ARISTA

White Paper

Network Observability In Scale Out AI Fabrics

Introduction

Scaling AI clusters to thousands of XPUs means even minor fabric inefficiencies, such as an oversubscribed link, path latency asymmetry, or optic failure, create a massive financial drain by leaving expensive XPUs idle. Because distributed training, inference, and physical AI workloads depend on continuous collective communication across Front End, Scale Out, and Scale Across fabrics, a single network bottleneck instantly inflates Job Completion Time (JCT) and Time to First Token (TTFT). Unlocking maximum AI infrastructure ROI requires shifting from isolated device monitoring to AI workload-aligned network observability.

This paper details how Arista EOS provides complete end-to-end network telemetry in Scale Out fabrics across the full XPU-to-XPU traffic path, empowering operators to eliminate congestion and drive peak cluster utilization across:

- **XPU NICs:** Host interfaces and local buffer state
- **Access Links:** Server-to-switch connectivity, optics health, and signal integrity
- **Switch Hardware:** Queue dynamics, buffer utilization, and microburst tracking
- **Inter-Switch Fabric:** Switch-to-switch links, optics health, ECMP load balancing, and hop-by-hop latency

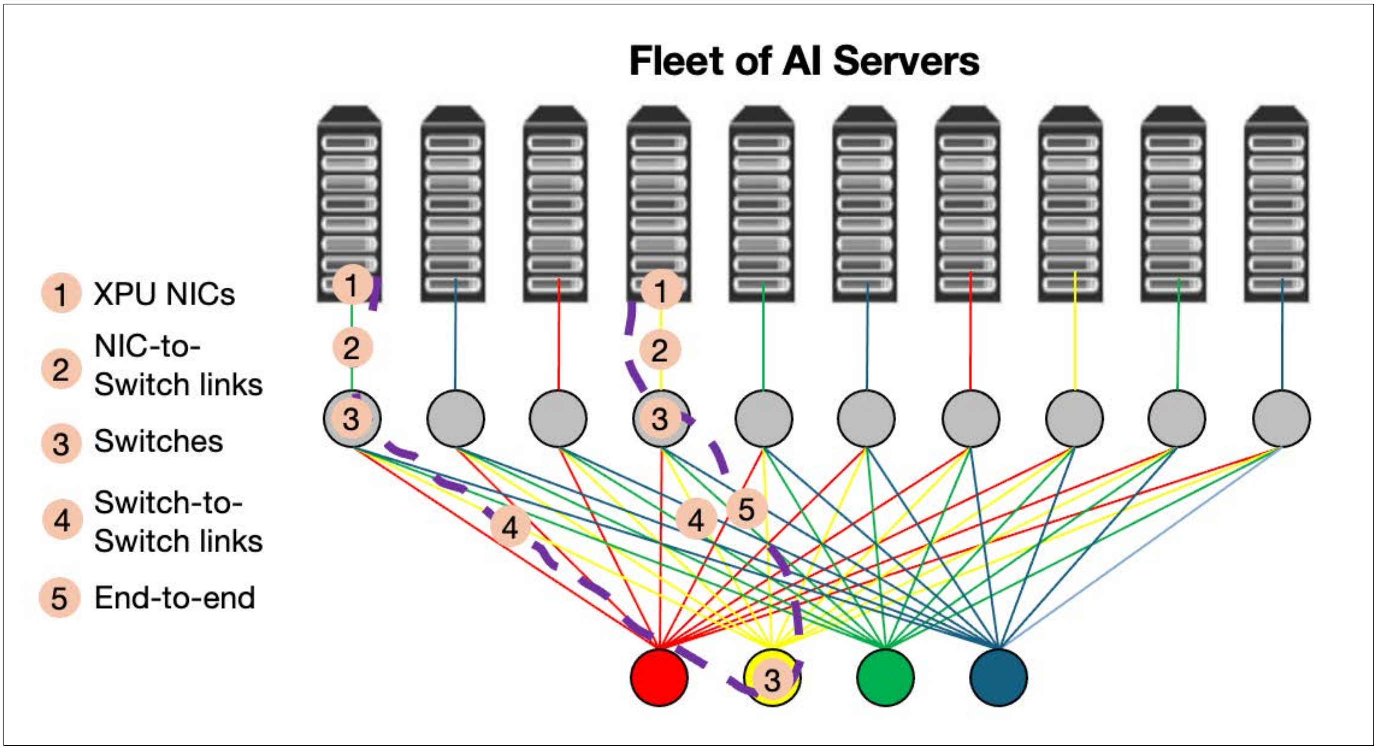


Figure 1. Points of interest for XPU-to-XPU traffic observability

XPU NIC Observability using EOS AI Agent

Siloed server and switch management poses severe risks in tightly coupled AI clusters, where a single server misconfiguration or failure can stall thousands of accelerators. Detection of the misconfiguration is like finding a needle in a haystack across hundreds or thousands of servers.

