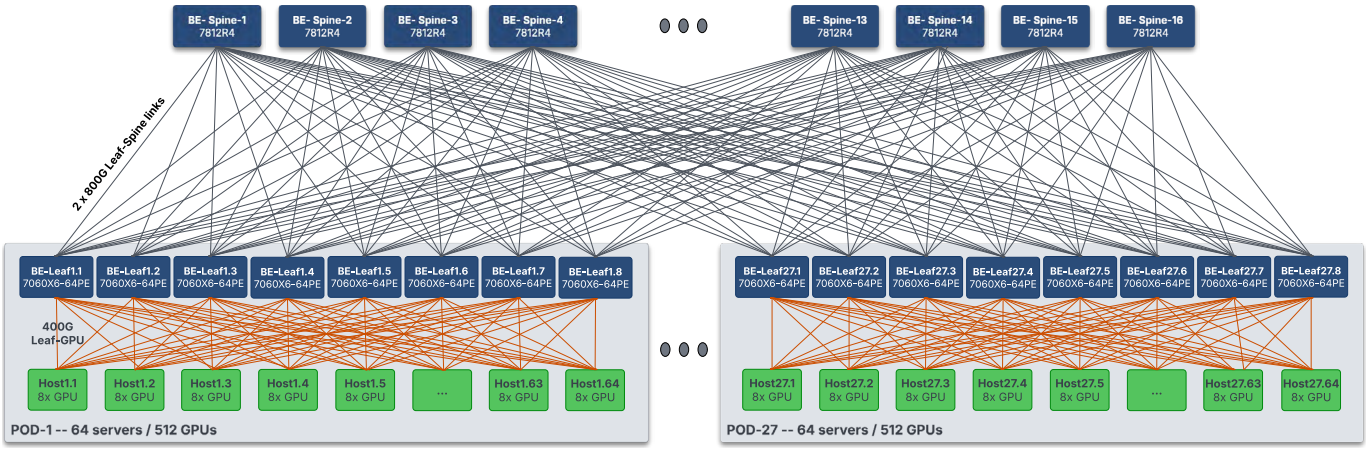

Figure 5: Two-tier with 16-way 7812R4 spine. Supports up to 13,824 GPUs. Shown with Rail-aligned PODs.

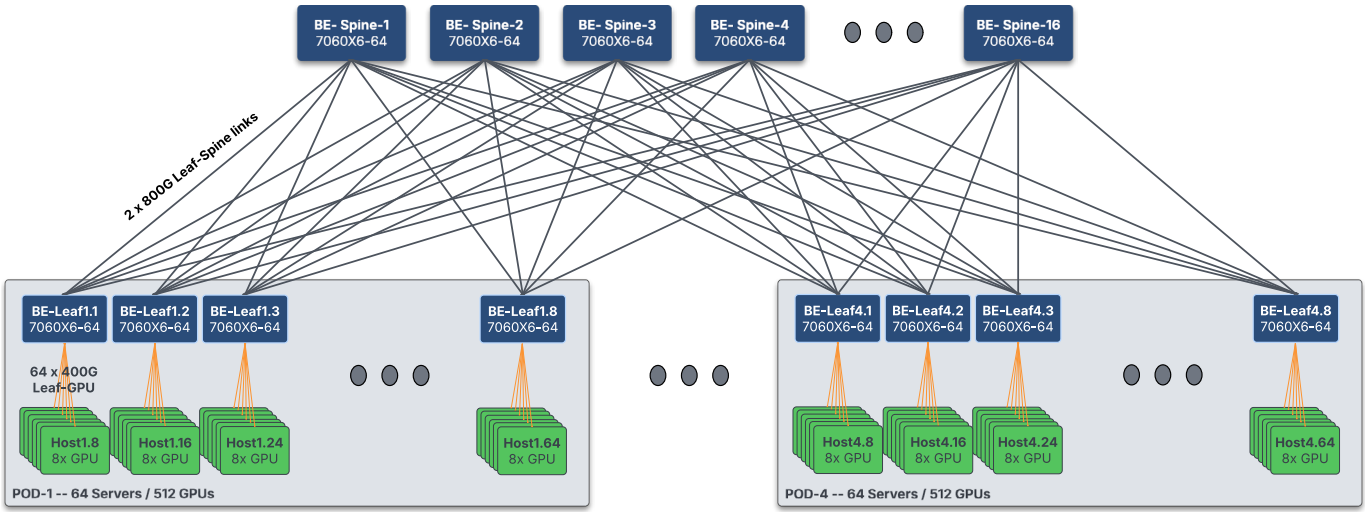

Figure 6: Two-tier with 16-way 7060X6 spine. Supports up to 2,048 GPUs. Shown with Rail-unaligned PODs.

