Solution Highlights

Open Standards

Arista and Broadcom are committed to supporting open, standards-based Ethernet and IP technologies to maximize choice, flexibility and performance for best-of-breed networks.

End-to-end Performance Optimized RoCE

Arista and Broadcom partnered to test and certify end-to-end RoCE with various congestion control mechanisms to ensure best-in-class performance.

Power Efficient NIC, Switches, and Interconnects

Lowest power and highest performance NICs, Switches, and Interconnects reduce infrastructure TCO, leaving more power for accelerators.

End-to-end Telemetry and Network Management

CloudVision™ integrates NIC and Switch management, with automated configuration and management of RoCE and congestion control on Arista switches and Broadcom NICs.

Simple Configuration

RoCE deployment is made simple by offering an End-to-end Performance Optimized baseline configuration that is easily tunable.

Increased Reliability

Lower power, less thermally challenged solutions result in improved MTBF so the network is highly reliable and always available for the Accelerators.

Overview

The rapid advancement of AI has driven an unprecedented need for AI data centers that can deliver optimal performance to support a variety of AI workloads. This necessitates maximizing network bandwidth and minimizing latency. To achieve the level of performance required for AI workloads, it is essential to optimize network performance and TCO with efficient, highly reliable designs that implement robust congestion control, load balancing and telemetry. Broadcom and Arista have collaborated to address these requirements by providing high-performance network hardware and fine-tuning key parameters to deliver end-to-end 400G and 800G AI network solutions. As AI networking evolves, so do the network switch deployments.

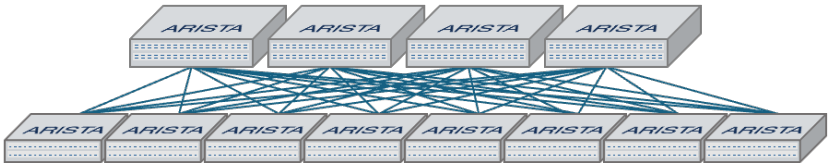

Fig 1: Arista switches in Tiered Leaf Spine/Plane based design

Arista's portfolio of fixed and modular systems enable clusters of tens of thousands of accelerators, retaining efficient 2-tier topologies which minimize complexity and costs while maximizing performance and reliability.

With commitment to open standards, Arista's portfolio provides customers with maximum choice in accelerators, NICs and storage while supporting all common cluster deployment configurations; e.g Clos including multi-stage, rail and plane topologies.

Features

To satisfy the high performance requirements of AI/ML cluster back-end cluster networks, a set of advanced features that are not commonly found in traditional networks are required. Broadcom and Arista provide a comprehensive set of features engineered to optimize performance, minimize congestion, and enhance overall network reliability and resilience for 400G and 800G AI network implementations.

- **RoCEv2**, the standard protocol for Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), is the optimal solution for maximizing data center performance. It enables direct memory access over a network and is essential for achieving the high-bandwidth and low-latency to maximize compute performance. Broadcom and Arista have established a partnership to support RoCEv2 across multiple generations of products.
- **Data Center Quantized Congestion Notification** (DCQCN - p: Probabilistic, d: Deterministic) DCQCN is an end-to-end congestion control mechanism for RoCE that enables the NIC and switch to actively detect and respond to network congestion. It is aligned with standards and delivers a robust baseline performance.
- **Priority Flow Control** (PFC), utilized by RoCE, is essential for establishing a lossless network. It prevents packet loss due to switch buffer overflows by pausing specific traffic classes during congestion, instead of halting all traffic on a link.
- **Equal-Cost Multi-Path** (ECMP) is essential for creating a congestion-free multi-hop network (Leaf + Spine). ECMP employs a hashing algorithm to route flows in a balanced manner as messages traverse from the Leaf through the Spine, ensuring solid baseline network performance.