The **Arista EOS AI Agent** eliminates the silos by deploying a lightweight daemon onto servers or SmartNICs. The NIC Controller Agent (NCA) running on the EOS switch periodically gathers the metrics from attached EOS AI Agents and makes them available through EOS APIs. The metrics collected include Ethtool counters, RDMA counters, PCIe counters, network counters, and PFC/DSCP/ECN configuration. This unified telemetry gives operators complete visibility of the flows entering the network to remediate fabric bottlenecks before they cause network pauses affecting the XPU utilization and AI job completion times. Key performance metrics collected using the EOS AI Agent are listed below.

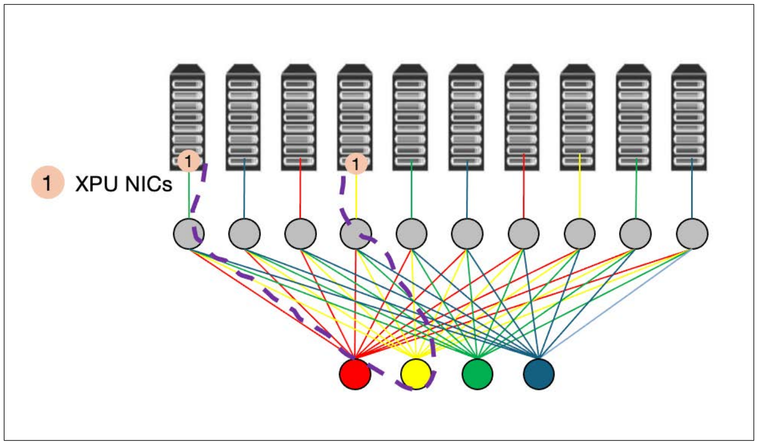


Figure 2. XPU NIC Observability

Area	Metrics	Possible Issue	Impact to AI Jobs
Physical Layer troubleshooting	Received packets dropped / discarded	Internal NIC buffers are overwhelmed	The network is outrunning the host, forcing the collective communication to drag out.
	Packets received with CRC errors / length errors	Caused by damaged cables, loose transceivers, or dirty optics	Introduces network jitter and tail latency; even a tiny corruption rate can inflate overall Job Completion Time (JCT).
	Buffer allocation errors	Out of memory for incoming traffic	A host-level configuration bottleneck starving the XPU synchronization.
Lossless Fabric Validation	PFC Frames received	Upstream switch buffers are saturated throttling the NIC	Pause frames spread backward idling XPUs.
	PFC Frames transmitted	NIC receive buffers are full	
Congestion Monitoring	Notification Point ECN marked packets received	Congestion in intermediate switches	Early warning of fabric saturation; if unmanaged, it will escalate into PFC generation, creating tail latency.
	Reaction Point ECN CNP sent by destination NICs back to the source	To throttle the traffic for alleviating congestion	Intentionally slows down data flows to prevent packet drops.
	Rising Reaction Point ECN CNP ignored	Source NIC is failing to throttle	Causes packet drops and increases job completion time.
	RDMA out of buffer / out of sequence packets	Packets are dropped or misordered due to buffer exhaustion	Freezes XPU execution; dropping network throughput to a crawl.
	RDMA adaptive retransmissions	Link flaps resulting in packet drops or path degradation forcing RDMA back to slow-start recovery	Destroys network throughput and induces severe tail latency, inflating JCT and TTFT.

Link Quality and Physical Layer Monitoring

Maintaining physical layer integrity is critical to avoiding tail-latency and cluster instability in high-performance AI Scale Out fabrics. Layer 1 (L1) anomalies often manifest as unstable links, packet drops, or complete port shutdowns. By leveraging Arista EOS interface telemetry and Digital Optical Monitoring (DOM), operators can proactively isolate and remediate these faults.

With the granular visibility and early detection capabilities of EOS, operators can either traffic engineer around the fault or schedule targeted maintenance to prevent performance degradation in AI workloads.