AMD Helios Rack-Scale Solution

AMD Helios is a next-generation rack-scale AI System. Rather than individual servers, Helios leverages the OCP Open Rack Wide (ORW) standard to deliver an integrated liquid-cooled rack-scale solution (Figure 7 below). Helios pairs 5th-generation Instinct MI455X GPUs with AMD Pensando™ Vulcano 800G AI NICs within the rack-scale architecture. Each Helios rack contains 18 compute trays. Each compute tray is configured with four MI455X GPUs and up to three AMD Pensando™ Vulcano 800G AI NICs per GPU for Scale Out, complemented by a 6th Gen EPYC Venice CPU and AMD Pensando Salina 400 DPU to create a complete system. Six in-rack switch trays provide a local Scale Up network that interconnects all 72 GPUs. The Scale Up network is built on an open networking approach, leveraging UALink over Ethernet (UALoE) to deliver up to 260TB/s of bandwidth.

Figure 7: AMD Helios rack-scale system and AMD Pensando™ Vulcano 800G AI NIC board

Multi-Planar Scale Out

Multi-planar designs

A Plane is an important concept in Scale Out architectures. A Plane is a parallel, completely independent network that connects to all GPUs. In a generalized multi-plane design, parallel independent planes are built and each GPU connects to all of the planes in parallel. The connectivity from each GPU may be achieved through multiple NICs or by attaching a single NIC to multiple planes, or both. In all cases, the planes are identical. There is no connectivity between the planes, so the NIC or host must manage both load balancing between planes and detection and handling of connectivity failures within planes. An example dual-plane network is illustrated in Figure 8 below.

