

Deploying Highly Available AI Fabrics with Arista EOS

Modern AI economics depend heavily on network resilience. As AI clusters expand, hardware fragility accelerates: [Meta Research](#) shows Mean Time Between Failures (MTBF) drops from 7.9 hours at 1,024 XPU to just 1.8 hours at 16,384 XPU. When projected to ultra-scale fabrics, physical link failures compound this problem even further: [SemiAnalysis](#) highlights that at a 100,000 XPU scale, even a conservative 5-year link MTBF can result in a training job failure every 26 minutes. AI infrastructure demands a highly available network fabric that seamlessly orchestrates traffic to eliminate bottlenecks, isolate localized faults, and unleash the full computational power of distributed XPU systems.

At scale, high-speed AI network fabrics experience disruptions across three primary vectors:

- **Software:** Operating system faults, control plane crashes, and routing process lockups.
- **Hardware:** Physical component failures (e.g., optical transceivers, cables) and link or port outages.
- **Maintenance:** Planned operations (software upgrades, hardware refreshes) and unplanned interventions (security patch deployments, emergency hardware swaps).

AI workloads depend on tightly synchronized collective operations across thousands of XPU. Consequently, even a sub-second network disruption can trigger cluster-wide training rollbacks, spike inference latency, and cause severe revenue or resource loss. Building a highly available AI fabric requires isolating the failure vectors at their source, preventing transient localized faults from cascading into fabric-wide compute stalls.

This paper outlines how Arista EOS delivers comprehensive network resilience to eliminate training job restarts, safeguard real-time inference, and maximize AI compute infrastructure ROI. By utilizing a single EOS image across Front End, Scale Out, Scale Up, and Scale Across fabrics, organizations radically simplify lifecycle operations, automation, and observability to achieve absolute high availability as the ultimate operational outcome.

Resiliency with Self-Healing Architecture of Arista EOS

Network software faults disrupt AI workloads differently depending on the execution phase. In distributed training, a single fault breaks synchronized collective operations, stalling all XPU until watchdog timers expire and force a costly checkpoint rollback. During real-time inference, faults cause localized packet loss and path disruption that lead to elevated tail latency and failed user requests. A resilient, highly available network operating system is therefore critical to maximize AI infrastructure productivity.

At the foundation of Arista high availability is EOS, a modular, state sharing network operating system built on a hardened Linux kernel. Traditional operating systems bundle routing protocols, hardware management, and monitoring utilities into monolithic architectures or tightly coupled threads, meaning a single software fault can crash the entire switch. Arista EOS is built on a resilient, scalable, and fault-tolerant architectural foundation using a central state database, as shown in Figure 1.

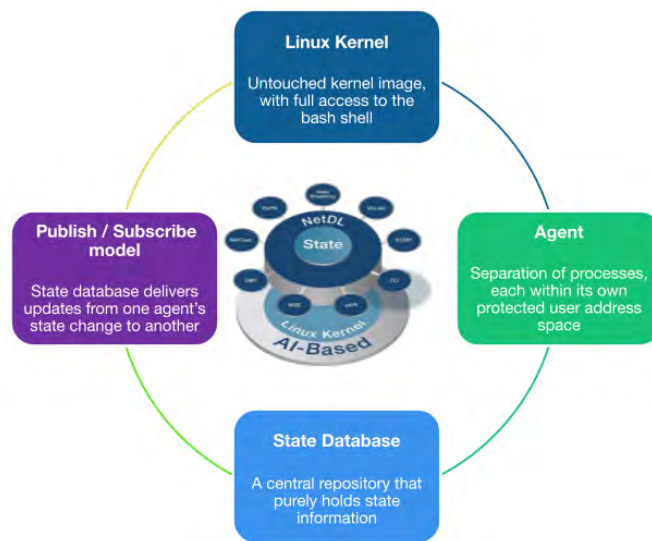


Figure 1: EOS architected for resiliency

The key architectural attributes of EOS, along with their operational benefits and direct impact on AI workloads, are summarized below:

Attribute	Benefit	Impact to AI fabric
Publish- Subscribe model	In-memory publish-subscribe state database that decouples system state from individual network processes, allowing for seamless fault recovery, rapid convergence, and live software patching.	The instant state-preserved convergence ensures AI jobs continue without interruption.
Control plane and data plane separation	EOS provides inherent separation between the control plane (software management) and the data plane (hardware forwarding) to help ensure that failures in the control plane do not affect traffic on the data plane.	Ensures hardware continues forwarding AI traffic at full wire speed even if control plane software experiences faults or restarts.
Agent isolation	Software agents within each plane run in complete isolation inside the network operating system, confining protocol issues to an individual process without spreading across the system.	The agent separation prevents localized software faults from cascading across the broader fabric while enabling targeted, non-disruptive bug fixes.
Memory and Message safety	Arista EOS uses a framework that enforces best practices for memory management, message passing, and other parts of the code that have historically been sources of vulnerabilities.	This design proactively eliminates memory corruption and security vulnerabilities, drastically reducing emergency patching cycles and unplanned fabric downtime.
Process Restartability	Decoupling system state from software processes allows individual agents to fail and restart independently without disrupting overall operations.	Failed agents recover quickly by resynchronizing directly from the state database, eliminating switch reboots.
Watchdogs	Hardware-assisted watchdogs maintain a continuous heartbeat with software agents, automatically executing a controlled reset if a fatal lockup occurs.	This real-time deadlock detection and surgical agent recovery prevent system freezes from cascading across the network and stalling multi-node training jobs.
Process updates	EOS agents can be patched live and restarted without disrupting switch operations or application data flow, ensuring zero user-perceptible downtime.	This enables live in-service software patching of individual OS components without interrupting data plane packet flow, eliminating scheduled cluster downtime.

Process restartability, watchdogs, and updates rapidly restore the EOS Agents by resynchronizing directly from the state database as shown in Figure 2. Because the underlying state remains preserved within the State Database (DB), the switching ASIC continues forwarding traffic at line rate during the sub-second process restart, rendering control plane recovery completely hitless to the data plane. This means the AI training and inference jobs communicate without any disruption.

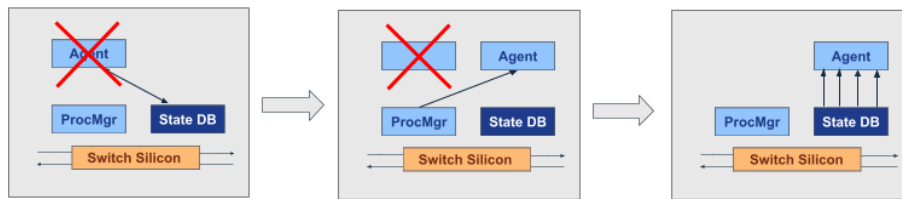


Figure 2: EOS Fault Containment and Self-Healing

Resiliency Against Hardware Failures

Unmitigated physical faults like hardware failures, optics signal degradation, and bit errors induce packet loss and network timeouts across AI fabrics. In AI training, this triggers tail-latency jitter and synchronization stalls that crash jobs and force costly checkpoint rollbacks; in inference, it causes latency spikes that degrade Time-to-First-Token performance and delay token generation.

To safeguard AI fabrics against hardware failures, signal degradation, and environmental risks, EOS employs proactive system and network level mechanisms. At the system level, EOS provides resiliency against memory failures protecting AI workloads from downtime. At the network level, EOS enables resilience to link failures through load balancing and path redundancy using Equal-Cost Multi-Path (ECMP), Dynamic Load Balancing (DLB), and multi-planar topologies paired with Multipath Reliable Connection (MRC).