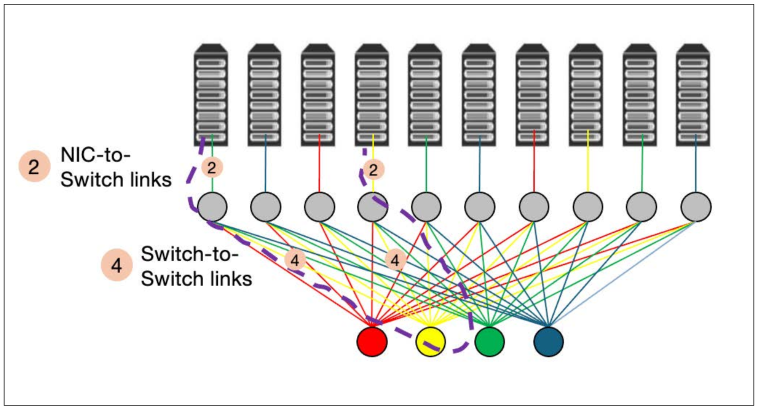


Figure 3. Link Monitoring

Area	Error / Alarm	Possible Issue	Impact to AI Jobs
Link Monitoring: Tracking FEC metrics, link configurations, and flaps gives proactive Layer 1 visibility to stop outages before they happen	Link flapping: Results in RDMA adaptive retransmissions.	Driven by low light levels, inconsistent RX signals, or faulty cabling. The port is placed into an error-disabled state after 5 flaps in 10s (default).	Instantly breaks the collective communication topology. The entire training job may abort, forcing a cluster-wide rollback to the last checkpoint.
	Configuration Mismatches	Speed, duplex, and Forward Error Correction (FEC) modes are misaligned, keeping the link down.	Shrinks total fabric bisection bandwidth. It prevents the node from passing health checks, causing job scheduling blocks or leaving the cluster under provisioned.
	Pre-FEC Bit Error Rate (BER): Number of errored bits relative to the total transmitted bits before they are corrected by FEC.	Anything greater than 10^{-7} signals that the link quality is actively degrading due to fiber contamination or aging transceivers.	Forces constant hardware error correction cycles, adding network jitter. This turns the affected host into a "straggler node" that stalls the entire synchronous training step.
	FEC Histogram: categorizes every received codeword based on the number of corrected symbol errors it contains.	Codewords with >= 16 errors are uncorrectable and cause packet loss due to degrading optics or failing SerDes lanes.	
	MAC Rx Local Fault: The local side is detecting an issue with the incoming signal. MAC Rx Remote Fault: The link peer (the remote switch/host) is detecting an issue with the incoming signal.	Loss of electrical signal, high amounts of FEC uncorrected codewords, and optics RX power low are common conditions.	Link is down to prevent blackholing traffic, resulting in lower bandwidth available.

Area	Error / Alarm	Possible Issue	Impact to AI Jobs
Digital Optical Monitoring (DOM): Always-on and tracks Optical Power (TX/ RX), transceiver temperature and voltage	Low RX Power	Dirty connector, excessive link loss, or fiber bend.	Triggers intermittent packet corruption and high adaptive retransmissions, drastically degrading network throughput.
	Low TX Power	Laser degradation, transceiver failure, or lane not activated.	Signals imminent link death; destabilizing the RDMA Queue Pairs (QPs) and risking a mid-run job failure.
	High temperature alarm	Inadequate rack cooling, fan failure or high ambient temperature.	May trigger the optic to thermally throttle its transmission speeds to prevent permanent damage, introducing sudden, severe tail latency into the collective communications.
	Voltage alarm	Power supply issues to the transceiver module.	May cause unrecoverable in-flight data loss and job failures.

Switch Counter and Queue Monitoring

AI workloads flow through the network on microsecond and millisecond timescales, creating a massive visibility gap when counters are polled on intervals of seconds. Additionally, because RoCEv2 is built as a lossless protocol, fabric degradation rarely manifests as explicit packet drops. Relying on legacy monitoring metrics means missing the true bottlenecks. Arista EOS solves this by providing hardware-accelerated monitoring tools that track switch queues and traffic counters at microsecond scales.