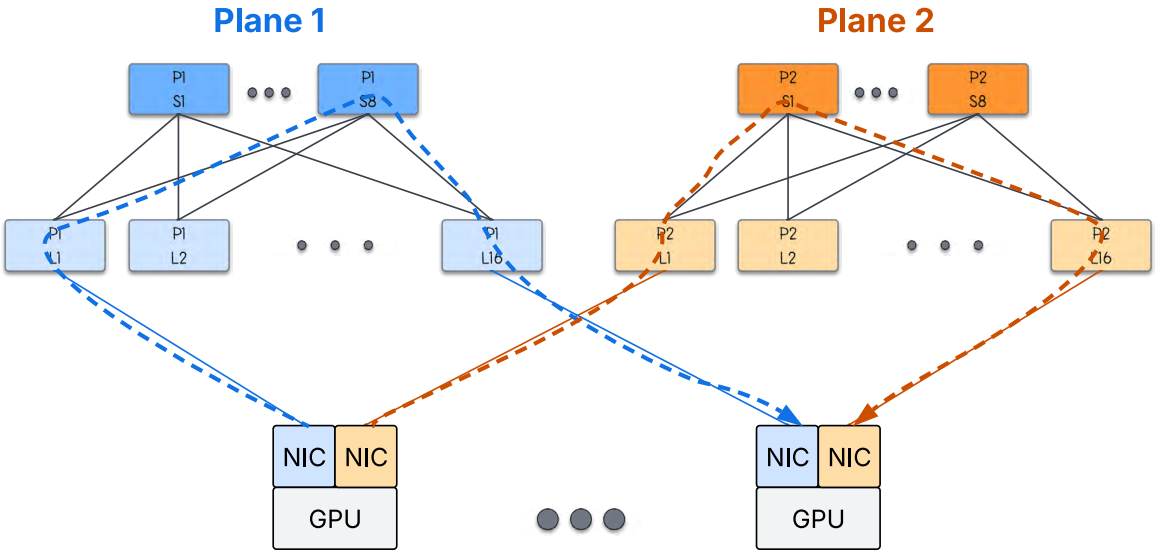

Figure 8: Simple dual-plane network with separate NIC per plane

Multi-plane solutions offer an immediate benefit in redundancy, as the system can survive many types of failure and the individual network planes are completely isolated from each other. Additionally, there are significant cost and scale benefits, because single-tier networks can support more clients and because the maximum scale of a two-tier network increases with the square of the radix of the network devices used. This is particularly relevant when an individual NIC is divided into multiple planes as illustrated in Figure 9 below. The reduction in port speed per-plane enables much larger scales in a multi-planar design. As an example, a 64 x 1.6Tbps switch such as the 7060XE7 can support up to 8,192 x 800 Gbps clients in a two-tier architecture. But the maximum scale increases rapidly when the port speed is reduced through a multi-plane fanout. The same switch can support 32,768 x 400G or 131,072 x 200G in a two-tier configuration. (And even larger sizes are possible with modular AI Spines.)

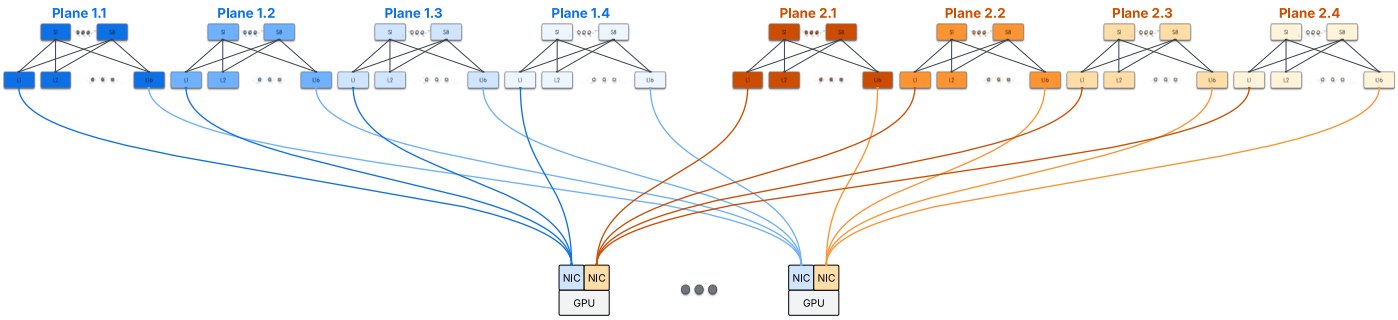

Figure 9: Multi-planar network with multiple NICs and multiple planes per NIC

AMD Helios is a fully multi-planar system[2]. Each MI455X GPU is attached to multiple AMD Pensando™ Vulcano 800 AI NICs, making the base design multi-planar by default. Additionally, each AMD Pensando Vulcano 800 AI NIC supports multi-plane breakout and can attach to one (800G), two (400G), or four (200G) planes. As Helios provides multiple NICs per GPU and those NICs also support multi-plane breakout, both multi-NIC planes and NIC-breakout planes can be implemented simultaneously. Load balancing and traffic management between NICs is implemented in AMD's ROCm system software. Load balancing and traffic management between planes on a single NIC is implemented in the Vulcano NIC as part of the network transport protocol: we discuss this in the next section.

Multipath transport protocols

Standard RoCEv2 is a single-path transport protocol. This means that communications related to each RDMA Queue Pair appear on the network as a single UDP flow in each direction. RoCEv2 does not natively support breaking up this flow to load balance across multiple paths within a single plane, let alone load balancing across multiple planes attached to a single NIC.

Within the open Ethernet ecosystem that both Arista and AMD are committed to, two solutions have emerged that support both multi-path and multi-planar operation for AI workloads. Both solutions are designed for a lossy network, removing the need for PFC. And they can adaptively load balance and spray traffic across many paths through the network, removing the challenges associated with load balancing large flows within the network in addition to providing a mechanism to distribute traffic across multiple planes.

