High-Performance Ethernet Networking for Artificial Intelligence Systems

Configuration & Deployment Guide

Table of contents

Table of contents

Introduction: 400G AI Networks

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 data movement between processing nodes, reliable transport is critical. As the future of metaverse applications evolves, the network needs to adapt to the humongous growth in data transfer due to data-intensive and compute-intensive applications. Broadcom's Ethernet Adapters (also referred to as Ethernet NICs) along with Arista Networks' switches (based on Broadcom's DNX and XGS family of ASICs) leverage RDMA (Remote Direct Memory Access) to eliminate any connectivity bottlenecks and facilitate a high-throughput, low-latency transport.

Server Architecture Basics

Before digging into the details of how to maximize the network performance, it is critical to understand the server and network architecture basics. The diagram below shows a very high-level architecture and key components of a standard AI server. The highest performance requirements in a network are typically in the "Scale-Out Fabric" portion of the server/network shown below and this is the focus of this Configuration and Deployment Guide.

Front-End fabric: The front-end fabric is the network infrastructure interconnecting different functions such as applications, storage, and in-band GPU cluster management. Based on performance needs, this fabric generally only requires one or two NICs per server.

Internal-AI fabric: The internal-AI fabric integrates the key hardware components, including the GPUs, CPUs, NIC, and storage. This is achieved through the use of PCIe switches.

Scale-up fabric: The scale-up fabric is for intra-note GPU-to-GPU communication (typically within the server only). This fabric is created using native interfaces of the GPU including AMD's Infinity Fabric, Ethernet, and Nvidia's NVLink.

Scale-out fabric: The scale-out fabric is the fabric used to interconnect AI servers to create clusters. This fabric is essential for distributed workloads required by AI models and requires a high-bandwidth, low-latency network. The scale-out fabric performance requirements typically demand one NIC per GPU with matching PCIe bandwidths (e.g. PCIe Gen5 x16 or 400Gbps). The NIC and GPU communicate through the PCIe switch using a highly efficient peer-to-peer protocol, giving the GPU non-blocking, low-latency access to the scale-out network, allowing fast data transfers to other GPUs in the cluster.

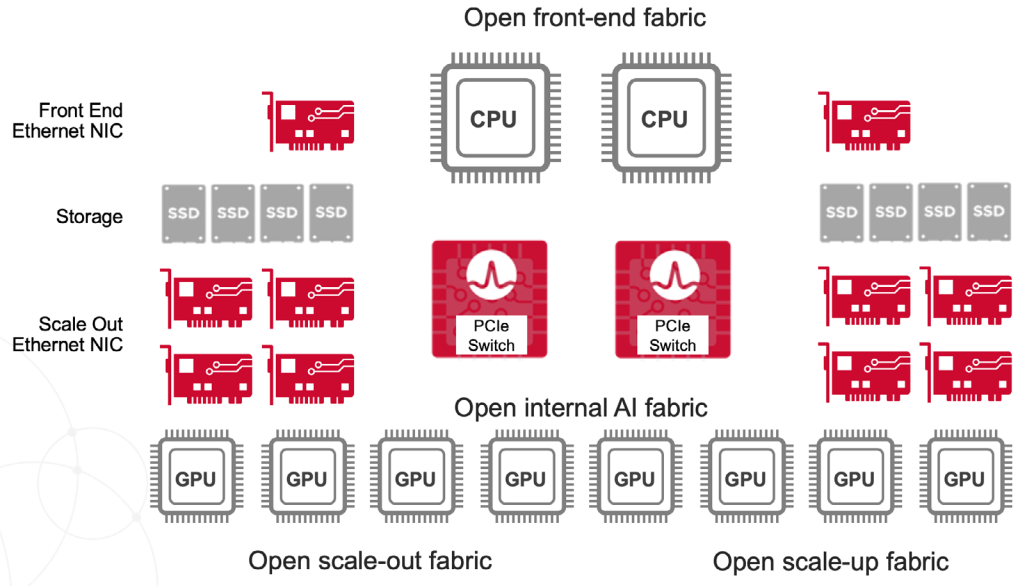

Figure 1: Basic AI server architecture

Network architectures for AI

As AI networking evolves, so do the network switch deployments. Arista supports all the network deployment configurations, e.g, CLOS including multi-stage, Planar (Rail Based).

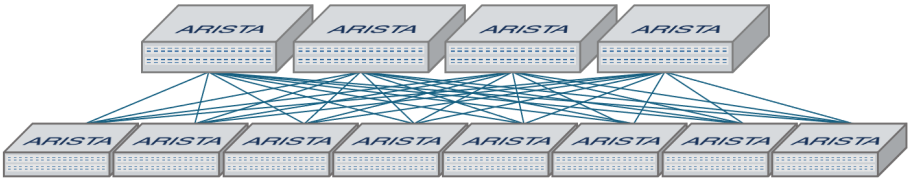

Figure 2: Arista switches in Tiered Leaf Spine/Plane based design

CLOS Topology: **Multi-stage architecture consists** of multiple stages of switches, providing equal-cost multi-paths for the traffic to flow.

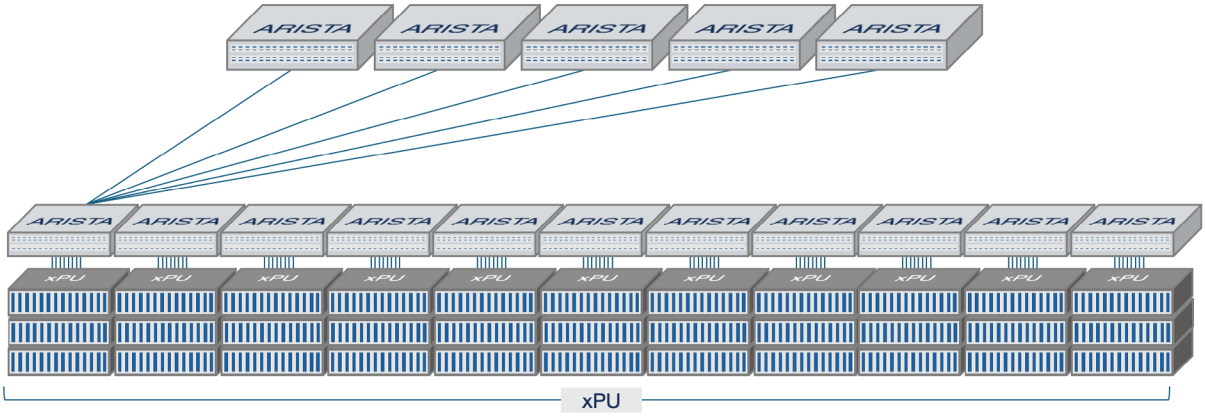

Figure 3: Leaf Spine Topology

Planar (Rail) Based topology: **Single-stage Architecture**, typically consists of a single layer of switches or routers connected in a linear or ring-like fashion. The single layers are interconnected using a higher layer switch.