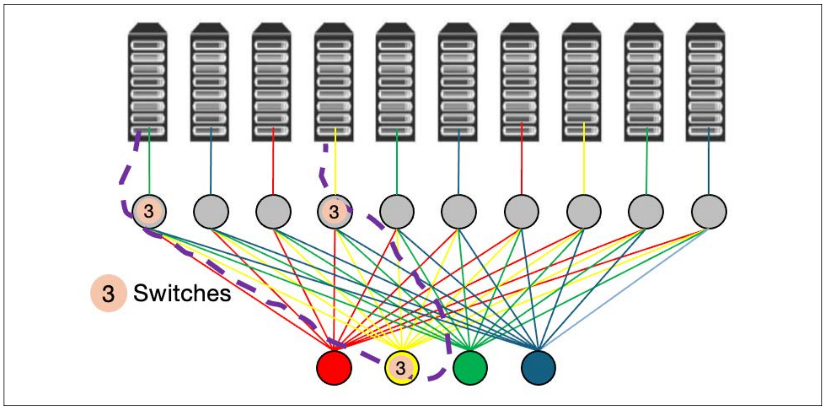


Figure 4. Switch Monitoring

Real Time Streaming and High Resolution Telemetry

Rather than relying on periodic polling, EOS streams granular, event-driven network state changes in real time via gNMI. By publishing network state updates the moment they happen, covering everything from interface health and routing tables to traffic counters, EOS gives operators immediate, actionable visibility into overall network health.

The **Arista EOS AI Analyzer** monitors interface traffic counters at an ultra-granular, 100-microsecond resolution. This hardware-driven capability captures transient bursts and "incast" events, providing Equal-Cost Multi-Path (ECMP) member utilization analysis over sub-millisecond timescales. This level of precision allows operators to observe micro-variations in traffic distribution that traditional tools miss. For example, during a 60-second test window, the AI Analyzer tracked interface statistics for an 8-way uplink ECMP group across 600,000 continuous samples. The resulting dataset was exported to a time-series database and visualized in Grafana as shown in the figure below. The telemetry reveals a stark contrast between legacy polling and microsecond tracking. The 100-microsecond granular visualization exposes significant traffic volatility and micro-congestions. By capturing fabric behavior on the exact timescales that AI workloads operate, the EOS AI Analyzer delivers the visibility required to validate load-balancing efficiency and optimize application-to-network orchestration.

Figure 5. 100us level traffic visibility with EOS AI Analyzer

Arista EOS Latency Analyzer (LANZ)

Latency Analyzer (LANZ) provides real-time reporting on interface congestion and queuing latency, shifting operators from reactive troubleshooting to proactive fabric management. By delivering high-resolution visibility into switch buffer utilization, LANZ exposes buffer pressure before application performance degrades, eliminating the guesswork in tuning Priority Flow Control (PFC) and Explicit Congestion Notification (ECN) thresholds. Correlating this LANZ telemetry with interface ECN, PFC, and queue drop counters reveals exactly how transient congestion impacts AI job performance, enabling precise buffer optimization and peak cluster efficiency.

Figure 6. Sample LANZ congestion event

Advanced Flow Telemetry

To complement microburst queue and counter tracking, Arista EOS supports standard flow monitoring tools such as sFlow and IPFIX. Utilizing hardware-accelerated engines, EOS handles aggressive sampling without impacting control-plane performance. In AI fabrics, sFlow helps identify:

- **Elephant Flows:** Sustained, high-bandwidth data bursts
- **Fabric Hotspots:** Unbalanced paths across the topology
- **Noisy Neighbors:** Non-essential traffic disrupting latency-sensitive queues

End-to-end Network Monitoring

In Arista EOS, monitoring for high-performance AI networks is engineered directly into the hardware and software architecture using **Inband Network Telemetry (INT)**. Arista EOS implements INT based on the Inband Flow Analyzer (IFA 2.0) specification.

Inband Network Telemetry is used to gather per flow telemetry information like path, per hop latency and congestion. This information is exported to the collectors using protocols like IPFIX and may be stored in a time series database for providing per flow historical telemetry information and network analytics.