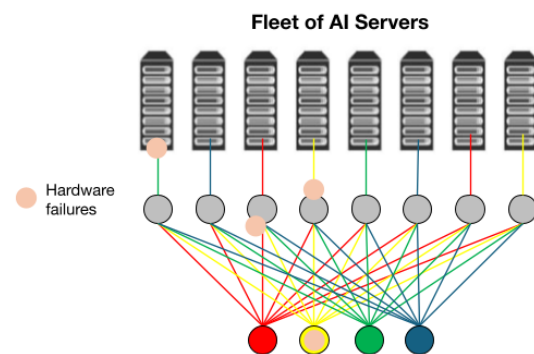


Figure 3: Component failures in different layers

System-Level Resiliency

At the system level, EOS detects physical link degradation, isolates unstable interfaces, and corrects memory bit flips to protect AI workloads from downtime.

Attribute	Benefit	Impact to AI fabric
Forward Error Correction (FEC) Monitoring	EOS detects physical link degradation before it escalates into uncorrectable packet loss or total link failure. EOS actively tracks pre-FEC BER metrics in real time. Rates exceeding 10^{-7} signal fiber contamination or transceiver degradation, generating proactive maintenance alerts.	While FEC mitigates bit errors, high AI data rates can quickly exhaust link correction margins, leading to packet loss, tail latency, and straggler nodes. EOS telemetry identifies degrading links early, allowing operators and automated systems to isolate failures before job completion times are impacted.
Link Flap Dampening	Mitigates the risk of failing optical links or loose cabling by automatically shifting excessively flapping interfaces into an error-disabled state.	This prevents localized physical layer issues from triggering repeated fabric-wide topology recalculations and routing instability across the AI fabric.
Memory bit error correction - Single Event Upset (SEU) Resiliency	EOS employs hardware-assisted background checks to continuously sweep internal control plane memory, routing tables, and ASIC registers, detecting and automatically repairing single-bit errors using Error Correcting Code (ECC) before they cause silent data corruption or system downtime.	Protects ultra-dense memory AI environments from cosmic radiation induced bit flips (Single Event Upsets), eliminating silent data corruption and avoiding system downtime across large-scale AI clusters.
Modular system - architectural resiliency	Modular systems like the Arista 7800R feature fully redundant, field-replaceable hardware components along with over-provisioned internal links. EOS continuously monitors internal connectivity between ingress silicon, fabric modules, and egress silicon, dynamically removing faulty links from the scheduling path and re-integrating them upon recovery without manual intervention. EOS supports Stateful Switchover (SSO) with sub-second supervisor failover by continuously mirroring control plane state, protocol topology databases, and hardware configuration maps to a standby unit. During switchover, EOS preserves BFD sessions to prevent interface flaps, holds VXLAN and EVPN control plane state stable, and maintains line-rate hardware forwarding.	Delivers hitless hardware recovery and maximizes cluster uptime. By automatically isolating degraded internal links without dropping traffic, it preserves predictable data plane performance and prevents packet loss or fabric disruption during high-density AI model training. Ensures complete control plane and data plane continuity during supervisor hardware failures. By avoiding overlay tunnel teardowns, packet loss, and fabric-wide routing reconvergence, it eliminates job-stalling network jitter and keeps synchronized AI workloads running without interruption.

Network-Level Redundancy

At the AI network level, EOS delivers resilient load balancing and path redundancy using Equal-Cost Multi-Path (ECMP) and Dynamic Load Balancing (DLB).

Attribute	Benefit	Impact to AI fabric
Resilient ECMP	Standard ECMP hashes traffic across multiple next-hops, but any link change forces a full recalculation that disrupts all flows. Resilient ECMP uses consistent hashing to assign flows to fixed paths so that link failures only redistribute traffic assigned to the failed interface, leaving flows on healthy paths completely unaffected.	In Front End AI fabrics, Resilient ECMP isolates link failures to affected flows, but redirecting elephant flows onto surviving paths risks buffer congestion. Consequently, DLB with RPR is recommended for Scale Out AI fabrics.
Dynamic Load Balancing (DLB) with Reactive Path Rebalancing (RPR)	DLB routes new flows to the least-congested link based on real-time port load and queue depth. RPR actively re-routes ongoing flows to better paths without waiting for idle gaps. Together, they rapidly migrate traffic to healthy paths during link failures, maximizing ECMP link utilization.	During link failures in AI fabrics, combining DLB and RPR rapidly shifts evicted and active flows to the least-congested surviving paths. This prevents post-failure buffer exhaustion, and protects Job Completion Times (JCT).