The first solution is ***Multipath Reliable Connection (MRC)***. OpenAI led a multi-vendor consortium (including contributions from AMD and Arista). The consortium published an open OCP specification for MRC along with a paper describing their experiences with using MRC to train frontier AI models. MRC extends RoCEv2 to add multipath support to a limited subset of RDMA operations used by AI Training workloads. MRC has already been proven for deployments based on AMD AI NICs and Arista switches. Initial Helios deployments will leverage MRC. For more information, please refer to our white paper: Multiplanar Scale Out fabrics with MRC.

[2]*Note that multi-planar deployments of MI3xx are also possible. Please contact us if you want to learn more.*

The second solution is ***Ultra Ethernet (UE)***. Both AMD and Arista were founding members of the Ultra Ethernet Consortium (UEC), which is a standards organization that was formed with the goal of enhancing Ethernet for the needs of AI and HPC while maintaining the interoperability that has made Ethernet successful. In 2025, the consortium issued an initial specification and whitepaper. The latest specification is V1.0.3 (July 2026). Ultra Ethernet Transport (UET) forms part of the specification and provides an alternative RDMA transport protocol intended to replace RoCE for all workloads and address many shortcomings in existing implementations. This makes Ultra Ethernet a more comprehensive solution for a wide range of workloads in addition to the workloads that are supported by MRC. MRC reuses key elements of Ultra Ethernet, including packet trimming, selective acknowledgments, and congestion control. For more information, please refer to our white paper: Demystifying Ultra Ethernet.

Validated Network Architectures

Helios is not yet deployed in production. However, AMD and Arista are collaborating closely in AMD's labs to bring up and test Helios networking in combination with Arista switches. This will ensure that validated designs are available prior to customer launch.

For the Front End network, Helios will typically be configured with the Salina DPU and both the designs and the lessons we have learned from multiple production deployments of MI3xx will apply.

For the Scale Out network, the base configuration pairs 2 x AMD Pensando™ Vulcano 800 AI NICs with each MI455X GPU. There is a natural fit between AMD Pensando Vulcano 800 AI NIC and Arista's 7060XE7 portfolio of 1.6T switches. Both platforms support the latest 200Gbps per lane optical standards, avoiding the need for gearboxes and making the most efficient use of fiber infrastructure. Additionally, 7060XE7 is available in both air- and liquid-cooled variants, allowing customers to deploy 100% liquid-cooled infrastructure if they wish.

We are jointly focused on validating two primary network architectures for Helios. For deployments up to 96 racks (6,912 GPUs), our reference architecture will be dual-plane (one per NIC) and support rail-aligned PODs containing three Helios racks (216 GPUs). This architecture is depicted in Figure 10 below.