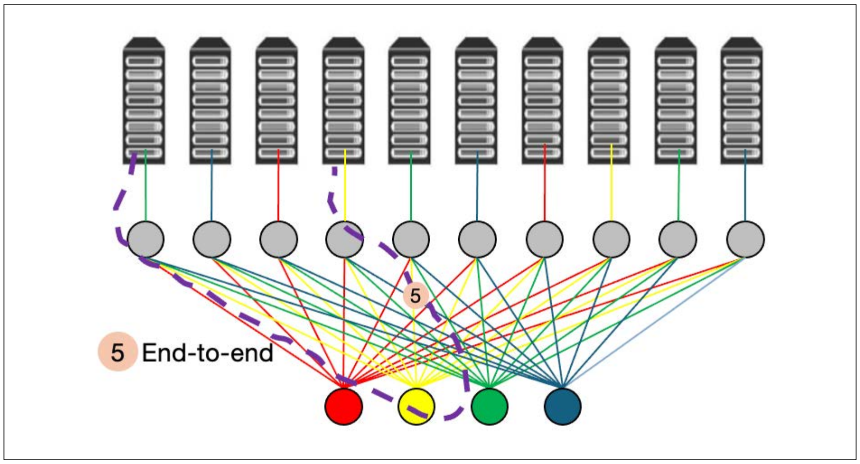


Figure 7. End-to-end Network Monitoring

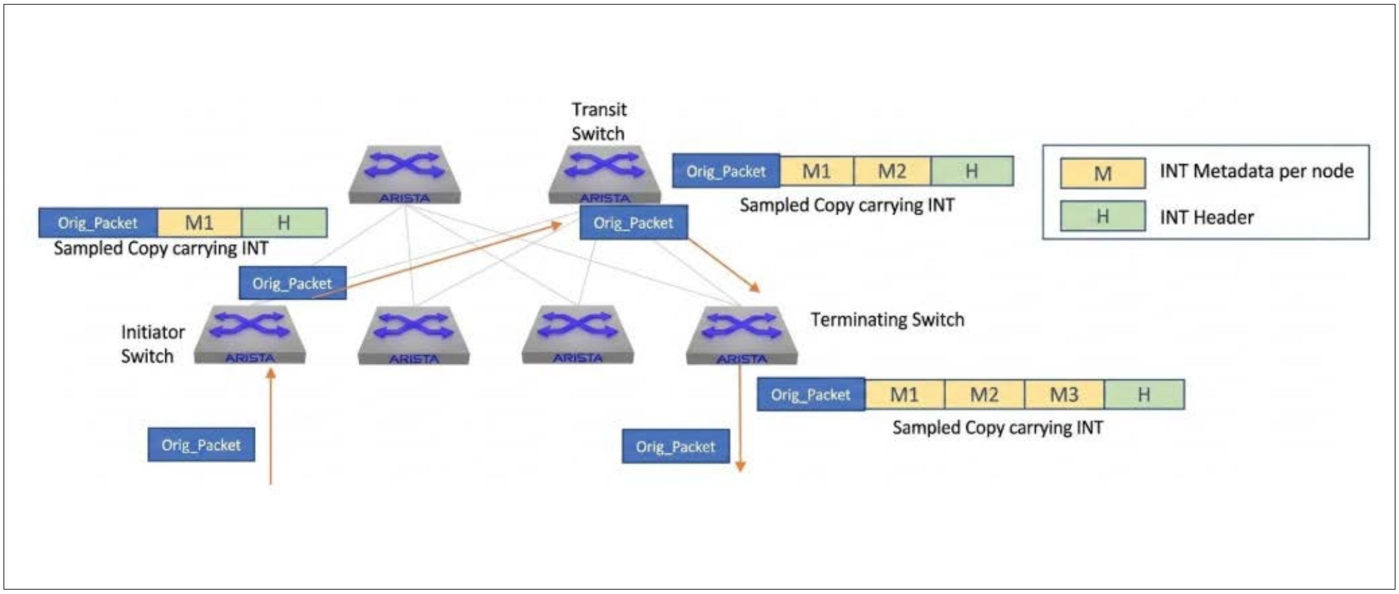


Figure 8. EOS Inband Network Telemetry

With INT, every participating Arista switch adds critical, real-time insights towards end to end network monitoring.

Metric	End to End Network Insight
Per-Hop Latency: EOS records precise ingress and egress timestamps at each hop.	Pinpoints Exact Bottlenecks: Subtracting ingress from egress timestamps provides the absolute residence time inside the switch down to the nanosecond, instantly identifying which device added delay and helps detect any anomalies.
Egress Queue Occupancy: Captures the exact buffer queue ID being used and if a threshold is passed, it flags emergent congestion.	Reveals transient queue buildup and buffer contention that traditional polling tools miss, explaining why and where packets experienced delay.
Path Tracking: Using Switch IDs, and Egress Port IDs embedded directly in the telemetry headers.	Allows operators to visualize the exact hop-by-hop path a flow took through an ECMP fabric and measure total end-to-end latency.