High Availability Using Multi-Plane Topology

Arista pioneered the universal leaf-spine architecture, establishing a foundation for ultra-low-latency, scalable, and resilient deployments. However, as AI workloads scale to 100,000+ parallel XPUs, traditional 3-tier single-plane Clos fabrics incur substantial overhead, doubling from 200% to 400% compared to 2-tier fabrics while driving up latency, power demands, and routing complexity. Leveraging high-radix switches allows this architecture to evolve into a multi-plane topology, as shown in Figure 4. Splitting high-speed host connections across multiple parallel network planes enables a 2-tier topology to connect massive XPU clusters while minimizing switch hops and tail latency.

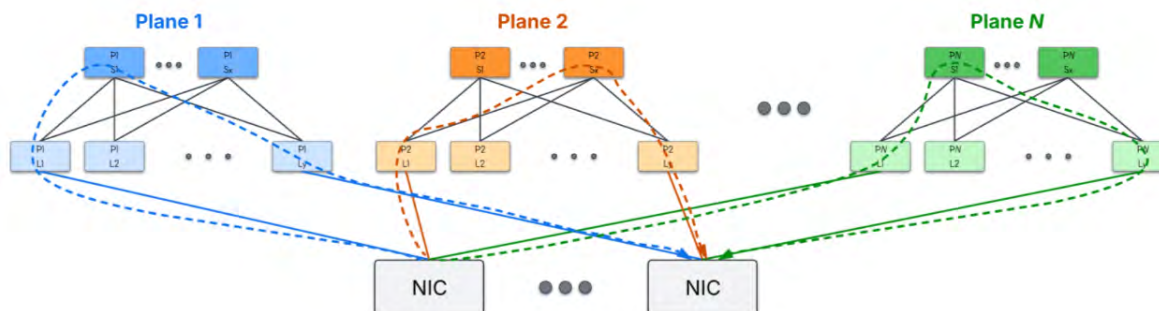


Figure 4: Multi-plane topology for resiliency

A multi-plane topology enhances AI fabric resiliency through two complementary mechanisms:

- Contained Blast Radius:** Segmenting physical traffic into dedicated, parallel planes prevents localized component, link, or protocol failures from impacting cluster-wide bandwidth. This design drastically minimizes the blast radius of network disruptions. For example, losing a leaf-to-spine link in an 800 Gbps single plane (64x800G with 32 leaf-spine links) deployment reduces network capacity by 3%, compared to 0.4% in an eight plane deployment with 100 Gbps per plane links. Similarly, losing a single NIC-to-leaf link forfeits only 12% of bandwidth (100G out of an 800G connection), allowing training jobs to seamlessly ride out link flaps on the remaining capacity.
- Multipath Reliable Connection (MRC):** Complements multi-plane topologies for resilient AI networking by providing adaptive load balancing and transparent fault recovery. When a link or switch fails, MRC dynamically reroutes traffic in real-time, preserving RoCEv2 memory semantics without requiring upper-layer application intervention. Once repairs complete, MRC automatically validates and reintegrates restored paths transparently, allowing long-running AI training jobs to progress uninterrupted without job stalls or restarts.

Hitless Operations and Lifecycle Management

Maintaining a massive AI fabric requires routine software patching, hardware maintenance, and heightened security updates to defend against emerging AI-powered cyber threats. Arista EOS Smart System Upgrade (SSU) provides a network-wide framework for maintenance, allowing network elements to be transparently inserted, updated, or removed without disrupting active AI workloads.

