Lossless Network for AI/ML/Storage/HPC with RDMA

Version 1

October 2022

BROADCOM

Table of contents

Introduction 4

RDMA Over Converged Ethernet (RoCE) 4

Deploying RoCE on Arista Switches 5

- Configuration and Verification of PFC on Arista Switches 6
- Configuration and Verification of ECN on Arista Switches 6
- PFC Watchdog 9

Deploying RoCE on Broadcom Ethernet NIC Adapters 9

- RoCE Congestion Control on Broadcom Ethernet NIC Adapters 10
- Installation guide for Broadcom Ethernet NIC Adapters 10
  - *Updating the Firmware on Broadcom Ethernet NIC Adapters* 10
  - *Configuring NVRAM* 11
  - *Host Requirements for Driver/Library Compilation* 11
  - *Installing the Layer 2 and RoCE Driver* 11
  - *Updating Initramfs* 12
  - *Installing the RoCE Library* 12
  - *Validating the RoCE Installation* 12
  - *Confirm Traffic Flow to the remote RoCE endpoint* 13
- Configuring Priority Flow Control on Broadcom NICs 14
- Configuring Congestion Control on Broadcom NICs 14

RoCE Performance Data 15

- OSU MPI Multiple Bandwidth / Message Rate (osu_mbw_wr) Test 16
- OSU MPI All to All (osu_alltoall) Latency Test 16
- OSU All Reduce (osu_allreduce) Latency Test 17
- GPCNet 17

RoCE Applications 18

- Peer Memory Direct 18
- SMB Direct 19
- iSCSI Extensions for RDMA 19
- NFS over RDMA 20

Table of contents

NVMe-oF over RDMA 21

Summary 22

Introduction

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. As the future of metaverse applications evolve, 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.

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 UDP/IP header and enables a routable RoCE. Broadcom's Ethernet Adapters support RoCEv2 in hardware and allows 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.

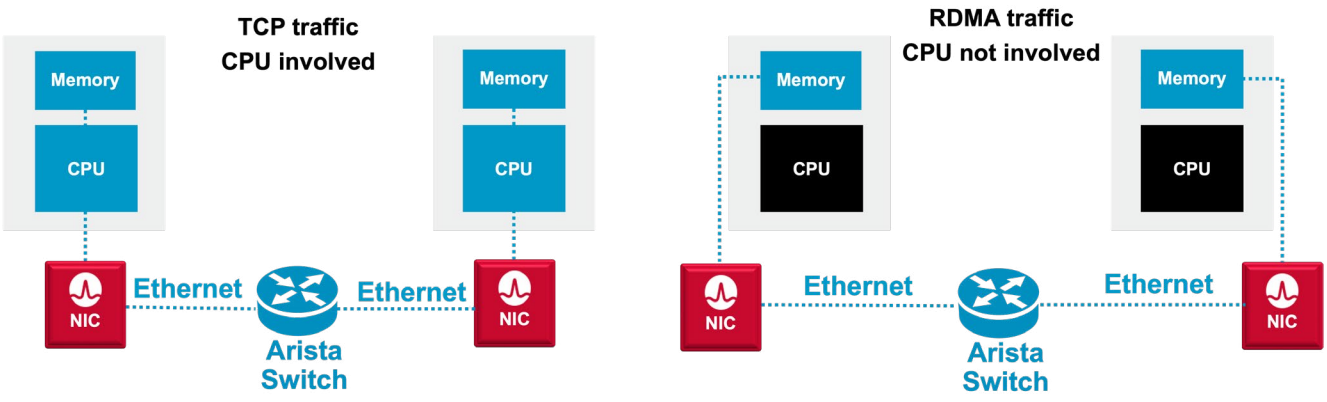


Figure 1: RoCE with Broadcom NICs and Arista Switches

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 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 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 sending rate is adjusted in response to CNPs