Unifying AI Fabric Observability

Arista offers multiple EOS telemetry consumption options.

- **Streaming Telemetry:** Replaces legacy SNMP polling with a proactive, push-based model. EOS continuously streams event-driven network states (such as interface counters, BGP events, packet drops) directly to time-series collectors with minimal CPU overhead, enabling real-time observability.
- **EOS API (eAPI):** Offers programmatic, JSON-formatted key-value access to manage and monitor EOS switches. This replaces brittle screen-scraping methods with robust APIs, making it effortless to ingest network telemetry into unified monitoring stacks.
- **Arista CloudVision:** Provides an out-of-the-box management platform powered by full-stack EOS telemetry. For custom builds, operators can leverage eAPI and streaming telemetry based on CloudVision's proven reference design (detailed below).

Arista CloudVision unifies network, host, and workload metrics into a single AI observability platform. Powered by Network Data Lake (NetDL), it aggregates multi-domain telemetry to give operators full cluster visibility. With the Arista Autonomous Virtual Assistant (AVA) built-in, operational teams can query network health in natural language, turning complex troubleshooting into immediate, actionable insight.

By unifying host and network telemetry, CloudVision enriches standard network events, packet flows, and queue metrics with specific AI Job IDs. This contextual mapping drives fast, job-centric troubleshooting that directly links network performance to AI workload health.

Job Health Auditing: An operator begins at a high-level application dashboard to identify an unhealthy or failed AI job.

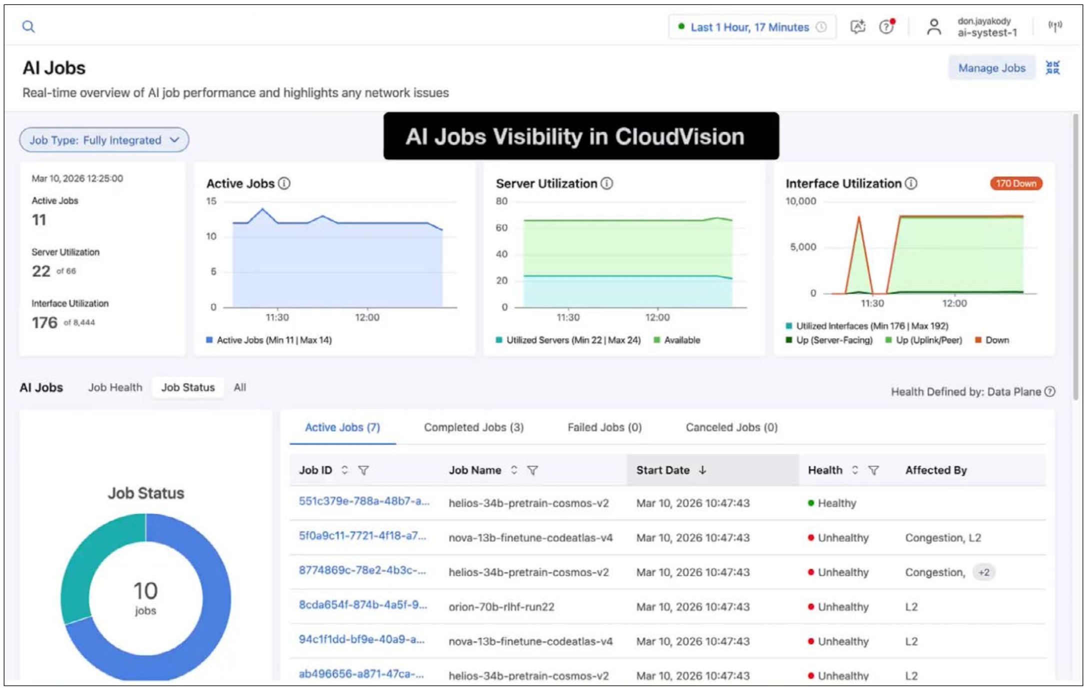


Figure 9. AI Jobs Dashboard in CloudVision

Congestion Correlation: CloudVision helps correlate the specific Job ID with fabric-wide telemetry, mapping it directly to active ECN markings, PFC pause frames, or LANZ queue anomalies occurring at that exact time.