Spine upgrades with maintenance mode: At the spine layer, EOS leverages fabric redundancy to gracefully drain and reroute traffic. By invoking a single EOS command, operators place a spine switch into maintenance mode. EOS uses standard route-metric adjustments to gracefully divert traffic to adjacent spine nodes. Once all interfaces are drained, the operator can safely execute code updates, swap hardware modules, or replace cabling without impacting active compute runs. After maintenance is complete, disabling maintenance mode gracefully re-inserts the spine switch into the forwarding topology.

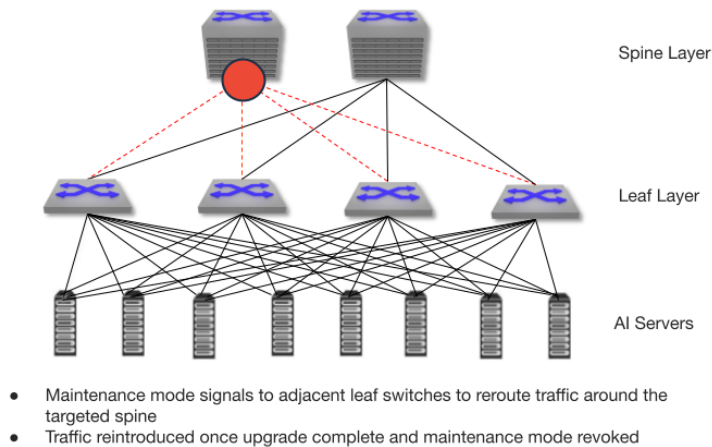


Figure 6: Spine Upgrade with maintenance mode

Leaf SSU with continuous connectivity: EOS Leaf SSU executes an ultra-fast operating system reload while preserving the active forwarding states inside the ASICs. The hardware continues forwarding data packets at line rate while the control plane boots up. Additionally, EVPN VXLAN SSU extends this capability across virtualized leaf-spine fabrics by holding encapsulation tables and VXLAN forwarding states intact across reboots, allowing switch upgrades without stalling active AI workloads.

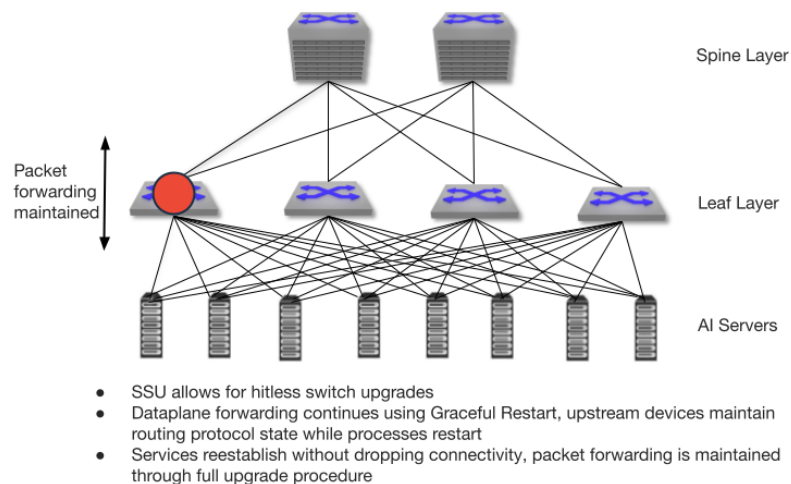


Figure 7: Hitless Leaf SSU

Protocol Graceful Restart: To prevent adjacent peers from tearing down routing tables during control plane reboots or supervisor switchovers, EOS allows Protocol Graceful Restart across BGP, OSPF, and IS-IS. When a switch initiates an update, it notifies its peers, which temporarily hold active routes as “stale but valid” in hardware while the control plane reconverges. If a specific routing agent requires a patch or restart, the switching ASIC continues forwarding traffic at line rate using pre-programmed hardware states, reconciling minor state changes as soon as the agent recovers.

Reducing Unplanned Activity with One EOS

Arista delivers simplified operations, hardened security, and high availability across the network through three integrated pillars: One Operating System (EOS), One Data Lake (NetDL), and One Management Engine (CloudVision with AVA).