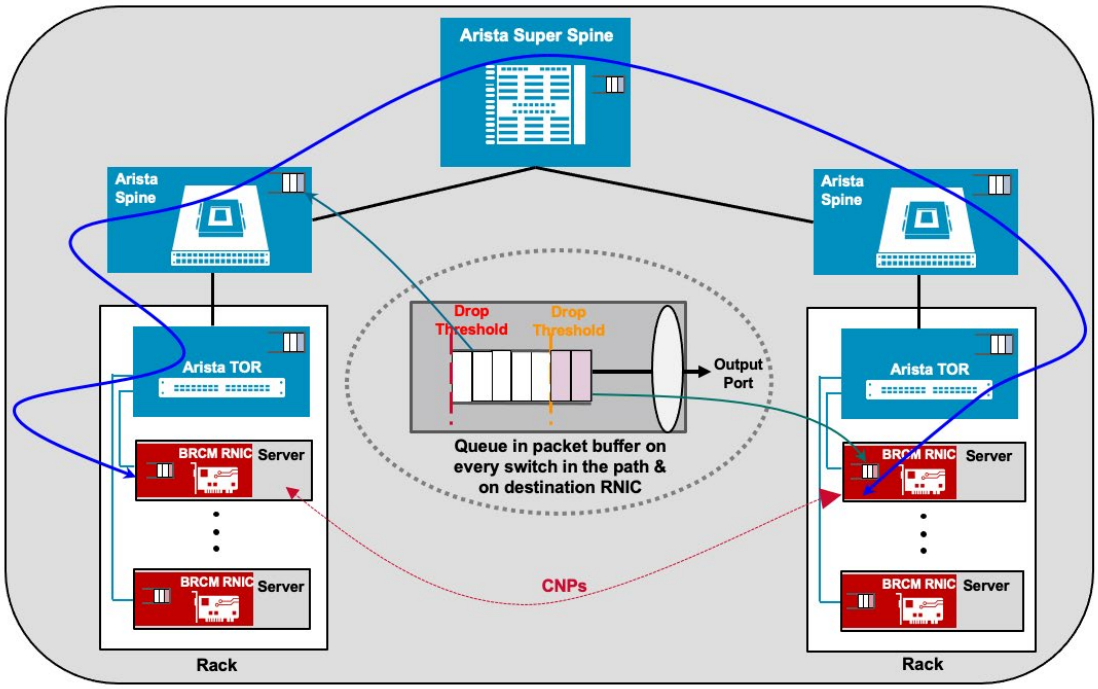


Figure 2: RoCE traffic in datacenter network

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 use of PFC and DCQCN to support RoCE deployments across 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 Switch	Description
7800R3	Highest density 100G/400G Deep Buffer, Lossless Modular Super Spine Switch
7280R3	High Performance 10/40/100/400G Data Center switch with Dynamic Deep Buffer
7060X	High Performance 10/40/100/400G Fixed Configuration Switch
7050X3	1 RU 10/40G Multilayer Leaf Switches

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 forced 100gfull
   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 interfaces ethernet < > priority-flow-control 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 defines 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 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 average queue size is greater than the maximum threshold, packets are marked for congestion.
- If average queue size is between minimum and 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
arista(config-if-Et6/1/1-txq-6)#random-detect ecn minimum-threshold 120 kbytes maximum-
threshold 150 kbytes max-mark-probability 100
arista(config-if-Et6/1/1-txq-6)#
```

- 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 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, 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 datacenter connectivity and AI/ML applications. Broadcom supports a broad portfolio of Ethernet NIC Adapters ranging from 1Gbps – 200Gbps 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 BCM575xx (Thor) ASIC and the adapters support 10GE, 25GE, 100GE and 200GE 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
BCM957504-N425G	BCM57504	4x 25G	SFP28
BCM957504-N1100G	BCM57504	1x 100G	QSFP56
BCM957504-N1100GD	BCM57504	1x 100G	DSFP
BCM957508-N2100G	BCM57508	2x 100G 1x 200G	QSFP56

Table 3: Broadcom PCIE NIC Adapters with RoCE support			
Part Number	ASIC	Ports	I/O
BCM957504-P425G	BCM57504	4x 25G	SFP28
BCM957508-P2100G	BCM57508	2x 100G 1x 200G	QSFP56

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-p and DCQCN-d, where DSQCN-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 rise 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). But 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 congested queue level starts to rise, only a small percentage of packets traversing the switch are ECN marked. Some of the flows which 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 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

This section talks about the procedures to install Broadcom Ethernet adapters and to configure RoCE.

Updating the Firmware on Broadcom Ethernet NIC Adapters

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

```
sudo bnxtnvm -dev=<host network interface name> install <firmware package>
```

Example:

```
sudo bnxtnvm -dev=ens2f0np0 install BCM957508-N1200FZ.pkg
```

Configuring NVRAM

To update the NVRAM configuration, use the bnxtnvm utility provided with the release.

Run bnxtnvm 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 bnxtnvm -dev=ens2f0np0 getoption=support_rdma:0
sudo bnxtnvm -dev=ens2f0np0 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 bnxtnvm -dev=ens2f0np0 setoption=support_rdma:0#1
sudo bnxtnvm -dev=ens2f0np0 setoption=performance_profile#1
```

NOTE: The portion of the command that is dark red changes depending on the name of the host network interface.

