Solution Highlights

Higher Bandwidth

Arista switches and Broadcom NICs provide high speed Ethernet connectivity suitable for data and compute intensive workloads.

Open Standards

Arista and Broadcom are committed to supporting open, standards based congestion control using mechanisms like PFC and ECN marking.

Pre-Certified end-to-end RoCE

Arista and Broadcom partnered to test and certify end-to-end RoCE with various congestion control mechanisms. Optimizing RoCE involves managing multiple traffic classes, tuning ECN thresholds, and other parameters. These settings have been vetted in the lab for different applications.

Simple Configuration

Arista and Broadcom together make deploying RoCE simple. Automation suites are available to configure and manage RoCE and congestion control on Arista switches and Broadcom NICs.

Lower Power

Arista and Broadcom are focused on providing power efficient Ethernet solutions on Switches and NICs. Lower power is extremely critical for datacenters, and it contributes to TCO savings.

Overview

Datacenter networking has evolved over the years and with the proliferation of AI/ML, disaggregated storage and High-Performance Computing (HPC), today's data centers require a high performance, low-latency network. With ever increasing database sizes and demand for high bandwidth for the movement of data between processing nodes, a reliable transport is critical.

With Ethernet being deployed ubiquitously and leading the industry with 400G and marching towards 800G, RDMA over Converged Ethernet (RoCE) is the preferred solution for modern datacenters as it provides the high performance of RDMA over Ethernet. RoCE bypasses the kernel to provide lower CPU utilization, lower latency and higher throughput than TCP, as illustrated in Figure 1.

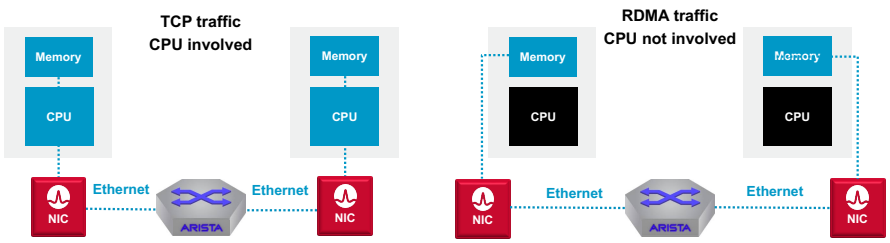

Figure 1: RoCE with Broadcom NIC and Arista switch

As RDMA is designed to operate over lossless transport, RoCE relies on congestion control mechanisms for best performance. In ethernet networks, packet losses are mainly seen due to congestion and buffer overflows. To mitigate that, efficient flow control and congestion control mechanisms are required in the network transport path. Broadcom's Ethernet Adapters (also referred to as Ethernet NICs) along with Arista Networks' switches leverage RDMA and supplement the necessary congestion control mechanisms to meet the high performance computation requirements of modern datacenters.

Features

Datacenter administrators need to configure switches and NICs in the network to provide the lossless transport RoCE requires and to prioritize RoCE traffic. The key congestion control mechanisms that the switch and NICs provide are DCQCN (Data Center Quantized Congestion Notification) and PFC (Priority Flow Control).

PFC helps to prevent packet loss due to any overflowing buffers on a switch. PFC as defined in IEEE standard 802.1Qbb is a link level flow control mechanism, supported on Arista switches and Broadcom Ethernet NIC adapters. PFC allows pausing traffic per queue/class instead of pausing all the traffic on a link during congestion. PFC helps ensure Lossless Ethernet by pausing and restarting flows depending on the congestion situation in the network. A limitation of PFC is that it restrains all traffic in the congested queue irrespective of whether it is destined for the congested path and can be subdued by DCQCN.