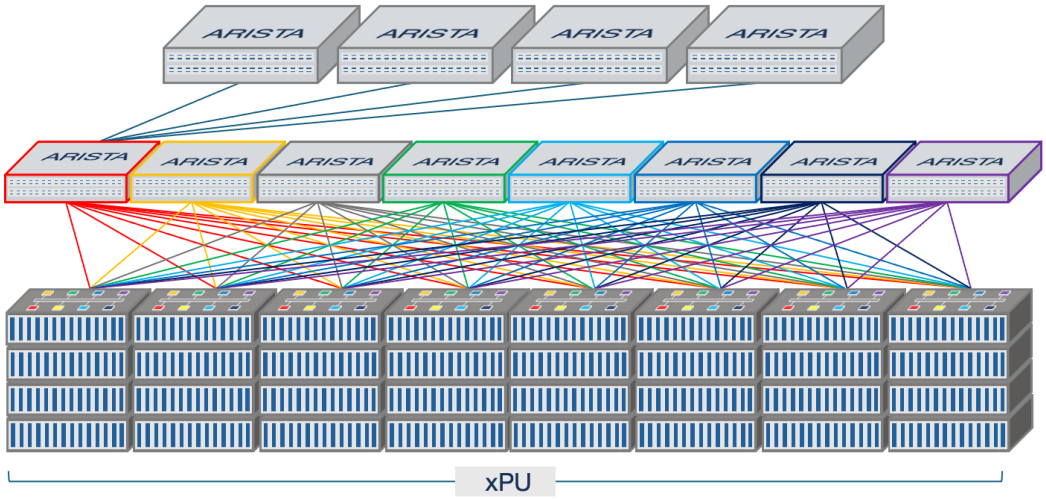

Figure 4:Planar/Rail-based Topology

RDMA Over Converged Ethernet (RoCE)

RoCE (RDMA over Converged Ethernet) is a network protocol that allows RDMA over an Ethernet network. RDMA helps to reduce the CPU workload as it offloads all transport communication tasks from the CPU to hardware and provides direct memory access for applications without involving the CPU. The second version of RoCE (RoCE-v2) enhances the protocol with a UDP/IP header and enables a routable RoCE. Broadcom's Ethernet Adapters support RoCEv2 in hardware and allow for higher throughput, lower latency, and lower CPU utilization, which are critical for AI/ML, Storage, and High-Performance compute (HPC) applications.

RoCEv2 provides three advantages:

- Operation on routed ethernet networks, ubiquitous in large data centers
- IP QoS – The DiffServ code point (DSCP), or alternatively VLAN PRI
- IP congestion control – The explicit congestion notification (ECN) signal

To eliminate potential packet loss and high latency on Ethernet networks, RoCEv2 uses congestion control mechanisms supported on Arista switches and Broadcom NICs such as Priority Flow Control (PFC), Explicit Congestion Notification (ECN), etc.

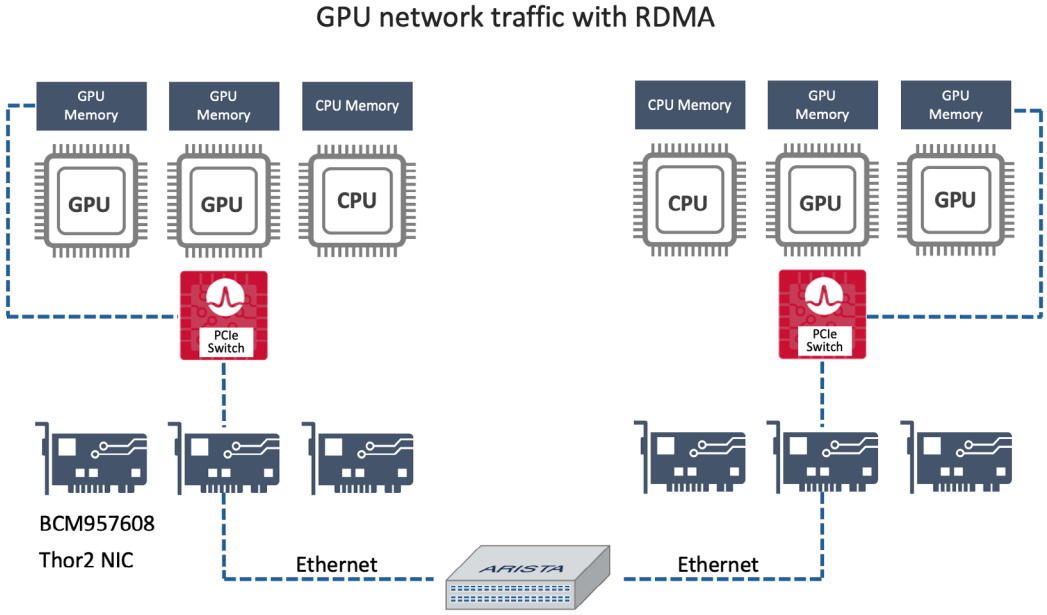

Figure 5: RoCE with Broadcom NICs and Arista Switch

RoCEv2 also defines a Congestion Notification Packet (CNP). RNICs send CNPs in response to ECN Congestion Experienced (CE) markings to indicate that the transmission rate should be reduced. ECN marking is done by switches along the path between the source and destination or by the receiving NIC. CNPs are associated with RoCE connections, providing fine-grained, per-connection congestion notification information. RoCEv2 only specifies the mechanism for marking packets when congestion is experienced and the format of the CNP response. It leaves the implementation of the congestion control algorithm unspecified, including the following information:

- When packets are ECN marked (at which queue level, and at what probability)
- When CNPs are generated in response to ECN
- How the sending rate is adjusted in response to CNP