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”
```

» Ubuntu/Debian

See the following commands for Ubuntu or Debian operating systems:

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

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:

```
BRCM_DRIVER_VERSION=1.10.2-222.0.95.1
tar xvf netxtreme-bnxt_en-${BRCM_DRIVER_VERSION}.tar.gz
cd netxtreme-bnxt_en-${BRCM_DRIVER_VERSION}.tar.gz
```

```
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. If any library is found, rename it.

   ```
   find /usr/lib64 –name libbnxt_re–rdmav\*.so;find /usr/lib –name libbnxt_re–rdmav\*.so;find /
   lib64 –name libbnxt_re–rdmav\*.so
   ```

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 source. Note that the portion of the command that is dark red below is release specific.

   ```
   BRCM_LIB_VERSION=222.0.95.1
   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
   ```

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 {} \;
   ```

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:                 222.0.95.0
node_guid:              be97:e1ff:feda:96d0
sys_image_guid:         be97:e1ff:feda:96d0
vendor_id:              0x14e4
vendor_part_id:         5968
hw_ver:                 0x2100
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
```

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 --report_gbits <ip address of remote end point>`

```
ib_write_bw -d bnxt_re0 --report_gbits 192.168.2.30
---------------------------------------------------------------------
                    RDMA_Write BW Test
Dual-port       : OFF                   Device         : bnxt_re0
Number of qps   : 1                     Transport type : IB
Connection type : RC                    Using SRQ      : OFF
TX depth        : 128
CQ Moderation   : 100
Mtu             : 4096[B]
Link type       : Ethernet
GID index       : 1
Max inline data : 0[B]
rdma_cm QPs     : OFF
Data ex. method : Ethernet
---------------------------------------------------------------------
local address: LID 0000 QPN 0x06c9 PSN 0xc5d95b RKey 0x2000212 VAddr 0x007fa3640ed000
GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:02:20
remote address: LID 0000 QPN 0x06cb PSN 0x448da7 RKey 0x2000308 VAddr 0x007f9edfec1000
GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:02:30
---------------------------------------------------------------------
#bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec] MsgRate[Mpps] Conflicting CPU
frequency values detected: 1200.407000 != 1286.066000. CPU Frequency is not max.
 65536      5000             96.92               93.53 0        .178400
```

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 3. 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, bnxt_setupcc.sh script can be used. The script is provided as part of the binary package from Broadcom.

1. Install the bnxtqos RPM package using the following command:

   ```
   sudo rpm -ivh bnxtqos-<version>.rpm
   ```

   The bnxt_setupcc.sh script is installed in the /usr/bin directory.

2. Configure the RoCE and CNP priorities using the following command:

   Command usage:

   ```
   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
   ```

The sample command shown in the previous example is configuring:

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

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 DCTCP. To change to DCQCN algorithm, use the following procedure.

NOTE: In addition to setting cc_mode to DCQCN, it is recommended to tune other advanced parameters for optimal performance.