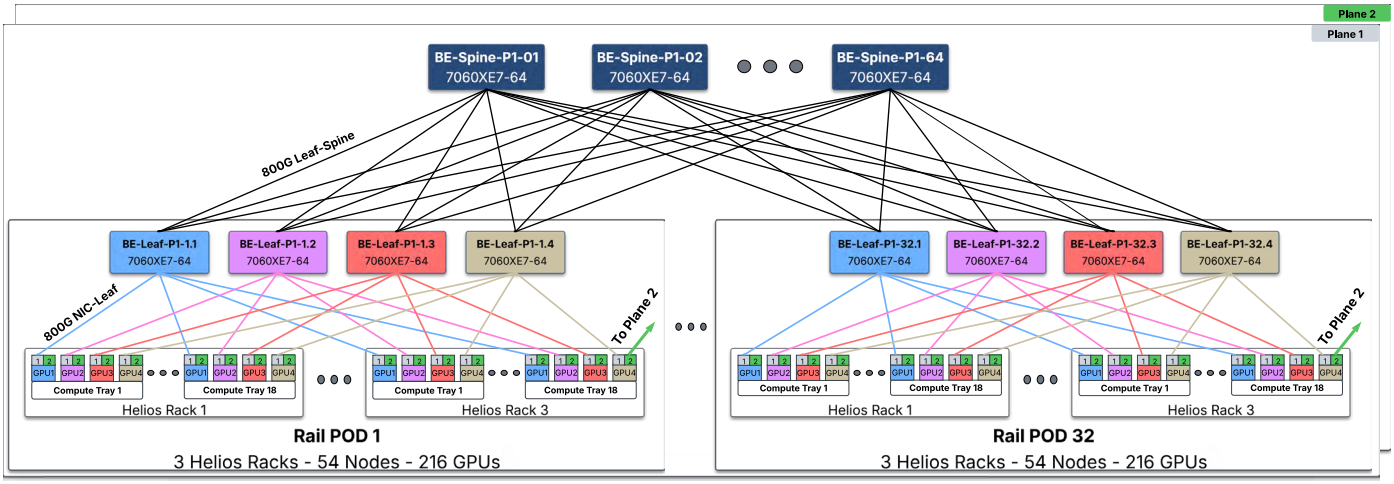

Figure 10: Helios dual-plane 7060XE7 design, for deployments up to 6,912 GPUs

The dual-plane architecture can be expanded to support up to 432 Helios racks (31,104 GPUs) by deploying the 7800R4 AI Spine in place of the 7060XE7 in the spine role (depicted below in Figure 11). The use of a 7800R4 AI Spine is also beneficial for customers who are deploying capacity incrementally, since the 7800 is a modular chassis. And the additional scale enabled by a 7800R4 spine enables larger deployments without the additional cabling and operational complexity associated with multi-plane breakout of individual NICs, as is required by the eight plane design that we discuss next.

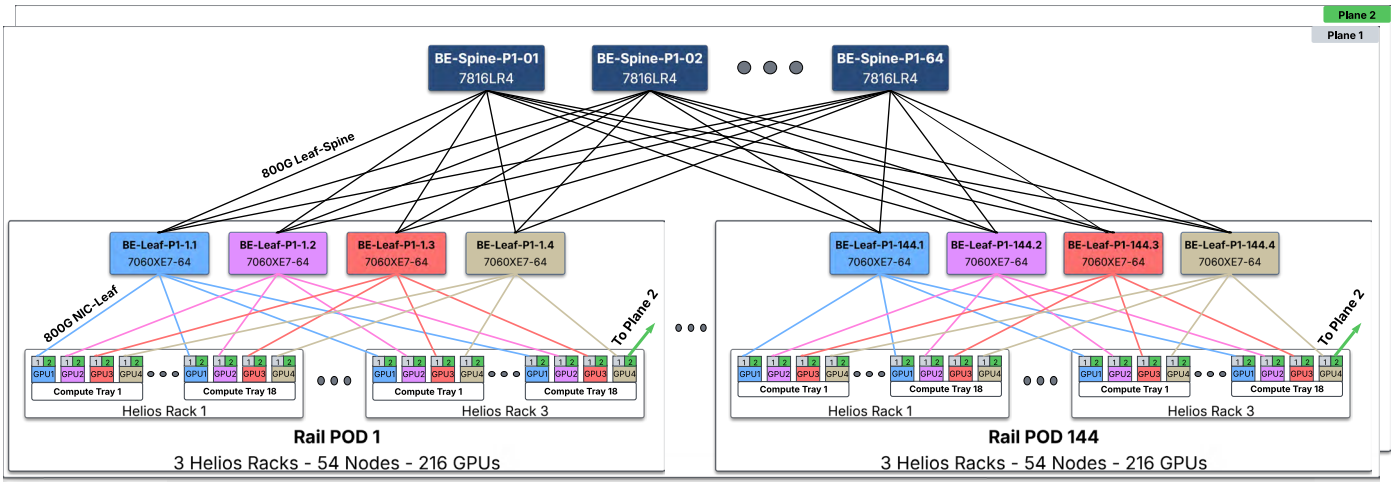

Figure 11: Helios dual-plane design with 7060XE7 leaf and 7800R4 spine, for deployments up to 31,104 GPUs.

For the largest deployments, the second reference architecture will leverage the multi-plane capabilities of the AMD Pensando™ Vulcano 800 AI NIC to provision four planes of 200G on each NIC, resulting in eight planes in total (as illustrated conceptually in Figure 9 above). This architecture is based on PODs of fourteen Helios racks (1,008 GPUs) and can scale smoothly up to 128 PODs (129,024 GPUs) as shown in Figure 12 below. If required, this design can be augmented with the 7800R4 AI Spine to support even larger scales (up to 580,608 GPUs).