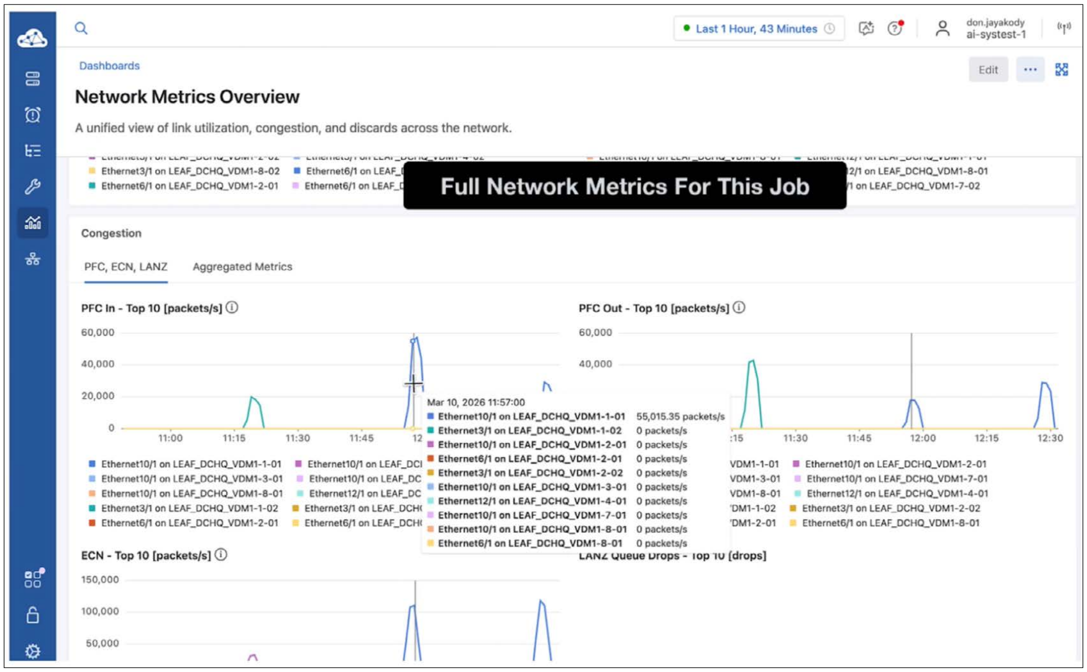


Figure 10. AI Job to network metrics correlation in CloudVision

Granular Flow Analysis: Using CloudVision operators can zoom directly down to the flow level to isolate specific device events, hardware drops, or path asymmetries responsible for the latency.

RoCEv2 Flows with Queue Pair (QP)