- Configuring DCQCN

  To configure DCQCN, 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
echo -n 1 > cc_mode
echo -n 0 > cp_bias_en
echo -n 1 > tr_update_mode
echo -n 7 > exp_crcp_ratio
echo -n 0x3ff > max_cp_cr_th
echo -n 0x3eb > reduce_cf_rtt_th
echo -n 1 > apply
```

- Configuring DCTCP

  To configure DCTCP, 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
echo -n 0 > cc_mode
echo -n 1 > cp_bias_en
echo -n 0 > tr_update_mode
echo -n 1 > exp_crcp_ratio
echo -n 0x280 > max_cp_cr_th
echo -n 0x15e > reduce_cf_rtt_th
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

For measuring performance numbers on the cluster with Broadcom NICs (100GE) and Arista switches (100GE/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

Server	Switch	NIC	Benchmarks
Model: Dell R740 CPU: Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz Thread(s) per core: 2 Core(s) per socket: 16 Memory Type: DDR4 - 2666 MT/s Memory: 96 GB (48GB/Socket) Kernel: 5.11.0-44 (Ubuntu 20.04)	Model: Arista DCS-7060DX4-32S-F (TH3 TOR) Hardware Revision: 11.01 Software Version: 4.27.2F	Model: Broadcom P2100G Driver Version: 223.1.25.0 Firmware Version: 223.1.25.0 Congestion Control OSU: DCQCN-p Congestion Control GPCNet: DCQCN-d	UCX: 1.10.0 OpenMPI: 4.1.1 GPCNet: 1.1 OSU: 5.8

A summary of the performance numbers from various benchmark tests are captured in Table 6 below.

Table 6: Broadcom NIC / Arista switch ROCE Performance		
Category	Test	Broadcom P2100 NIC
Baseline	OSU Benchmark osu_mbw_mr, 128KB, 4 nodes, 16 PPN	48.6 Gbps (389 Gbps)
Collectives	OSU Benchmark Blocking osu_alltoall latency, 16 nodes, 16 PPN, 128KB message	47.8 ms
	OSU Benchmark Blocking osu_allreduce latency, 16 nodes, 16 PPN, 128KB message	393 us
Congestion Control	GPCNet Benchmark Random Ring Two-sided Latency under congestion, 16 nodes, 16 PPN (99 percentile latency)	9.3 us
	GPCNet Benchmark Multiple Allreduce Latency under congestion, 16 nodes, 16 PPN (99 percentile latency)	39.2 us

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.

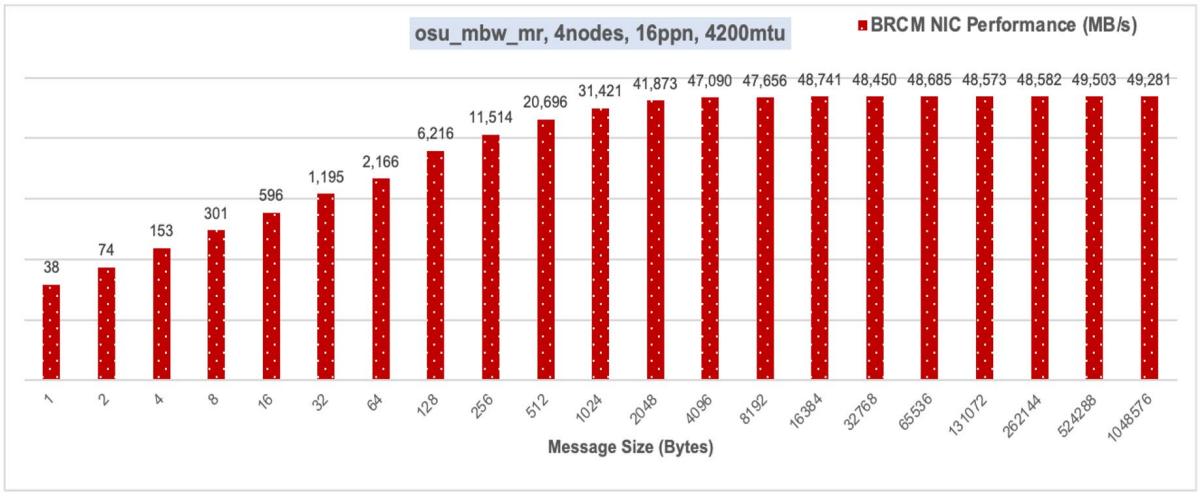


Figure 3: osu_mbw_wr benchmark test with Broadcom NICs and Arista Switches