- **Programmable UDP Source Port**, Complementary to ECMP, the Programmable UDP Source Port feature enables granular control at the Queue Pair level. Despite ECMP, some network congestion may occur; if detected, the Programmable UDP Source Port feature in the NIC allows for changing the source port to avoid congestion, thereby enhancing flow control and optimizing load balancing.
- **Dynamic Load Balancing** (DLB), also known as Adaptive Routing, is a feature that dynamically alleviates congestion within the network. DLB identifies congestion and enables the switch to reroute queue pairs or flows to paths with minimal or no traffic, ensuring maximum performance.
- **Cluster Load Balancing** (CLB), provides an application aware approach to handling traffic for large AI training workloads. CLB monitors the demands of the workload to ensure that all uplinks are utilized optimally and that there is no oversubscription on any of the links going from spine to leaf.
- **Programmable Congestion Control** (PCC): Broadcom's PCC facilitates a custom Congestion Control (CC) algorithm that can be developed alongside the existing DCQCN CC algorithm to meet specific performance optimization objectives, offering ultimate flexibility and tuning.

Arista EOS (Extensible Operating System) and Datacenter Switches

Arista EOS® (Extensible Operating System) delivers a high-bandwidth, low-latency, lossless network that can scale to support hundreds or thousands of XPUs at 100G/200G/400G/800G speeds, addressing the challenge of interconnecting XPUs for modern AI applications. EOS enables a premium lossless network through traffic management configurations, adjustable buffer allocation schemes, and PFC and DCQCN for RoCE deployments. Arista's DLB and CLB features maximize network forwarding efficiency by minimizing or avoiding congestion. Latency Analyzer (LANZ), a feature of EOS, monitors interface congestion and queuing latency with real-time reporting to simplify the configuration of appropriate PFC and ECN thresholds. This visibility into network buffer utilization allows for correlation between application performance and network congestion events, which in turn supports optimal configuration of PFC and ECN values.

Arista Portfolio	Product Description
7060X5 and 7060X6	High Density 400G and 800G Fixed Switch Portfolio for AI and DC
7280R and 7800R	High Performance 400G and 800G Dynamic Deep Buffer Platforms
7700R4	800G Distributed Etherlink Switch for Accelerated Computing

Table 1: Arista Datacenter Switches

Broadcom Ethernet NIC Adapters

Broadcom offers a broad portfolio of Ethernet NIC Adapters with port speeds ranging from 1 Gbps to 400 Gbps, delivering best-in-class performance, hardware acceleration, and offload capabilities that result in higher throughput, higher Central Processing Unit (CPU) efficiency, and lower workload latency for Ethernet/IP as well as RoCE traffic.

Broadcom's latest AI-focused Ethernet adapters are based on BCM576xx (Thor2) ASIC and support 400GE, 200GE, 100GE, and 25GE and are available in both Open Compute Project (OCP) and Peripheral Component Interconnect Express (PCIe) form factors. All BRCM576xx NICs are optimized for AI applications and support key features including RoCEv2, DCQCN, PFC, and PCC to ensure maximum network bandwidth and the lowest latency. Additionally, leveraging advanced silicon processes, Broadcom's adapters are the lowest-power 400G interfaces available today, reducing overall power and cooling demands, reducing costs and improving network reliability. Lastly, Broadcom's AI NICs support the most diverse cabling options available today for the network inter-connectors which have a direct impact on the network power, cost, and reliability.