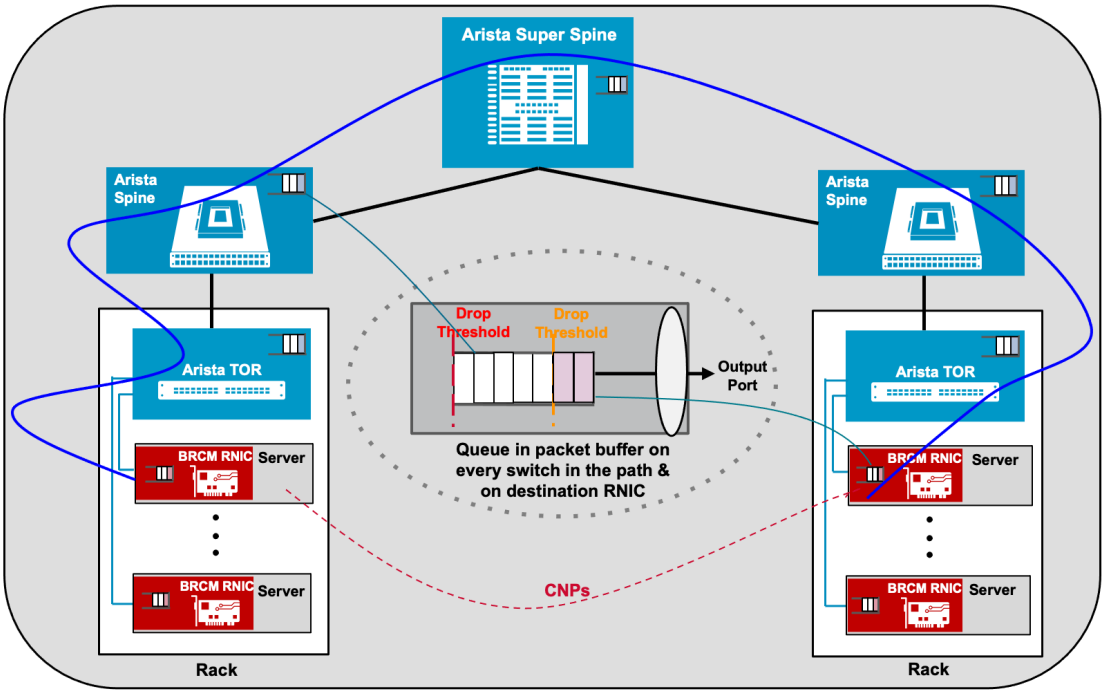

Figure 6: RoCE traffic in datacenter network

Deploying RoCE on Arista Switches

Arista Extensible Operating System (EOS®) is the core of Arista cloud networking solutions for next-generation data centers and high-performance computing networks. Arista EOS 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 the use of PFC and DCQCN to support RoCE deployments across the Arista 7280R series, 7800R series, 7050X series, and 7060X series. The exact Arista Switch to be used depends on the specific use case.

Table 1: Arista Datacenter Switches for RoCE use cases

Arista Portfolio	Description
7060X5 and 7060X6	High Density 800G Fixed Switch Portfolio for AI and DC
7280R3 and 7800R3	High Performance 400G and 800G Dynamic Deep Buffer Platforms
7700R4	800G Distributed Ethernet Switch for Accelerated Computing

The Arista 7280R and 7800R series are based on the Broadcom Jericho chipset families. Equipped with deep buffers and Virtual Output Queueing scheduling mechanisms, these ensure lossless transmission of end-to-end data. The 7280R series is the fixed configuration family of switches, while the 7800R series is the modular line of switches.

The Arista 7050X and 7060X series are based on the Broadcom Trident and Tomahawk chipset families respectively. Supporting rich feature support and low latency, the 7050X and 7060X series are perfect for highly efficient and robust deployments.

General installation and configuration of Arista switches is available here.

Once end-to-end network connectivity is established, Priority Flow Control (PFC) or Explicit Congestion Notification (ECN) can be enabled to ensure lossless transport for RoCE traffic.

Configuration and Verification of PFC on Arista Switches

PFC is one of the most important aspects of successful RoCE deployments. PFC specifies a link-layer flow control mechanism between directly connected peers. It uses the 802.3 PAUSE frames to implement flow-control measures for multiple classes of traffic. Switches can drop the less important traffic and notify the peer devices to pause traffic on specific classes so that critical data is not dropped and allowed to pass through the same port without any restrictions.

This Quality of Service (QoS) capability allows differentiated treatment of traffic based on the CoS/priority and eases congestion by ensuring that critical I/O is not disrupted and that other non-critical traffic that is more loss-tolerant can be dropped. Each priority is configured as either drop or no drop. If a priority that is designated as no-drop is congested, the priority is paused. Drop priorities do not participate in pause.

PFC Configuration

The CLI command to enable PFC on the interface is "priority-flow-control mode on" and "priority-flow-control priority <TC> no-drop" enables PFC on that Transmit Queue.

- Enable PFC on the interface.

```
arista(config)#interface ethernet 3/1/1
arista(config-if-Et3/1/1)#priority-flow-control mode on
```

- Enable PFC for specific TCs.

```
arista(config-if-Et3/1/1)#priority-flow-control priority 0 no-drop
```

The above command should be issued for all the TC's that the user wants to enable PFC on.

Example Configuration

The following configuration shows how PFC can be configured for TC3 and TC4 under interface Ethernet 2/1/1 on an Arista switch.

```
interface Ethernet2/1/1
mtu 9000
speed 200G-4
no switchport
priority-flow-control mode on
priority-flow-control priority 3 no-drop
priority-flow-control priority 4 no-drop
!
```

Show Commands

1. show priority-flow-control interfaces ethernet <>

```
arista#show priority-flow-control interfaces ethernet 3/1/1
The hardware supports PFC on priorities 0 1 2 3 4 5 6 7
PFC receive processing is enabled on priorities 0 1 2 3 4 5 6 7
The PFC watchdog timeout is 1.0 second(s)
The PFC watchdog recovery-time is 2.0 second(s) (auto)
The PFC watchdog polling-interval is 0.2 second(s)
The PFC watchdog action is drop
The PFC watchdog override action drop is false Global PFC : Disabled
E: PFC Enabled, D: PFC Disabled, A: PFC Active, W: PFC Watchdog Active
Port       Status Priorities Action      Timeout Recovery   Polling      Note
                                         Interval/Mode      Config/Oper
-------------------------------------------------------------------------------
Et3/1/1        E - -  01       -            -   - / -          - /-            DCBX
```

```
   disabled
   Port              RxPfc               TxPfc
   Et3/1/1            0                   0
2. show priority-flow-control interfaces ethernet < > counters
   arista#show priority-flow-control interfaces ethernet 3/1/1
   counters

   Port    RxPfc    TxPfc
   Et3/1/1   0        0
3. show priority-flow-control interface ethernet < > status
   arista#show priority-flow-control interfaces ethernet 3/1/1
   status
   The hardware supports PFC on priorities 0 1 2 3 4 5 6 7
   PFC receive processing is enabled on priorities 0 1 2 3 4 5 6 7
   The PFC watchdog timeout is 1.0 second(s)
   The PFC watchdog recovery-time is 2.0 second(s) (auto)
   The PFC watchdog polling-interval is 0.2 second(s)
   The PFC watchdog action is drop
   The PFC watchdog override action drop is false
   Global PFC : Disabled
   E: PFC Enabled, D: PFC Disabled, A: PFC Active, W: PFC Watchdog Active
   Port      Status Priorities Action      Timeout Recovery      Polling      Note
                                           Interval/Mode         Config/Oper

   ------------------------------------------------------------------------------------
   Et3/1/1     E - -    01       -      -      - / -         - / -       DCBX
   Disabled
```

Configuration and Verification of ECN on Arista Switches

Explicit Congestion Notification (ECN) is an extension to TCP/IP that provides end-to-end notification of impending network congestion prior to loss. Two Bits (bit 0 and bit 1) in the ToS byte of the IP header are used for ECN. That is, ECN bits in the ToS byte define a packet in 4 different ways:

- 00 - (default) indicates packet is non-ECN capable
- 01 - indicates packet is ECN capable
- 10 - indicates packet is ECN capable
- 11 - indicates Congestion Occurred somewhere in the network

ECN is an optional feature that is only used when both endpoints support it. ECN should be considered complementary to PFC for lossless network behavior and is therefore an integral component of RoCE. ECN bits are marked on traffic in certain classes when the configured buffer thresholds are exceeded.

ECN operates over an active queue management (AQM) algorithm - Weighted Random Early Detection (WRED) to detect congestion on the network device and mark ECN-capable traffic with ECN flag.