OSU MPI All to All (osu_alltoall) Latency Test

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

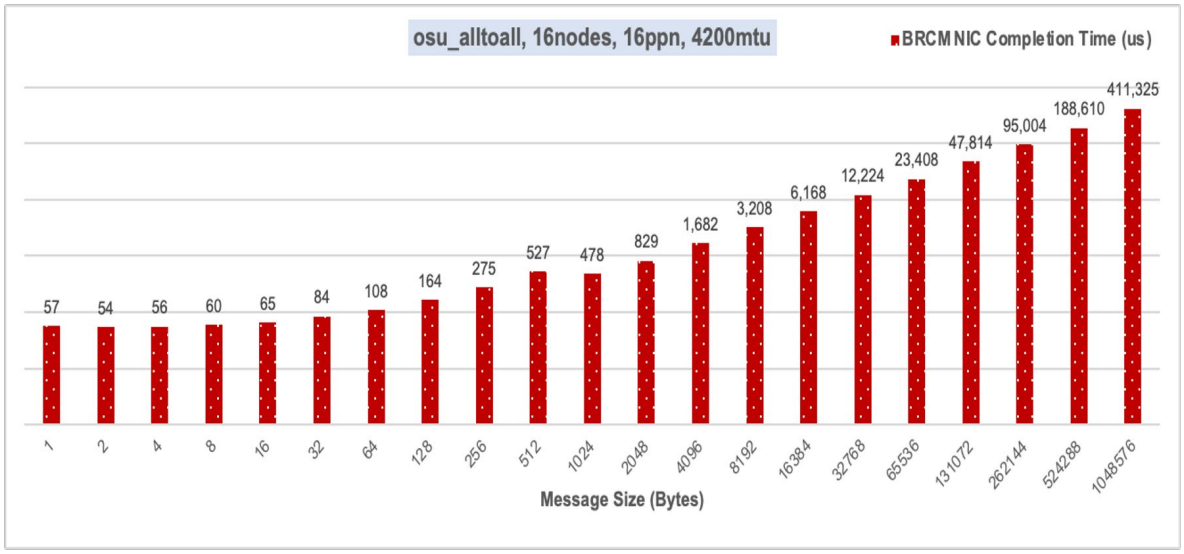


Figure 4: osu_alltoall latency test with Broadcom NICs and Arista Switches

OSU All Reduce (osu_allreduce) Latency Test

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

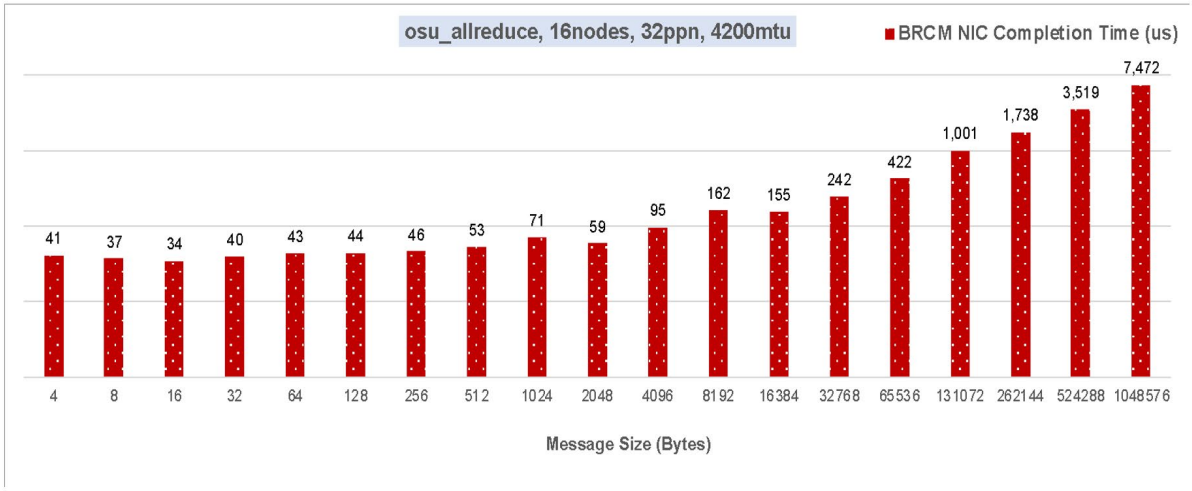


Figure 5: OSU All Reduce Latency Test with Broadcom NICs and Arista Switches

GPCNet

Global Performance and Congestion Network Test (GPCNet) is a generic, topology agnostic bench-mark suite that captures the complex workloads anticipated on multitenant HPC networks. Broadcom NICs support different congestion control algorithms (dcqcn-d and dcqcn-p) for different applications. In this case, dcqcn-d maintains a shallow buffer leading to far lower completion time under congestion.

Table 7: GPCNet Benchmark test with Arista Switch and Broadcom NICs			
gpcnet, 16 nodes, 16 ppn, 4200 mtu			
		Broadcom NIC dcqcn-d	Broadcom NIC dctqcn-d
Isolated / Unloaded	All Reduce: Lat:Avg	15.6 us	15.6 us
	All Reduce: Lat:99%	24.6 us	24.6 us
	RR Two-Sided Latency: BW:Avg	1144.7 MiB/s/rank	1245.2 MiB/s/rank
	RR Two-Sided Latency: BW:99%	748.2 MiB/s/rank	905.9 MiB/s/rank
	RR Two-Sided Latency: Lat:Avg	2.4 us	2.4 us
	RRTwos: Lat:99%	4.6 us	4.6 us
Congestion	All Reduce: Lat:Avg	23.3 us	41 us
	All Reduce: Lat:99%	39.2 us	99.3 us
	RR Two-Sided Latency: BW:Avg	734.9 MiB/s/rank	710.6 MiB/s/rank
	RR Two-Sided Latency: BW:99%	276.4 MiB/s/rank	214.2 MiB/s/rank
	RR Two-Sided Latency: Lat:Avg	4.1 us	10.1 us
	RR Two-Sided Latency: Lat:99%	9.3 us	46.4 us

RR = Random Ring Communication Pattern

RoCE Applications

There are several new technologies that will allow data centers to benefit from performance improvements provided by RoCE. These include:

Peer Memory Direct

General Purpose GPU (GPU) and dedicated AI accelerators have been gaining a lot of traction in machine learning where it takes a massive amount of computing performance to train the sophisticated deep neural networks. This training can take days to weeks. Due to vast amount of data for training, it is desirable to distribute computations across multiple systems with accelerators, which result in increased data exchange for data loading to the accelerators prior to the computations and for data shuffling after the computations. This loading/shuffling process repeats as the training processes are iterative.