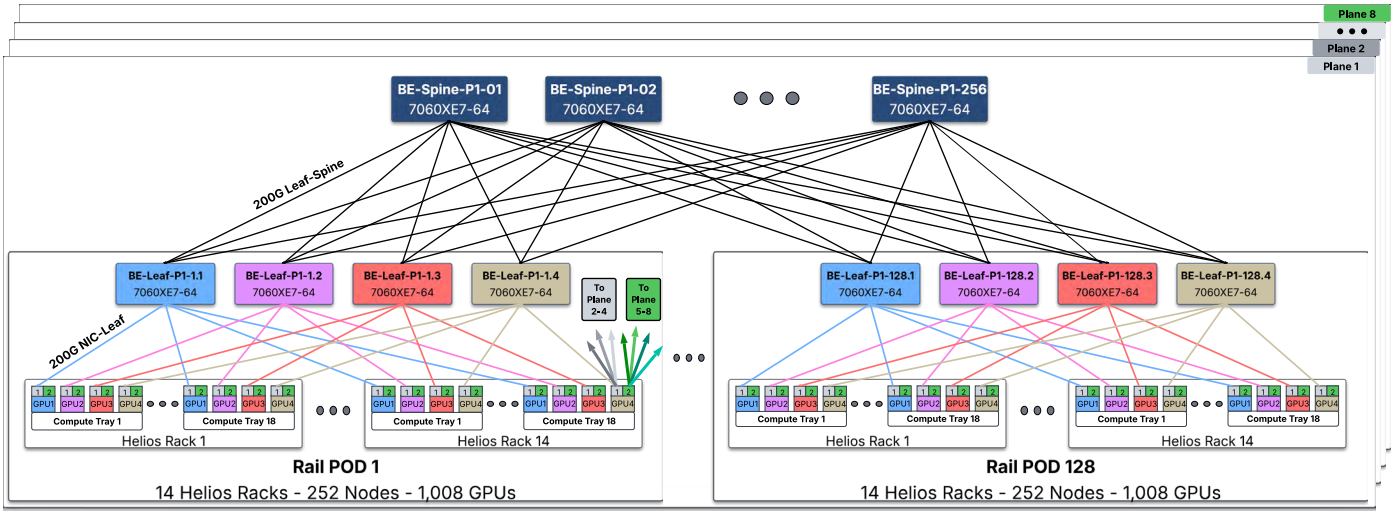

Figure 12: Helios eight-plane 7060XE7 design for deployments up to 129,024 GPUs

Both configurations will leverage MRC as an enhanced RDMA transport. For the dual-plane design, we will validate MRC's multipath capabilities within each plane along with ROCm support for multiple NICs. For the eight-plane design, we will additionally validate MRC's multi-planar support. In parallel to the joint work with AMD, Arista is also validating MRC performance on AMD Pensando™ Pollara 400 AI NICs in conjunction with MI3xx GPUs in our AI Center of Excellence, which we will discuss below.

Building AI networks with Arista

Arista is a leader in AI Networking, across all types of AI Network. This section briefly presents Arista's solutions and the benefits of building and operating AI networks with Arista. For more depth, refer to https://www.arista.com/en/solutions/ai-networking and our AI Networking white paper.

Etherlink AI Portfolio

Arista's Etherlink portfolio offers high-performance standards-based Ethernet systems with smart features that mesh perfectly with AMD-based AI Clusters. Smart AI-Aware features include a portfolio of specialist load balancing and congestion control features to ensure reliable packet delivery and protect AI performance. Etherlink features are supported across a broad range of 800G and 1.6T systems. Etherlink is compatible with both Ultra Ethernet Transport (UET) and Multipath Reliable Connection (MRC).

Arista offers fixed, modular, and distributed network platforms for AI, scaling from small clusters to topologies supporting over 100,000 accelerators. The 7060X series provides 400G, 800G, and 1.6T systems up to 102.4Tbps. The 7800R4 series delivers up to 460Tbps with 576x800G or 1152x400G ports, enabling large, efficient single-switch clusters. The 7800R4 may also be deployed in an AI Spine role to deliver extremely large two-tier networks. Additionally, the 7800 supports internet-scale routing and the key encryption, traffic segmentation and traffic engineering features that are required for Front End and Scale Across networks.