Note: ECN is only used when both endpoints support it and are willing to use it. Packets are ECN marked based on WRED as follows:

- If the average queue size (ie. the number of packets in the queue) is below the minimum threshold, packets are queued as in normal operation without ECN.
- If the average queue size is greater than the maximum threshold, packets are marked for congestion.
- If the average queue size is between the minimum and the maximum queue threshold, packets are either queued or marked. The proportion of packets that are marked increases linearly from 0% at the minimum threshold to 100% at the maximum threshold.

ECN Configuration

- ECN is configured at the egress Tx-Queue of an Interface

```
arista(config)#interface ethernet 6/1/1
arista(config-if-Et6/1/1)#tx-queue 6
# Example of probabilistic threshold
arista(config-if-Et6/1/1-txq-6)#random-detect ecn minimum-threshold 500 kbytes maximum-threshold 1500 kbytes max-mark-probability 25
# Example of deterministic threshold
arista(config-if-Et6/1/1-txq-6)#random-detect ecn minimum-threshold 256 kbytes maximum-threshold 256 kbytes max-mark-probability 100
```

- Enable ECN Counters under the Tx-Queue

```
arista(config-if-Et6/1/1-txq-6)#random-detect ecn count
```

- Enable ECN Counter feature in Hardware

```
arista(config)#hardware counter feature ecn out
arista(config)#show hardware counter feature | grep -i ECN
ECN       out    Jericho2: 1  up
```

- On DCS-7280R, DCS-7280R2, DCS-7500R, DCS-7500R2, DCS-7280R3, DCS-7500R3 and DCS-7800R3, the following CLI is required to allocate counter resources for ECN counters.

```
arista(config)# [no | default] hardware counter feature ecn out
```

PFC Watchdog

Priority Flow Control (PFC) Watchdog feature monitors the switch interfaces for priority-flow-control pause storms. If such storms are detected, it takes actions such as:

- Disable reacting to received Pause frames
- Stop sending packets to these interfaces and drop any incoming packets from these interfaces.

PFC Pause storm reception is usually an indication of a misbehaving node downstream, and propagating this congestion upstream is not desired. Note that the granularity of monitoring is per port and per priority.

Configuration

Configure the interval to poll the queues:

```
arista(config)# priority-flow-control pause watchdog default polling-interval ?
<0.1 - 30> Polling interval in seconds
```

Configure the interval after which the port should start dropping packets on congested priorities:

```
arista(config)# priority-flow-control pause watchdog default timeout ?
<0.2-60> Timeout value in seconds
```

Configure the interval after which stuck ports, and priorities when clear of PFC Pause storm should recover and start forwarding:

```
arista(config)# priority-flow-control pause watchdog default recovery-time ?
```

```
<0.2-60> Recovery time in seconds
```

Configure the PFC Watchdog action to be drop:

```
arista(config)# priority-flow-control pause watchdog action drop
```

If the drop action is not configured, the default action is to stop reacting to PFC Pause frames received on the (port, priority) experiencing the PFC Pause storm.

Show commands

```
# show priority-flow-control counters watchdog
Port   TxQ    Total times  Total times
              stuck        recovered
---------------------------------------
Et3    UC1    6      6
```

Deploying RoCE on Broadcom Ethernet NIC Adapters

Designed for cloud scale and enterprise environments, Broadcom Ethernet NIC Adapters are the ideal solution for network connectivity for high-performance computing, secure data center connectivity, and AI/ML applications. Broadcom supports a broad portfolio of Ethernet NIC Adapters ranging from 1Gbps – 400Gbps port speeds and delivers best-in-class performance, 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. RoCE is supported on Ethernet adapters based on BCM576xx (Thor2) ASIC and the adapters support 400GE speeds. The Broadcom Ethernet NIC adapters with RoCE support are available in both OCP and PCIE form factors and are summarized in Table 2 and Table 3 below.

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

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

Table 3: Broadcom PCIE NIC Adapters with RoCE support

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

RoCE (RDMA over converged Ethernet) is a complete hardware offload feature supported on Broadcom Ethernet NIC controllers, which allows RDMA functionality over an Ethernet network. RoCE helps to reduce CPU workload as it provides direct memory access for applications, bypassing the CPU.

RoCE Congestion Control on Broadcom Ethernet NIC Adapters

Broadcom Ethernet NIC adapters support two congestion control (CC) modes, DCQCN-Pand DCQCN-D, where DCQCN-P utilizes Probabilistic ECN marking policy, with marking probability increasing linearly within a range of congested queue levels, while DCQCN-D utilizes Deterministic ECN marking policy as in DCTCP where 100% of the packets are marked when congested queue level rises above a configured threshold.

In both modes, the NIC performs very similar operations and utilizes the same infrastructure to control the rate of each flow (Queue Pair, or QP, in RoCE terminology). However since the number of ECN marked packets and hence CNPs differ, the computation of congestion level is different.

In DCQCN-P there are fewer CNPs than in DCQCN-D since when the congested queue level starts to rise, only a small percentage of packets traversing the switch are ECN marked. Some of the flows that do receive CNPs reduce their rate while others do not. If congestion persists, a higher percentage of packets are marked, and more flows possibly receive a signal from the network and reduce their rate. Thus, when there are many competing flows, the congested queue level may rise to a higher level until stabilizing in comparison with DCQCN-D. On the other hand, since there are more CNPs with DCQCN-D, there is a higher load on the NIC in processing the stream of CNPs and accessing the associated flow context.

The CC algorithm in Broadcom Ethernet NIC adapters has been enhanced relative to the original DCQN paper due to several issues in the original algorithm. For more details, refer to the congestion control for RoCE whitepaper for Broadcom Ethernet adapters.

Installation Guide for Broadcom Ethernet NIC Adapters

Broadcom Ethernet User Guide, available publicly, provides detailed instructions on how to install RoCE on Broadcom Ethernet Network Adapters.

Updating the Firmware on Broadcom Ethernet NIC Adapters

The following niccli command is used to update the adapter firmware on Broadcom Ethernet NIC adapters. Note that the niccli command requires sudo or root access.

```
sudo niccli –i <index> install <firmware package>
```

Example:

```
sudo niccli –i 1 install BCM957608-P1400G.pkg
```

Configuring NVRAM

To update the NVRAM configuration, use the niccli utility provided with the release. Run niccli version to check the version you are using.

- Ensure that RDMA is enabled for the specific PF.
- For RoCE performance, the performance profile NVM CFG must be set to RoCE (value 1). NOTE: A host reboot is required for the new settings to take effect.

Verify the RDMA and performance settings with the following commands:

```
sudo niccli –i 1 nvm –getoption support_rdma –scope 0
sudo niccli –i 1 nvm –getoption performance_profile
```

The output value for the support_rdma parameter should read Enabled and the value for performance_profile should read RoCE.

To enable RDMA for the specific PF and to set the performance profile to RoCE, use the following commands:

```
sudo niccli –i 1 nvm –setoption support_rdma –scope 0 –value 1
sudo niccli –i 1 nvm –setoption performance_profile –value 1
```

Reboot the system after setting the NVRAM options.