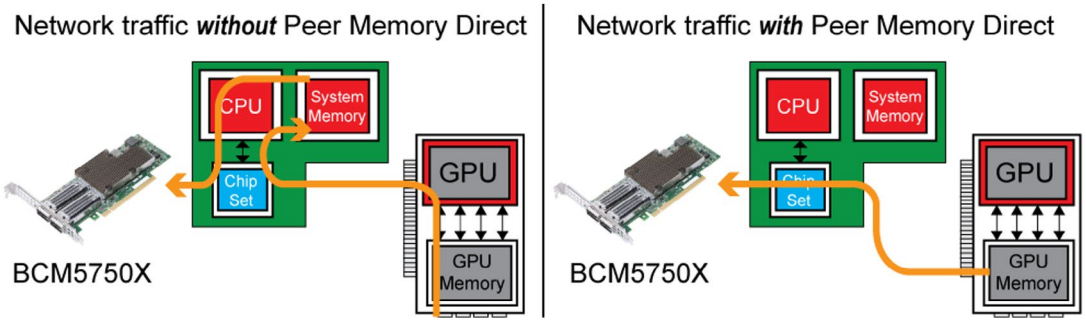


Figure 6: Peer Memory Direct Data Transfer

A closer look at these data flows reveals that the net data exchanges are moved between the accelerators on the same server or across different servers. Both types of flow indicate that a better data movement isolation can be put into place to isolate memory pressure to the memory fabrics, memory controllers and DDR memory. This is the motivation of Peer Memory Direct or AI Direct.

For the exchanges between different servers, RDMA capability on NICs is critical to offer a low latency and highly efficient and yet CPU bypass transport. Traditional RDMA software stack enables direct memory exchanges between process' memories which are server system DDR memory. Then the CPU is involved in moving data from system memory to accelerator memory. These additional movements add latencies and are subject to the limit of system memory bandwidth. Peer Memory Direct technology circumvents this by taking advantage of PCIe peer-to-peer (P2P) transfer and thereby eliminates CPU bandwidth and latency bottle necks. It also eliminates system memory copies and CPU overhead for transferring data to/from GPU memory. For configuring Peer Memory Direct with Broadcom NICs, refer to the link.

SMB Direct

Server Message Block (SMB) is an application-layer network protocol that provides shared access to files, printers, and serial ports. Microsoft provides support for high performance storage networks using RoCE and Microsoft uses SMB in this scenario.

This enables a remote file server to work like local storage with applications that use Microsoft SQL Server and Microsoft Storage Server.

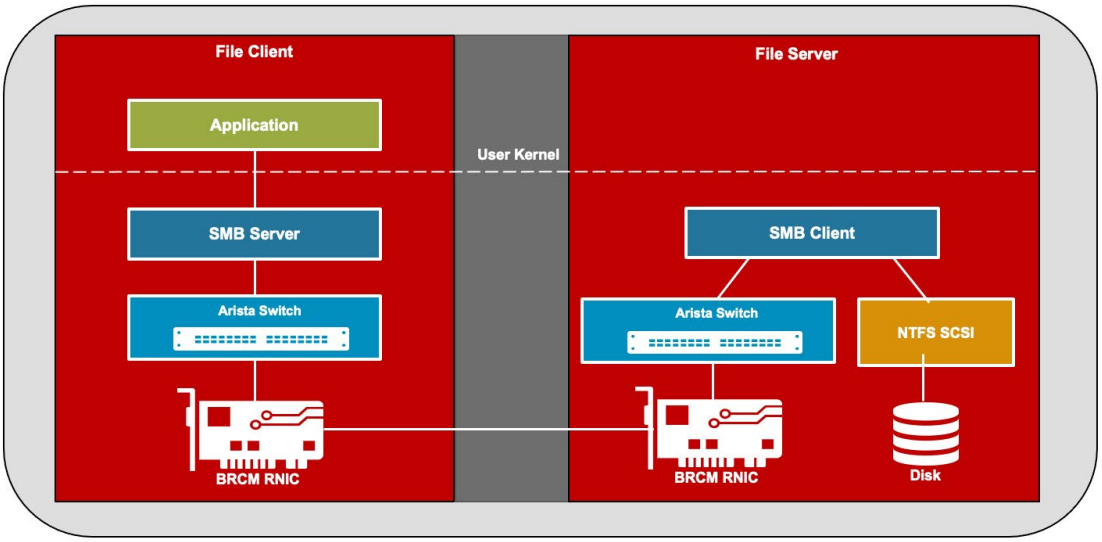


Figure 7: SMB Direct with Arista Switch and Broadcom NICs