EOS as a foundation for AI

Arista Extensible Operating System (EOS) serves as the foundation for next-generation data centers and high-performance AI networks, offering industry-leading reliability and the lossless environment essential for accelerated computing. Built on a unique multi-process state-sharing architecture, EOS separates network state from protocol processing to ensure maximum uptime and system stability. It offers extensive programmability through a rich set of APIs and SDKs. Key features include real-time telemetry for network-wide visibility, zero-touch provisioning, and support for containerized or virtualized deployments, all aimed at reducing operational costs and increasing business agility.

By utilizing a single binary software image across all Arista platforms—including AI-optimized hardware like the 7800R4, 7060X6, and 7060XE7 series—EOS ensures operational consistency and simplifies the deployment of complex AI fabrics. Its programmable nature and support for advanced security, segmentation, and policy capabilities allow organizations to automate workflows and scale their AI compute resources while maintaining rigorous performance and safety standards.

AI Operations

Arista's platforms with EOS provide extensive telemetry capabilities for AI networking. These capabilities combine with Arista CloudVision for AI Networks to create a data-driven networking architecture where EOS provides the telemetry and CloudVision provides the intelligence and centralized management. CloudVision for AI Networks adds workload job visibility and analysis to the normal network-oriented views. Server CPU and NIC metrics can be integrated through Arista AI Agent, providing deep visibility into AI workload operation to improve cluster deployment time and operational stability. And Arista's Autonomous Virtual Assistant (AVA) provides AI-driven support to operators.

These solutions can be flexibly combined with two open-source automation frameworks to deliver a complete solution for end-to-end provisioning, deployment, and operations. Arista AVD (Architect Validate Deploy) provides a collection of tools and best practices for deploying and managing Arista networks using infrastructure-as-code principles. Complementing AVD, the Arista Network Test Automation (ANTA) framework allows on-demand validation of the network state against intended configurations, ensuring the network operates reliably and as expected. Arista solutions provide open APIs for customers who wish to integrate their own systems.

Smart System Upgrade (SSU)

Arista Smart System Upgrades are designed to address one the most challenging operational tasks in AI fabrics: infrastructure maintenance and device upgrades. Whether to deploy new features, install new hardware, or to mitigate security vulnerabilities, maintenance that causes network downtime can be hugely impactful to job completion times and overall efficiency, impacting the ROI of AI clusters. SSU is the solution to that problem, building on the robust multi-process state-sharing architecture of EOS. Combining protocol-based graceful shutdown, insertion, and unique redirection techniques, traffic can be diverted around the devices under maintenance. And in the critical leaf layers facing the XPUs, packet forwarding simply continues during device upgrades with workloads unaffected. Operations teams and network administrators now have the ability to perform seamless maintenance to any infrastructure element. Through CloudVision, organizations can fully automate these upgrade operations. To learn more, refer to our SSU White Paper.

Maintenance Mode

Maintenance Mode is a framework that allows for easy maintenance of a switch or specific elements of a switch. Arista's Maintenance Mode provides a set of commands with a wide range of flexibility that make our network operations lives a bit simpler. With Maintenance Mode we support the removal and reintroduction of either a whole switch, portions of the switch or even specific BGP neighbors in a graceful operation that minimizes network disruption. The feature leverages industry standards including Graceful BGP Session Shutdown (RFC 8326). Please refer to the Arista TOI and Community Central article for further details.

Arista's AI Center of Excellence

In parallel with our joint work with AMD in their labs, Arista has also collaborated with AMD, Supermicro, and others to build an AI Center of Excellence. This lab supports our work in validating solutions prior to production deployments as well as proving the efficiencies and differentiators of the Arista + AMD solution.

It is deeply valuable for Arista engineers to be able to directly work in the lab, test solutions, and obtain the proof that comes from seeing traffic on the wire. Arista's COE lab recently conducted 800GbE testing with the 7060X6 and AMD's MI350 accelerators combined with the Pollara AI NIC. Figure 13 below shows the lab configuration and Table 1 below shows real-world performance benchmarking that demonstrates Arista's fabric achieving unrivaled performance across a variety of cluster traffic patterns.