Host Requirements for Driver/Library Compilation

Compiling the driver and library has dependencies on build packages such as automake, libtool, make, gcc, and so forth. The following packages are recommended based on the OS distribution being used.

» CentOS/Redhat/Fedora

See the following commands for CentOS, Redhat, and Fedora operating systems:

```
dnf group install “Development Tools”
dnf group install “Infiniband Support”
yum install –y libibverbs-devel qperf perftest infiniband-diags make gcc kernel kernel-devel
autoconf aclocal libtool libibverbs-utils rdma-core-devel
```

» Ubuntu/Debian

See the following commands for Ubuntu or Debian operating systems:

```
apt install –y automake autoconf libtool bc bison build-essential flex libibverbs-dev
ibverbs-utils infiniband-diags perftest ethtool
```

Installing the Layer 2 and RoCE Driver

This section describes how to install the Layer 2 communication (L2) and RoCE driver. The installation tarball contains the netxtreme-bnxt_en-<version>.tar.gz file. This file includes both the L2 and RoCE drivers. Install the drivers using the following commands:

```
Go to the directory in the release
cd <release>/drivers_linux/bundle
BRCM_DRIVER_VERSION=1.10.3-232.1.132.0
tar xvf netxtreme-bnxt_en-${BRCM_DRIVER_VERSION}.tar.gz
cd netxtreme-bnxt_en-${BRCM_DRIVER_VERSION}
make
sudo make install
sudo depmod -a
```

Updating Initramfs

Most Linux distributions use a ramdisk image to store drivers for boot-up. These kernel modules take precedence, so the initramfs must be updated after installing the new bnxt_en/bnxt_re modules. For CentOS, Redhat, and Fedora operating systems, use sudo dracut -f and for Ubuntu/Debian operating systems use `sudo update -initramfs -u.`

Installing the RoCE Library

This section describes how to install the RoCE library. The installation tarball contains the libbnxt_re- <version>.tar.gz file. This file includes the libbnxt_re RoCE library.

Execute the following steps.

1. To avoid potential conflicting library files, remove or rename the libbnxt RoCE library from the Linux distribution using the following command. The command is a single command and tries to locate the inbox libbnxt_re library in one of the previous directories. It may be necessary to run it as a sudo user.
   ```
   find /usr/lib64 /usr/lib /lib64 -name "libbnxt_re-rdmav*.so" -exec mv {} {}.inbox \;
   ```
2. Build and install the userspace RDMA library from the source using the following commands. See Host Requirements for Driver/Library Compilation for information regarding host package dependencies that are required for building the RoCE library from the source. Note that the portion of the command that is dark red below is release-specific.
   ```
   cd <release>/drivers_linux/bnxt_rocelib
   BRCM_LIB_VERSION=232.1.132.0
   tar xvf libbnxt_re-${BRCM_LIB_VERSION}.tar.gz
   cd libbnxt_re-${BRCM_LIB_VERSION}
   sh autogen.sh
   ./configure --sysconfdir=/etc
   make
   sudo make install all
   sudo sh -c "echo /usr/local/lib >> /etc/ld.so.conf"
   sudo ldconfig
   sudo cp bnxt_re.driver /etc/libibverbs.d/
   ```

3. Record the md5sum of the library that was built to verify that the correct library is running using the following command.

```
find . -name "*.so" -exec md5sum {} \;
```

4. Use the following commands to identify the path of the libbnxt_re library being used on the host and then calculate its md5sum. The md5sum should match the md5sum of the built libraries in the previous step.

```
strace ibv_devinfo 2>&1 | grep libbnxt_re | grep -v 'No such file'
md5sum <path of the lbbnxt_re library> shown by the last command
```

Validating the RoCE Installation

After the drivers and libraries are installed, perform the following steps to validate RoCE installation.

Confirming the GUID for the RoCE Interface

The node GUID indicates that RoCE has been successfully configured on the system. There are two commands that can be used to confirm the GUID for the RoCE interface:

- ibv_devices – indicates if the GUID is available.
- ibv_devinfo – indicates if the GUID is available and provides additional details about the RoCE interface.

```
#ibv_devices
device node GUID
bnxt_re0 be97e1ffeda96d0

# ibv_devinfo
hca_id: bnxt_re0
transport: InfiniBand (0)
fw_ver: 232.1.132.0
node_guid: d604:e6ff:fe7e:5bbc
sys_image_guid: d604:e6ff:fe7e:5bbc
vendor_id: 0x14e4
vendor_part_id: 5984
hw_ver: 0x11
phys_port_cnt: 1
port: 1
state: PORT_ACTIVE (4)
max_mtu: 4096 (5)
active_mtu: 4096 (5)
sm_lid: 0
port_lid: 0
port_lmc: 0x00
link_layer: Ethernet
```

Installing RoCE QOS Configuration Package (bnxt_re_conf)

RoCE QOS config package (bnxt_re_conf) sets up udev rules and scripts for configuring RoCE QoS parameters – PFC, CC, RoCE and CNP DSCP values, and so forth. Bnxt_re_conf involves an udev rule(90-bnxt_re.rules) that is triggered when the bnxt_re device is loaded and it invokes a wrapper script(bnxt_re_conf.sh) that would take values of the required parameters from a configuration file and will run the configuration using bnxt_setupcc.sh.

The bnxt_re_conf pkg is distributed in a variety of formats (debian, RPM, and source tarball). Depending on the OS distro being used, the appropriate pkg format can be used.

```
cd <release>/drivers_linux/bnxt_re/bnxt_re_conf
dpkg -i bnxt_re_conf_232.0.155.5-1_all.deb
```

or

```
rpm -Uvh bnxt_re_conf-232.0.155.5-1.noarch.rpm
```

It is recommended to reboot the host if the bnxt_re_conf pkg is installed the first time or the contents of the /etc/bnxt_re/bnxt_re.conf file are modified. This allows the RoCE QOS settings to be applied correctly to each NIC.

After installing the bnxt_re_conf pkg, ensure the pkg is correctly installed by checking for the presence of the /etc/bnxt_re/bnxt_re.conf file. The file contents should match the configured QOS values. The default values are shown as follows:

```
# cat /etc/bnxt_re/bnxt_re.conf
ENABLE_FC=1
FC_MODE=3
ROCE_PRI=3
ROCE_DSCP=26
CNP_PRI=7
CNP_DSCP=48
ROCE_BW=50
UTILITY=3
```

The correct RoCE QOS application on each RoCE NIC can be verified via the following Broadcom-provided NICCLI tool command:

```
# sudo niccli -i 1 getqos
IEEE 8021QAZ ETS Configuration TLV:
PRIO_MAP: 0:0 1:0 2:0 3:1 4:0 5:0 6:0 7:2
TC Bandwidth: 50% 50% 0%
TSA_MAP: 0:ets 1:ets 2:strict
IEEE 8021QAZ PFC TLV:
PFC enabled: 3
IEEE 8021QAZ APP TLV:
APP#0:
Priority: 7
Sel: 5
DSCP: 48
APP#1:
Priority: 3
Sel: 5
DSCP: 26
APP#2:
Priority: 3
Sel: 3
UDP or DCCP: 4791
TC Rate Limit: 100% 100% 100% 0% 0% 0% 0% 0%
```

When the NIC is configured for RoCE on a host along with the RoCE QOS configuration pkg (bnxt_re_conf), the NIC is automatically configured for DCQCN-P along with the following settings.