SMB 3.0 added the SMB Direct feature that works with network adapters that support RoCE. This capability provides high-performance remote file access for servers and is ideal for use cases such as virtualization and databases. SMB Direct high-performance capabilities are also very beneficial for live migrations with Hyper-V deployments.

The combination of SMB Direct and RoCE adapters provide:

- Scalable, fast and efficient storage access
- High throughput with low latency
- Minimal CPU utilization for I/O processing
- Load balancing, automatic failover and bandwidth aggregation using SMB Multichannel

iSCSI Extensions for RDMA

Performance for Internet Small Computer System Interface (iSCSI) storage has also been enhanced with iSCSI extensions for RDMA (iSER). The iSER protocols are defined in RFCs 5047 and 7145 and enable RDMA to be used to transfer data directly between memory buffers for computers and storage devices.

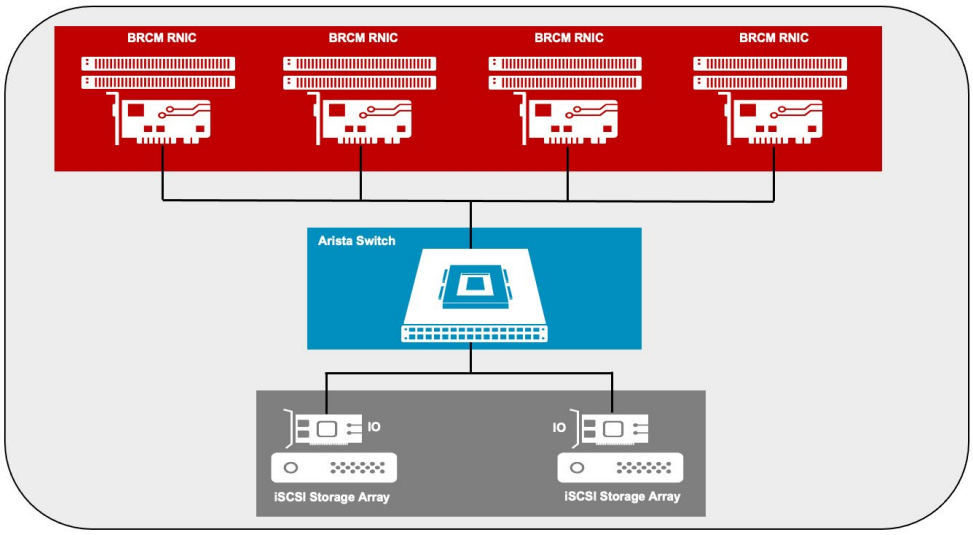


Figure 8: iSER with Arista Switch and Broadcom NICs

iSER promises to provide significant performance improvements over iSCSI due to eliminating the TCP/IP processing overhead, this becomes significant with increased Ethernet speeds of 10/25/100GbE and beyond. iSER will provide higher throughput for storage applications, lower latency and more efficient use of server and storage controller processing resources.

NFS over RDMA

Network file system (NFS) is a distributed file system protocol that allows users on client computers to access files over a network as if it was local storage. NFS is an open standard defined with request for comments (RFCs) that enable ongoing development and implementation of new technologies. One focus area has been the remote procedure call (RPC) layer for NFS that provides communication between the client and server. RDMA support has been added to the RPC layer with RFCs 5532, 5666 and 5667 to provide enhanced data transfer performance.

Using RoCE for NFS over RDMA has the potential for similar performance benefits as SMB Direct for increasing performance of applications servers that use network file storage. NFS clients and servers can expect higher throughput at smaller data block sizes as well as increased I/O operations per second (IOPS), lower latency and reduced NFS client and server CPU consumption.