Part Number	ASIC	Ports	I/O
BCM957608-N1400GD	BCM57608	1x 400G	QSFP112-DD
BCM957608-N2200G	BCM57608	2x 200G	QSFP112

Table 2: Broadcom OCP3.0 NIC Adapters with RoCE support

Part Number	ASIC	Ports	I/O
BCM957608-P1400GD	BCM57608	1x 400G	QSFP112-DD
BCM957608-P2200G	BCM57608	2x 200G	QSFP112

Table 3: Broadcom PCIe NIC Adapters with RoCE support

When selecting cabling for servers and switches in a data center, it is imperative to consider the interconnect type. This decision affects the data center's reliability, power usage, cooling requirements, and overall cost. Broadcom and Arista have collaborated to offer various pre-qualified cabling options that ensure seamless integration for a complete end-to-end solution. These options include DAC copper cables (up to 5 meters), Active Electrical Cables, Optical Cables, and Linear Pluggable Optic (LPO) Cables, which provide reduced power consumption and enhanced reliability compared to traditional Optical solutions.
Additional details for Arista Optics and Q&A for reference

Cable	Distance	Power	Reliability	Cost	MPN
Copper Cable (DAC)	5m	Low	High	Low	Amphenol: DJERGN-0003
Active Electrical Cable (AEC)	7m	Medium	Medium	Medium	Credo: CAC82X321A2N-CO-HW
VSR Optical Transceiver	50m	High	Low	High	Switch: Arista OSFP-800G-2VSR4 NIC: Eopotlink EOLQ-854HG-01-M
DR Optical Transceiver	500m	High	Low	High	Switch: Arista OSFP-800G-2XDR4 NIC: Hisense LMQ3621S-PC1
DR Linear Pluggable Optic (LPO)	500m	Medium	Medium	Medium	Switch: Arista LPO-800G-2DR4 NIC: Eoptolink EOLQ-134HG-5H-MSL

Table 4: Comparison Metrics for different Cable/Optics Types

Arista CloudVision

Arista CloudVision is a multi-domain management platform that uses cloud networking principles to simplify network operations. It's built on Arista's Network Data Lake (NetDL) architecture, aggregating data from across an enterprise. Uses AI and machine learning (ML) to analyze network data and provide insights, updates, and alerts. It also uses predictive insights from Arista Autonomous Virtual Assist (Arista AVA).

Arista AI Agent

The AI Agent integrates NICs and Arista's EOS network operating system to help manage and monitor switches and NIC connections and debug server-level issues. Arista's AI Agent and CloudVision software work together to provide a unified view of network and server statistics, which can help network engineers troubleshoot issues more efficiently.

Benefits of using the AI Agent and CloudVision together include:
Improved troubleshooting: Network engineers can correlate network events with server-side issues.
Real-time insights: CloudVision provides real-time insights, updates, and alerts.
Anomaly identification: CloudVision uses AI/ML to identify real anomalies and distinguish them from noise.
Network operations: CloudVision's AI/ML algorithms help improve network operations.

Summary

Arista and Broadcom are committed to fulfilling the evolving requirements of today's AI applications, and future workloads. This commitment entails implementing a robust, pre-configured solution that delivers a highly scalable 400G or 800G end-to-end optimized network. The partnership prioritizes the integration of power-efficient and reliable NICs, switches, and interconnects to maximize network availability and accelerator efficiency. This rigorously tested and validated solution ensures rapid deployment, enabling AI workloads to stood up and become productive as quickly as possible.

References

- Arista Cloud Grade Routing Products
- Arista Hyper-Scale Data Center Platforms
- Arista EOS Quality of Service
- Arista Priority Flow Control (PFC) and Explicit Congestion Notification (ECN)
- Arista Configuration Guide
- Arista EOS Software Downloads
- Arista AI Networking
- Arista Cloud Vision
- Broadcom Ethernet Network Adapters
- Broadcom Ethernet NIC RoCE Features
- Broadcom Ethernet NIC Configuration Guide
- Broadcom Ethernet NIC Firmware and Drivers Downloads
- Broadcom RoCE Configuration Guide
- Congestion Control for Large-Scale RDMA Deployments
- RoCE Deployment Guide

Headquarters

5453 Great America Parkway
Santa Clara, California 95054
408-547-5500

Support

support@arista.com
408-547-5502
866-476-0000

Sales

sales@arista.com
408-547-5501
866-497-0000