- RoCE v2 packets are marked with a DSCP value of 26 and use Priority 3 internally
- CNP packets are marked with a DSCP value of 48 and use Priority 7 internally
- PFC is enabled for Priority 3 traffic

Confirm Traffic Flow to the Remote RoCE Endpoint

If the RoCE endpoint is currently configured, traffic flow can be verified by using one of the perftest package utilities.

Command usage:

```
ib_write_bw  -d bnxt_re0 -q 2 -F --report_gbits
 ---------------------------------------------------------------------------------------
 RDMA_Write BW Test
 Dual-port       : OFF Device         : bnxt_re0
 Number of qps   : 2 Transport type : IB
 Connection type : RC Using SRQ      : OFF
 PCIe relax order: ON
 ibv_wr* API     : OFF
 CQ Moderation   : 1
 Mtu             : 4096[B]
 Link type       : Ethernet
 GID index       : 3
 Max inline data : 0[B]
 rdma_cm QPs : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0x2c04 PSN 0x9b1afe RKey 0x4000002 VAddr 0x007fa691285000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:01:10
 local address: LID 0000 QPN 0x2c03 PSN 0x3a1c10 RKey 0x4000002 VAddr 0x007fa691295000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:01:10
 remote address: LID 0000 QPN 0x2c04 PSN 0x3ef2c9 RKey 0x4000002 VAddr 0x007f7ecc568000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:01:20
 remote address: LID 0000 QPN 0x2c03 PSN 0x15f7f RKey 0x4000002 VAddr 0x007f7ecc578000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:01:20
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
 65536      4486898          0.00               392.07   0.747809
---------------------------------------------------------------------------------------
```

Configuring Priority Flow Control on Broadcom NICs

Broadcom's RoCE driver enables three traffic classes (L2, RoCE, and Congestion Notification Packet (CNP)). Loading the driver automatically sets up the default RoCE/CNP Priority Code Point (PCP) priorities and Differentiated Services Code Point (DSCP) values. Priority Flow Control (PFC) and Congestion Control (CC) are enabled by default and the default values are listed in Table 4. No other configuration is required on the host if the upstream switches are configured with these default values.

Table 4: Broadcom NIC - Default PCP and DSCP Values

Traffic Type	Default PCP value	Default DSCP value
RoCE	3	26
CNP	7	48

The default traffic classes are:

- TC0 (L2 Traffic)
- TC1 (RoCE Traffic)
- TC2 (CNP Traffic)

In the absence of L2 traffic, the full bandwidth is allotted for RoCE traffic.

To change the default values of PCP and DSCP to match the user's network settings, the parameters in /etc/bnxt_re/bnxt_re.conf can be changed.

For example if we need to change the below QOS settings on the NIC, the below parameters in /etc/bnxt_re/bnxt_re.conf need to be changed

- RoCE PCP priority 5 and DSCP value 32
- CNP PCP priority 6 and DSCP value 36

Modify the file cat /etc/bnxt_re/bnxt_re.conf with the values below:

```
ROCE_PRI=5
ROCE_DSCP=32
CNP_PRI=6
CNP_DSCP=36
```

After making the above changes reboot the host for the changes to take into effect. The changes made by this method are persistent across reboots

Alternatively, the above parameters can also be changed by using the below command using bnxt_setupcc.sh:

```
sudo bnxt_setupcc.sh -d <x> -i <RoCE interface> -m <x> -s <RoCE DSCP value> -p <CNP DSCP value> -r <RoCE PCP value> -c <CNP DSCP value>
```

Example:

```
sudo bnxt_setupcc.sh -d bnxt_re0 -i ens4f0np0 -m 3 -s 32 -p 36 -r 5 -c 6 -u 3
```

The changes made by this method are not persistent across reboots.

To check the configuration, run the following commands:

#sudo niccli -i 1 getqos

```
---------------------------------------------------------------------------

NIC CLI v232.0.148.0 - Broadcom Inc. (c) 2024 (Bld-106.52.39.138.16.0)

---------------------------------------------------------------------------

IEEE 8021QAZ ETS Configuration TLV:
         PRIO_MAP: 0:0 1:0 2:0 3:0 4:0 5:1 6:2 7:0
         TC Bandwidth: 50% 50% 0%
         TSA_MAP: 0:ets 1:ets 2:strict
IEEE 8021QAZ PFC TLV:
         PFC enabled: 5
IEEE 8021QAZ APP TLV:
   APP#0:
    Priority: 6
    Sel: 5
    DSCP: 36
```

APP#1:

Priority: 5

Sel: 5

DSCP: 32

APP#2:

Priority: 5

Sel: 3

UDP or DCCP: 4791

TC Rate Limit: 100% 100% 100% 0% 0% 0% 0% 0%

#sudo niccli -i 1 dscp2prio

NIC CLI v232.0.148.0 - Broadcom Inc. (c) 2024 (Bld-106.52.39.138.16.0)

---

dscp2prio mapping:

priority:5

dscp:32

priority:6

dscp:36

#sudo niccli -i 1 listmap -pri2cos

NIC CLI v232.0.148.0 - Broadcom Inc. (c) 2024 (Bld-106.52.39.138.16.0)

Base Queue is 0 for port 0

Priority TC Queue ID

---

0 0 4

1 0 4

2 0 4

3 0 4

4 0 4

5 1 0

6 2 5

7 0 4

Configuring Congestion Control on Broadcom NICs

To adjust the congestion control parameter, the Broadcom RoCE driver relies on the kernel configfs. The default congestion control algorithm is DCQCN-P. To change to DCQCN-D algorithm, use the following procedure.

NOTE: Switch ECN values will need to be modified. Option -P uses a probabilistic marking while -D needs to modify the switch to use deterministic (100%) marking. For example

- -P: 500/1500/25%
- -D: 256/256/100%
- Configuring DCQCN-D
  To configure DCQCN-D, use the following commands:

```
mkdir -p /sys/kernel/config/bnxt_re/bnxt_re0
cd
/sys/kernel/config/bnxt_re/bnxt_re0/ports/1/cc/
echo -n 0 > cc_mode
echo -n 1 > apply
```

- Configuring DCQCN-P
  To configure DCQCN-P, use the following commands:

```
mkdir -p /sys/kernel/config/bnxt_re/bnxt_re0
cd /sys/kernel/config/bnxt_re/bnxt_re0/ports/1/cc/
echo -n 1 > cc_mode
echo -n 1 > apply
```

- Viewing the Current Congestion Control Parameters
  To view the currently configured congestion control parameters, use the following commands:

```
mkdir -p /sys/kernel/config/bnxt_re/bnxt_re0
cd
/sys/kernel/config/bnxt_re/bnxt_re0/ports/1/cc/
echo -n 1 > advanced
echo -n 1 > apply
cat apply
```

RoCE Performance Data

RoCE Performance Data Measurement Configuration

For measuring the 400GE performance data, 4 Arista switches and 64 nodes are used.