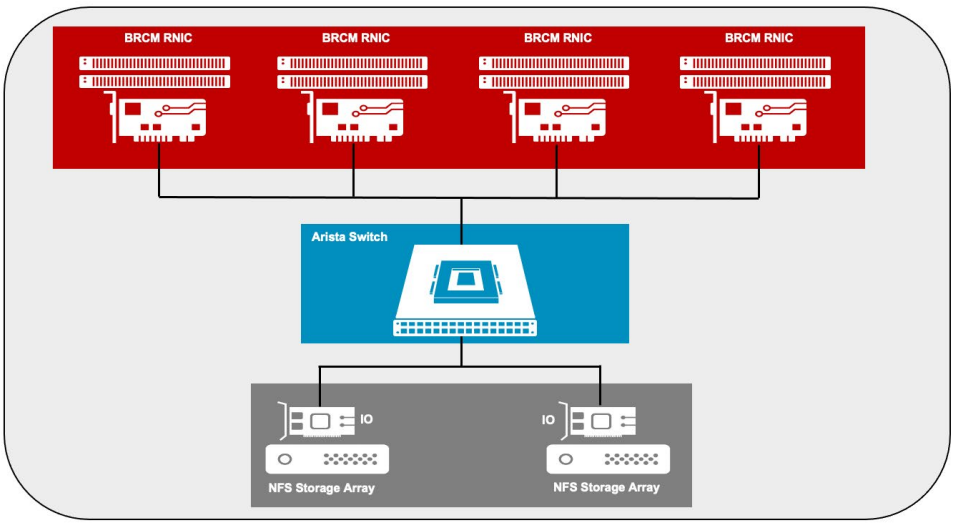


Figure 9: NFS over RDMA with Arista Switch and Broadcom NIC

NVMe-oF over RDMA

The NVMe protocol is not limited to simply connecting a local flash drive inside a server, it may also be used over a network. When used in this context, a network "fabric" enables any-to-any connections among storage and server elements. NVMe over Fabrics (NVMe-oF) is enabling organizations to create a very high-performance storage network with latency that rival direct attached storage. As a result, fast storage devices can be shared, when needed, among servers. NVMe over fabric is an alternative to SCSI on Fibre Channel or iSCSI, with the benefit of a lower latency, a higher I/O rate and an improved productivity.

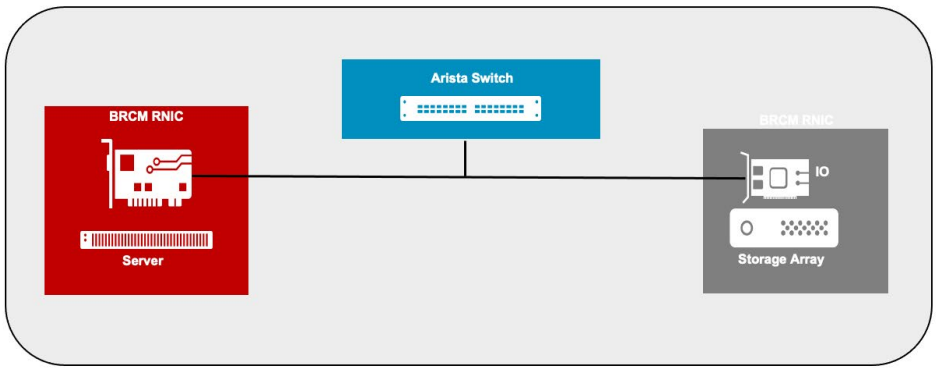


Figure 10: NVMe-oF over RDMA with Arista Switch and Broadcom NICs

Summary

RDMA is a proven technology to provide greater performance and scalability. With the heavy compute requirements associated with intensive AI/ML & storage workloads, RoCE is a fundamental component in ensuring a reliable, end-to-end transport solution for modern data centers. Arista and Broadcom are committed to support open standards-based congestion control mechanisms on the switches and NICs and are at the forefront to meet the requirements of datacenter applications to deliver a reliable, high throughput, low latency network. Deployments leveraging Arista switches and Broadcom NICs also benefit from the power efficient ethernet solutions which are extremely critical for data centers today and contribute to TCO savings.

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 Broadcom RoCE Datasheet
- 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
- Broadcom Ethernet NIC Congestion Control
- Congestion Control for Large-Scale RDMA Deployments
- Configuring Peer Memory Direct with Broadcom NICs

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 © 2022 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. October 10, 2022