One EOS Image Across All AI Fabrics

Standardizing on a single EOS binary across Front End, Scale Out, Scale Up, and Scale Across fabrics provides operational benefits:

- **Simplified Automation:** Standardizes DevOps workflows across all fabrics using a single integration model, streamlining end-to-end CI/CD testing and configuration.
- **Hardened Security:** Eliminates security gaps inherent to multi-NOS environments by consolidating software hardening and compliance into a unified platform.
- **Streamlined Upgrades:** Accelerates software qualification cycles, significantly reducing deployment risk and operational overhead.

NetDL: Unified Network-Wide Observability

High-availability AI fabrics require continuous, network-wide observability to detect performance anomalies proactively, contain failure blast radiuses rapidly, and ensure uncompromised workload resiliency.

Arista NetDL (Network Data Lake) serves as a unified multi-tenant time-series database. By capturing real-time EOS state streaming alongside third-party telemetry, flow metrics, packet captures, routing data, and AI job performance, NetDL eliminates traditional monitoring silos. This comprehensive data foundation speeds root-cause analysis during critical incidents and powers AI-driven network operations.

CloudVision AVA: Proactive AI-Driven Operations

Operating on top of NetDL, CloudVision AVA (Autonomous Virtual Assist) transforms network operations from reactive troubleshooting to proactive remediation:

- **Predictive Anomaly Detection:** Analyzes multi-domain telemetry to catch pre-failure indicators (such as degrading optics or latency jitter) before they trigger fabric outages.
- **Noise Reduction:** Correlates multi-source evidence to filter false positives, drastically lowering Mean Time to Detect (MTTD) and Mean Time to Resolve (MTTR) network issues.
- **Automated Investigation:** Delivers pre-packaged root-cause analysis and automated compliance auditing, maximizing network uptime while minimizing manual intervention.

Summary

To deliver high availability at scale, Arista EOS organizes its self-healing capabilities across different layers, as illustrated in Figure 8:

- **System & Software Resiliency:** At the foundational layers, pre-FEC error tracking, SEU memory correction, and modular state isolation fix degradations proactively, while hitless SSU and sub-second SSO ensure control-plane recoveries never impact the data plane.
- **Fabric & Planar Architecture:** At the network layers, combining DLB with RPR and deploying multi-plane topologies with MRC mitigate failure-induced hotspots, maintain uniform link utilization during outages, and prevent RoCEv2 packet loss or XPU stragglers.
- **Operational Management:** At the management layer, standardizing on a single EOS image across all fabrics (backed by NetDL and CloudVision AVA) transitions operations from reactive firefighting to predictive, automated anomaly remediation.



Figure 8: End-to-end resilient AI Network with EOS

In ultra-scale AI deployments, where component failures can stall tens of thousands of XPUs, network resilience directly determines infrastructure profitability. Arista EOS eliminates costly bottlenecks through a multi-layered approach to high availability, combining a self-healing, state-sharing architecture with intelligent, hardware-accelerated traffic management and hitless lifecycle operations via Smart System Upgrade (SSU). Running a single EOS image across every AI fabric layer simplifies automation, accelerates qualification cycles, and provides the operational stability required to keep massive XPU clusters running at peak efficiency.

References

- [The Many Facets of AI Fabrics](#)
- [Three Genius Ideas for AI Fabrics](#)
- [AI Networking Whitepaper](#)
- [AI Networking Solution Guide](#)
- [Multiplanar Scale Out fabrics with MRC](#)
- [Network Observability In Scale Out AI Fabrics](#)
- [EOS Architecture Whitepaper](#)

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



Copyright © 2026 Arista Networks, Inc. All rights reserved. CloudVision, and EOS are registered trademarks and Arista Networks is a trademark of Arista Networks, Inc. All other company names are trademarks of their respective holders. Information in this document is subject to change without notice. Certain features may not yet be available. Arista Networks, Inc. assumes no responsibility for any errors that may appear in this document. 10/12/26 02-0121-01