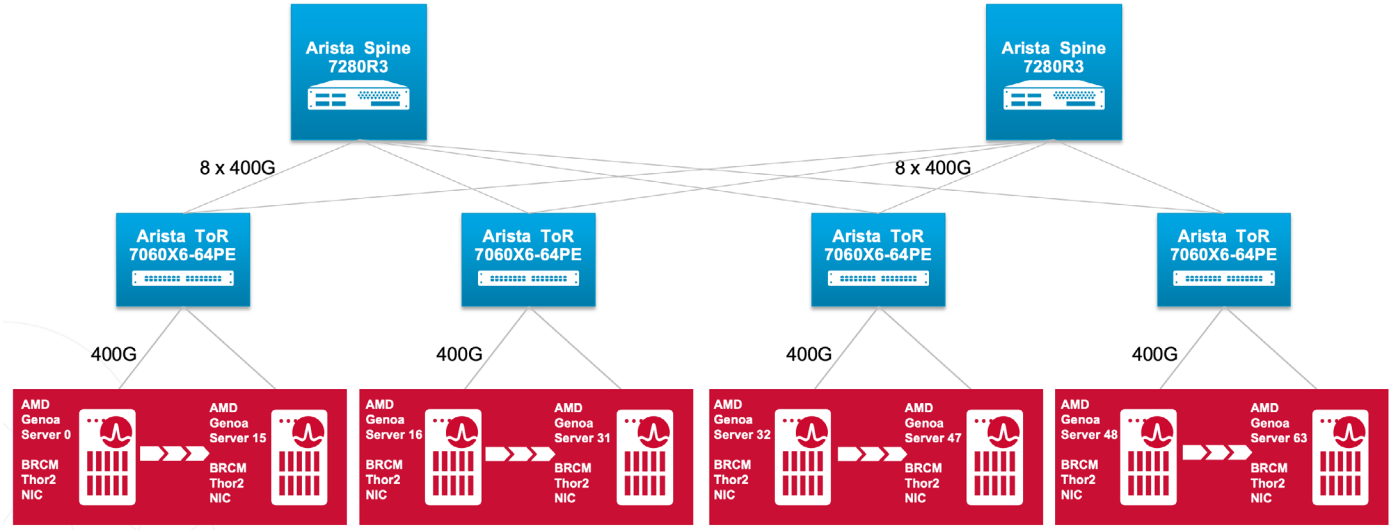

Figure 7: Roce Setup

For measuring performance numbers on the cluster with Broadcom NICs (400GE) and Arista switches (200GE/400GE) for this user guide, OSU MPI Benchmark and GPCNet Benchmark tests have been used. The cluster configuration used for the benchmark tests is captured in Table 5 below.

Table 5: Cluster Configuration for Performance Tests on 400GE

Server	Switch	NIC	Benchmarks
Model: HPE ProLiant DL385 Gen11 CPU: AMD(R) EPYC 9534 64-Core Processor CPU @ 2.6045GHz Thread(s) per core: 1 Core(s) per socket: 64 Memory Type: DDR5 -4800 MT/s Memory: 384 GB (192GB/Socket) Kernel: 5.14.0-284.30.1.el9_2.x86_64(Red Hat Enterprise Linux release 9.2 (Plow))	Model: Arista DCS-7060X6-64PE(TH5 TOR) Hardware Revision: 11.20 Software Version: 4.33.2F	Model: Broadcom P2200G Driver Version: 232.0.144.0 Firmware Version: 232.0.145.0 Congestion Control: DCQCN-p	UCX: 1.15.0 OpenMPI: 4.1.6 OSU: 7.3 LAMMPS: lammps-2Aug2023 HPCG: 3.1

RoCE Performance Data Overview

For measuring performance numbers on the cluster with Broadcom NICs (400GE) and Arista switches (200GE/400GE) for this user guide, OSU MPI Benchmark and GPCNet Benchmark tests have been used. The cluster configuration used for the benchmark tests is captured in Table 5 below.

Table 6: Broadcom NIC / Arista switch ROCE Performance on 400GE

Category	Test	Benchmarks
OSU Baseline Benchmarks	Throughput - osu_mbw_mr, 64KB, 2 nodes, 64 PPN, aggregate BW	792 Gbps
	Unload latency - osu_lat, 2B, 2 nodes (single switch)	3.6 us
OSU Collectives Benchmarks	osu_alltoall latency, 64 nodes, 64 PPN, 64K message (blocking)	475 ms
	osu_allreduce latency, 64 nodes, 64PPN, 64K message (blocking)	251 us
LAMMPS	Large-scale Atomic/Molecular Parallel Simulator	1.86E-09
	Rhodo Scaled Benchmark - OpenMPU CPU/atom/steps	
	Chain Scaled Benchmark - OpenMPU CPU/atom/steps	8.42E-11
HPCG	High-Performance Conjugate Gradient GFLOP/s	5,473

OSU MPI Multiple Bandwidth / Message Rate (osu_mbw_wr) Test

The focus of the multi-pair bandwidth and message rate test is to evaluate the aggregate uni-directional bandwidth and message rate between multiple pairs of processes. Each of the sending processes sends a fixed number of messages (the window size) back-to-back to the paired receiving process before waiting for a reply from the receiver. This process is repeated for several iterations. The objective of this benchmark is to determine the achieved bandwidth and message rate from one node to another node with a configurable number of processes running on each node.

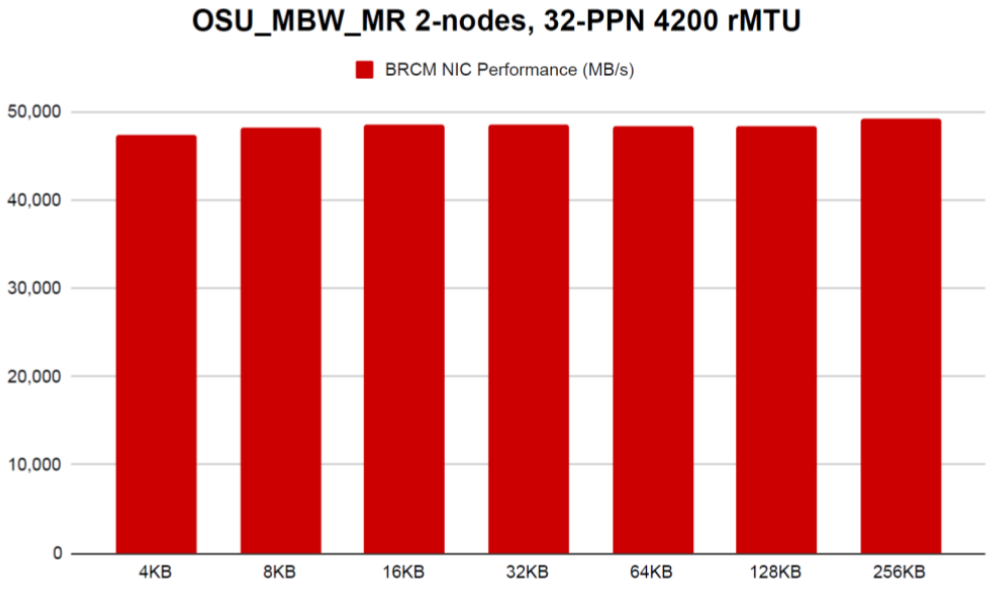

Figure 8: osu_mbw_wr benchmark test with Broadcom NICs and Arista Switches on 400GE

OSU MPI All to All (osu_alltoall) Latency Test

This benchmark test measures the min, max, and average latency of operation across N processes, for various message lengths, over many iterations and reports the average completion time for each message length.