DCQCN is an end-to-end congestion control mechanism for RoCE. With DCQCN, Arista switches and Broadcom NICs mark IP packets using the ECN (Explicit Congestion Notification) Congestion Encountered (CE) field as the switch packet buffers cross a configured threshold. This provides an indication of congestion before buffers are full and packets are dropped. When the destination NIC receives a packet marked with ECN Congestion Encountered, it generates an RDMA Congestion Notification Packet (CNP) to the sender and thereby instantly propagates the congestion information to the source NIC indicating it to reduce the transmission rate for that flow. Broadcom NICs provide independent transmission rate control for every RDMA flow.

Arista EOS (Extensible Operating System) and Datacenter Switches

Modern AI applications need a high-bandwidth, lossless, low-latency, scalable, multi-tenant network that can interconnect hundreds and thousands of GPUs at speeds of 100Gbps, 200 Gbps, 400Gbps, 800Gbps and beyond. Arista EOS® (Extensible Operating System) provides all the necessary tools to achieve a premium lossless, high bandwidth low latency network. EOS supports traffic management configuration, adjustable buffer allocation schemes and use of PFC and DCQCN to support RoCE deployments. Without visibility into network buffer utilization, configuring appropriate PFC and ECN thresholds can be challenging. Arista EOS offers an easy solution called Latency Analyzer (LANZ) which tracks interface congestion and queuing latency with real-time reporting. This helps correlate the performance of the application with network congestion events allowing PFC and ECN values to be optimally configured to best suit the requirements of the application.

Arista Switch	Product Description
7800R Series	Highest density 100G/400G Deep Buffer, Lossless Modular Super Spine Switch
7280R Series	High Performance 10/40/100/400G Data Center switch with Dynamic Deep Buffer
7060X Series	Highest Performance and power efficient 10/40/100/400G Fixed Configuration Switch
7050X Series	Industry Leading Performance 10/40/100/400G Fixed Configuration Switch

Table 1: Arista Datacenter Switches

Broadcom Ethernet NIC Adapters

Designed for cloud scale and enterprise environments, Broadcom Ethernet Adapters are the ideal solution for high performance computing, secure datacenter connectivity and machine learning. Broadcom supports a broad portfolio of Ethernet NIC Adapters ranging from 1Gbps – 200Gbps port speeds and delivers best-in-class performance and hardware acceleration and offload capabilities that result in higher throughput, higher CPU efficiency, and lower workload latency for TCP/IP as well as RoCE traffic. Broadcom has optimized the DCQCN implementation on its NICs with two congestion control modes, DCQCN-p and DCQCN-d, DSQCN-p utilizes Probabilistic ECN marking policy, with marking probability increasing linearly as the congestion level in switch queue increases. DCQCN-d utilizes Deterministic ECN marking policy where 100% of the packets are marked when congested queue level rise above a configured threshold. DCQCN-d is optimized for applications with competing workloads as it leads to lower overall queue depths. RoCE is supported on Ethernet adapters based on BCM575xx (Thor) ASIC and the adapters support 10GE, 25GE, 100GE and 200GE. The NIC adapters are available in both OCP and PCIE form factors.

Part Number	ASIC	Ports	I/O
BCM957504-N425G	BCM57504	4x 25G	SFP28
BCM957504-N1100G	BCM57504	1x 100G	QSFP56
BCM957504-N1100GD	BCM57504	1x 100G	DSFP
BCM957508-N2100G	BCM57508	2x 100G	QSFP56
BCM957508-N2200G	BCM57508	2x 200G	QSFP56

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

Part Number	ASIC	Ports	I/O
BCM957504-P425G	BCM57504	4x 25G	SFP28
BCM957508-P2100G	BCM57508	2x 100G	QSFP56
BCM957508-P2200G	BCM57508	2x 200G	QSFP56

Table 3: Broadcom PCIE NIC Adapters with RoCE support

Summary

Designed for modern high performance compute workloads RoCE is a proven technology shown to provide greater performance and scalability for datacenter applications. With robust congestion control on the switches and NICs, Arista and Broadcom are at the forefront to meet the requirements of modern datacenter applications to have a reliable, high throughput, low latency network with low CPU overload.

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

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

www.arista.com

ARISTA