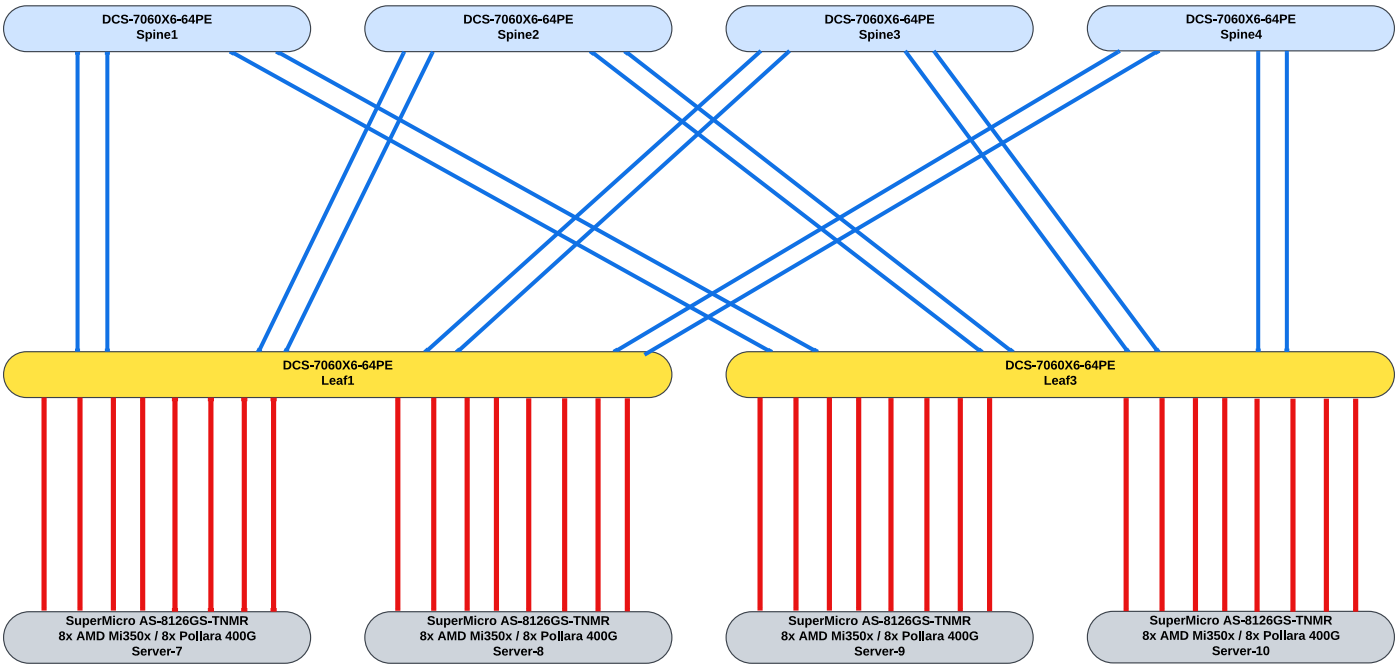

Figure 13: Recent AI COE testbed focused on testing MI350X/Pollara with an 800GbE 7060X6 fabric

RCCL Collective	8×800G Default Configuration	8×800G (CLB/DLB + PFC/ECN Tuning)
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

Table 1: Performance results for a variety of benchmarks: MI350X, Pollara, 800GbE 7060X6 leaf-spine

Conclusion

The combination of AMD GPUs, EPYC CPUs, and AMD AI NICs with Arista-powered AI Fabrics provides customers with world-class performance based on best-of-breed solutions all within an open ecosystem. The deep partnership between AMD and Arista across several generations of technology means that customers can deploy validated designs that have been repeatedly tested in production.

Arista and AMD continue to jointly push the boundaries of what is possible within the open Ethernet ecosystem. Helios deployments will demonstrate the power and performance of new open solutions powered by industry collaboration and led by AMD and Arista.

Arista's Etherlink AI portfolio delivers a holistic solution for optimized AI networking, combining AI-optimized hardware systems, industry-leading reliability with EOS, and a comprehensive AI Operations platform to build and manage AI Centers at every scale.

References

- AI Network Solution Guide
- AI Networking white paper
- Three Genius Ideas for AI Fabrics
- Multiplanar Scale Out fabrics with MRC

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office 1390
Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062