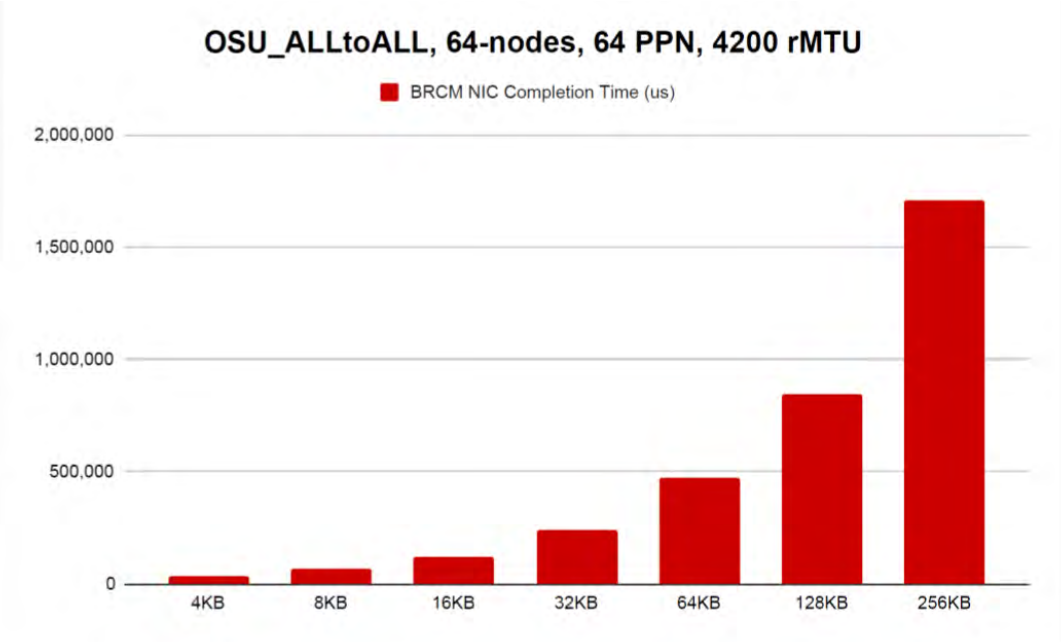

Figure 9: osu_alltoall latency test with Broadcom NICs and Arista Switches on 400GE

OSU All Reduce (osu_allreduce) Latency Test

Like osu_alltoall, osu_allreduce benchmark test measures the min, max, and average latency of operation across N processes, for various message lengths, over many iterations and reports the average completion time for each message length.

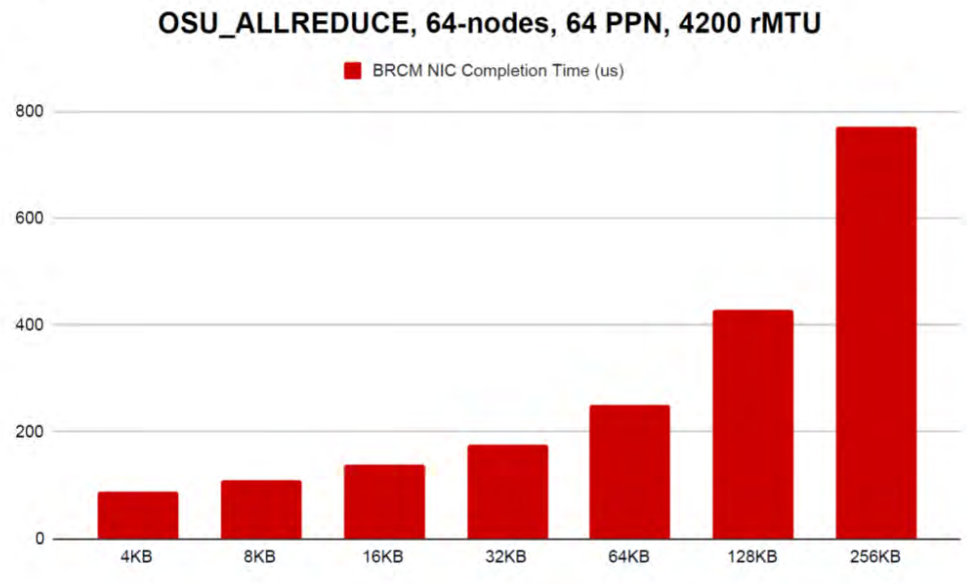

Figure 10: OSU All Reduce Latency Test with Broadcom NICs and Arista Switches on 400GE

Cabling

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. Table 7 outlines the different connectivity options and compares key deployment metrics to consider.

Table 7: Comparison Metrics for different Cable/Optics Types					
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

LPO Technology Primer

LPOs eliminate a DSP that is normally present in the optical transceivers and AOC cables. The elimination of the DSP means that no signal processing or modulation is performed and the output signal is a linear representation of the input signal. Any required signal correction is handled by the NIC SerDes and/or the remote network switch SerDes.

The elimination of DSP processing in the transceiver leads to lower cost and lower power dissipation. The LPO technology allows for 40% lower power compared to DSP-based optical transceivers, and lower latency. LPO modules also run at lower temperatures, which significantly improves reliability.

LPO functionality and interfaces are defined by the two main specifications, OIF CEI-112G-LINEAR-PAM4 and 100G-DR-LPO MSA. The CEI-112G-LINEAR specification defines the 112 Gb/s chip-to-module, near package or co-packaged PAM4 electrical interface for use in the range 36 to 56 GBd. The LPO MSA specification defines the 100 Gb/s/lane 53.125 GBd PAM4 optical interfaces, optical links using standard single-mode fiber, and host-module electrical interfaces for hosts with DSP based SerDes.

Summary

Arista and Broadcom are committed to fulfilling the evolving requirements of AI applications, both presently and in the future. This commitment entails implementing a robust, pre-configured solution that delivers a highly scalable 400G end-to-end optimized network. Furthermore, our 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 be operational in a minimal timeframe.

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 CloudVision

Arista optics Details, Q&A

Arista Broadcom RoCE Datasheet

Broadcom Ethernet Network Adapters

Broadcom Ethernet NIC Configuration Guide

Broadcom Ethernet NIC Firmware and Drivers Downloads

Broadcom RoCE Configuration Guide

Broadcom Ethernet NIC Congestion Control

Congestion Control for Large-Scale RDMA Deployments

Configuring Peer Memory Direct with Broadcom NICs

Broadcom cable solutions guide

Cable/Interconnect Compatibility for BCM957608 (Thor 2) Adapters

Cable/Interconnect Compatibility for BCM9575XX, BCM9574XX, and Earlier Adapters

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