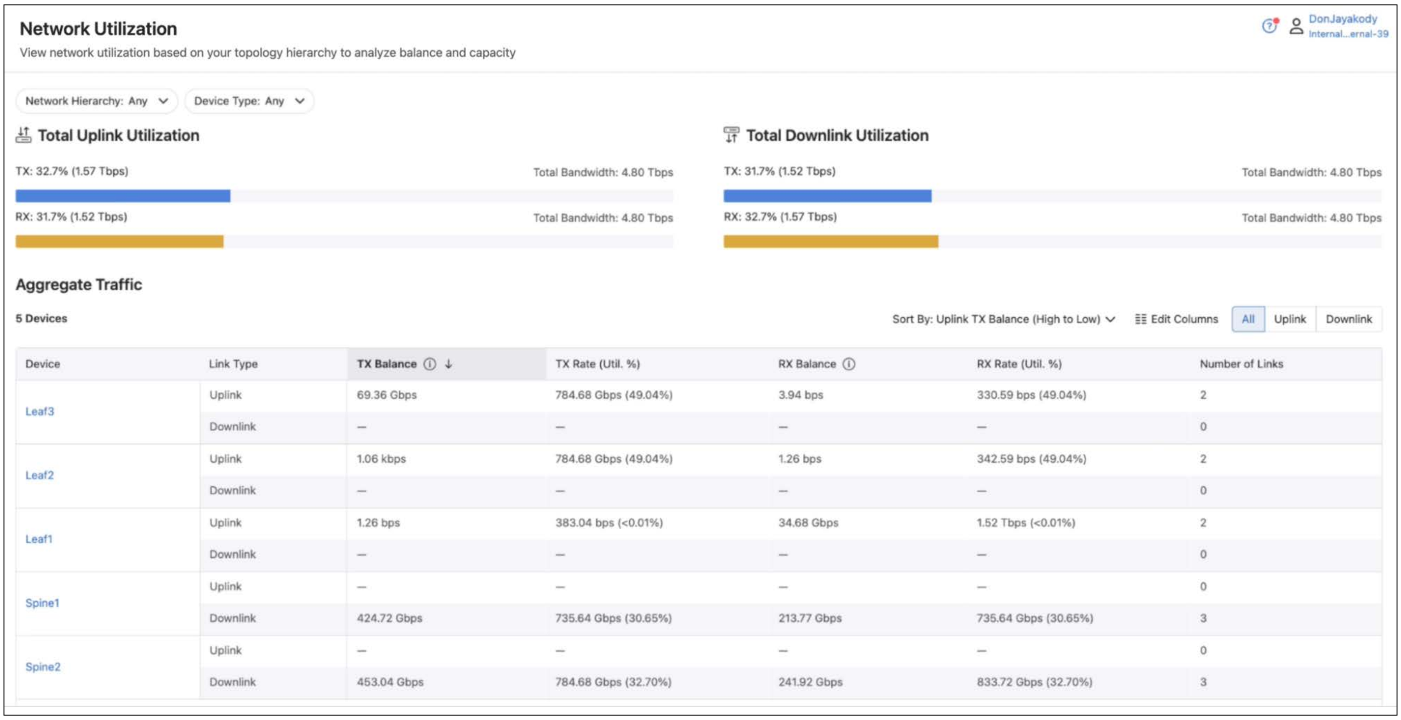
Top Flows

Sort by: CNP (High to Low)

Host 1		Host 2	Hops	CNP	NACK	Bytes	Packets
172.20.254.19	⇄	172.20.253.81	1+	42	23	20.5 TB	16.4 G
QP: 0x14FE13	←	UDP: 14785	1+	42	23	5.0 KB	50
UDP: 12785	→	QP: 0x14FE23	1+	—	—	10.2 TB	8.2 G
UDP: 11321	→	QP: 0x12FE11	1+	—	—	10.3 TB	8.2 G
10.81.117.103	⇄	10.11.12.13	1	12	22	30.5 TB	2.9 G

Figure 11. Granular flow visibility in CloudVision

Traffic Imbalance Detection: CloudVision indicates how well balanced AI traffic is between ECMP links. TX / RX Balance shows the standard deviation of traffic rates to measure balance across links, higher value indicates a more uneven traffic distribution. This helps detect how well traffic load balancing works in the Scale Out fabric.



Device	Link Type	TX Balance	TX Rate (Util. %)	RX Balance	RX Rate (Util. %)	Number of Links
Leaf3	Uplink	69.36 Gbps	784.68 Gbps (49.04%)	3.94 bps	330.59 bps (49.04%)	2
	Downlink	—	—	—	—	0
Leaf2	Uplink	1.06 kbps	784.68 Gbps (49.04%)	1.26 bps	342.59 bps (49.04%)	2
	Downlink	—	—	—	—	0
Leaf1	Uplink	1.26 bps	383.04 bps (<0.01%)	34.68 Gbps	1.52 Tbps (<0.01%)	2
	Downlink	—	—	—	—	0
Spine1	Uplink	—	—	—	—	0
	Downlink	424.72 Gbps	735.64 Gbps (30.65%)	213.77 Gbps	735.64 Gbps (30.65%)	3
Spine2	Uplink	—	—	—	—	0
	Downlink	453.04 Gbps	784.68 Gbps (32.70%)	241.92 Gbps	833.72 Gbps (32.70%)	3

Figure 12. Network Utilization view in CloudVision

Summary

The exponential growth of Large Language Models (LLMs) and Physical AI is driving compute demands to unprecedented levels. As AI clusters scale to thousands of XPUs, minimizing Job Completion Time (JCT) and Time to First Token (TTFT) requires real-time, end-to-end network observability. While application teams expand raw compute capacity, network engineering teams must continuously optimize the interconnect to prevent cluster-wide bottlenecks.

Arista EOS bridges this gap by delivering raw performance with operational simplicity, scale, and visibility through job-aligned telemetry. This empowers operators to fine-tune congestion control, optimize workload placement, and maximize overall token throughput. Powered by a single, unified OS across the entire portfolio, Arista EOS eliminates fragmented management layers, unlocking full-stack observability eliminating silent XPU starvation, safeguarding multi-week training investments, and delivering the deterministic, low-latency performance required to successfully monetize real-time AI services at scale.

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