PURESTORAGE®

ARISTA

TECHNICAL GUIDE

Deploying NVMe-oF for Enterprise Leaf-Spine Architecture

Best practices and configuration guidelines for deployment of NVMe/RoCE on a Layer 3 leaf-spine data-center network.

Contents

ARISTA

Executive Summary

NVMe/RoCE provides an enhanced storage protocol for accessing NVMe based storage targets over ethernet networks, but implementations have been hampered by the need to create isolated storage networks. This paper tests the viability of deploying NVMe/RoCE on a Layer 3 leaf-spine data-center network and provides best practices and configuration guidelines for deployment.

This is illustrated by showcasing a proof of concept environment that measures the performance in a leaf-spine architecture against the more commonly used single switch (single hop) architecture. We performed these tests using a Pure Storage® FlashArray//X, Arista 7050X3 Series switches, and Mellanox ConnectX-5 host adapters. The testing uses the Flexible I/O tester tool to emulate storage data traffic. All configuration settings are detailed in the Appendix.

Assumed Reader Knowledge

This paper showcases lossless transport for storage-area networks (SANs) using NVMe-oF with Pure Storage® FlashArray//X, Mellanox ConnectX-5 adapters, and Arista switches. The lossless transport of storage traffic in this context refers to the ability of the host, array, and network switches to actively prevent traffic from being dropped by using various flow control mechanisms, like the pause mechanism used in priority flow control (PFC). This paper does not focus on the underlying operation of the technologies. Therefore, readers are assumed to be thoroughly familiar with the terminologies and concepts of iSCSI, Data Center Bridging Capability Exchange (DCBX), Link Layer Discovery Protocol (LLDP), Explicit Congestion Notification (ECN), PFC, Remote Direct Memory Access (RDMA) versions 1 and 2, RDMA over Converged Ethernet (RoCE), NVMe-oF, Border Gateway Protocol (BGP), Ethernet VPN, and Virtual Extensible LAN.

A Short History

Storage capacity demands linearly increased year-over-year from 2010 to 2017. Since 2017, there has been an exponential increase in storage capacity demand due to scale-out and web-native applications. Before the introduction of these new applications, enterprises relied on storage area networks (SANs) to provide scaled capacity.

Traditional enterprise applications didn't require the performance offered by scale-out applications that commonly leverage direct-attached storage (DAS). For that reason, Ethernet SANs, using Fibre Channel over Ethernet (FCoE) or iSCSI as a transport worked well because of the relatively low-performance demand. With the introduction of solid-state drives (SSDs) using flash memory as a medium, storage arrays became more performant. But they still relied on the legacy SCSI protocol to access the data on the drives.

The NVMe workgroup was formed In 2009 to address the shortcomings of SCSI as the protocol for flash drives. In 2011, the group published the NVMe specification for PCIe. This standard and the subsequent commercial release of NVMe products in 2012 created real excitement due to the performance advantages over SCSI as a protocol to access flash devices. Shortly after publishing the PCIe specification, the NVMe workgroup began work to extend the protocol across network fabrics and the NVMe over Fabrics (NVMe-oF) specification was published in 2016. This provided a mechanism for high-performance access to NVMe devices over various network fabrics.

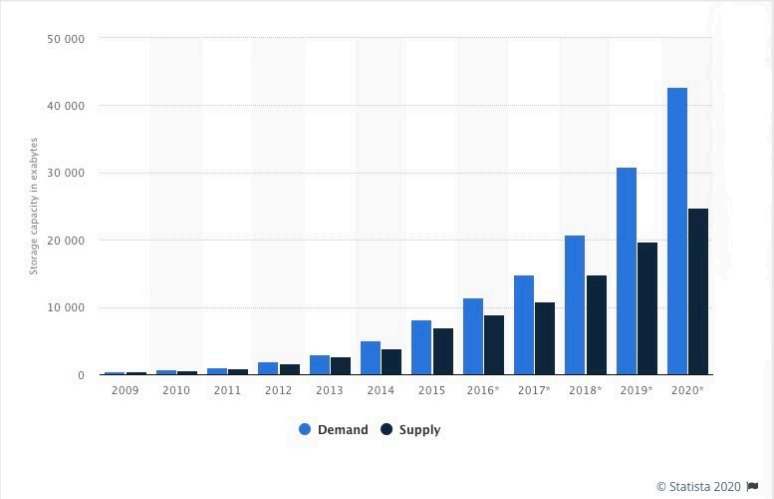


Figure 1. Storage capacity supply-demand comparison

With the Ethernet network speeds increasing to 40G and 100G in 2010, the wider adoption of these higher speeds in data centers, applications moving to virtualized environments, and the scale-out application boom, the introduction of NVMe-oF in SANs has become a certainty. With current SCSI storage protocols, data packets are processed through the host's CPU and the operating system. Using current SCSI storage protocols to provide a high volume of data between the storage arrays and compute hosts, requires the protocol stack to have the significant processing power and reduces the overall storage performance of the compute hosts. As a result, data access latency may suffer. NVMe-oF—which can use RoCEv2 and other congestion-control mechanisms over traditional data-center networks—is an ideal technology for the current demands of high performance, low latency, and low cost in SANs.

This paper describes methods for deploying NVMe/RoCE as a transport in a two-tier Leaf-Spine data center layer-3 BGP EVPN network. This is illustrated by showcasing a proof of concept environment that measures the performance in a leaf-spine architecture to that of the more commonly used single switch (single hop) architecture.

These tests are performed using a Pure Storage® FlashArray™//X, Arista 7050X3 Series switches, and Mellanox ConnectX-5 host adapters. The testing uses the Flexible I/O tester tool to emulate storage data traffic. All configuration settings are described in detail in the Appendix section.

Pure System Architecture

Pure Storage is an innovator in using flash and an early adopter of NVMe standards. The FlashArray//X products provide an end-to-end NVMe solution including NVMe-oF (referred to as DirectFlash™ Fabric) support to the hosts. FlashArray components support an all-NVMe architecture, including the PCIe-enabled chassis, DirectFlash Modules, NVMe NVRAM, and IO cards that support NVMe-oF, as well as the option to add an expansion shelf using NVMe-oF (Figure 2).

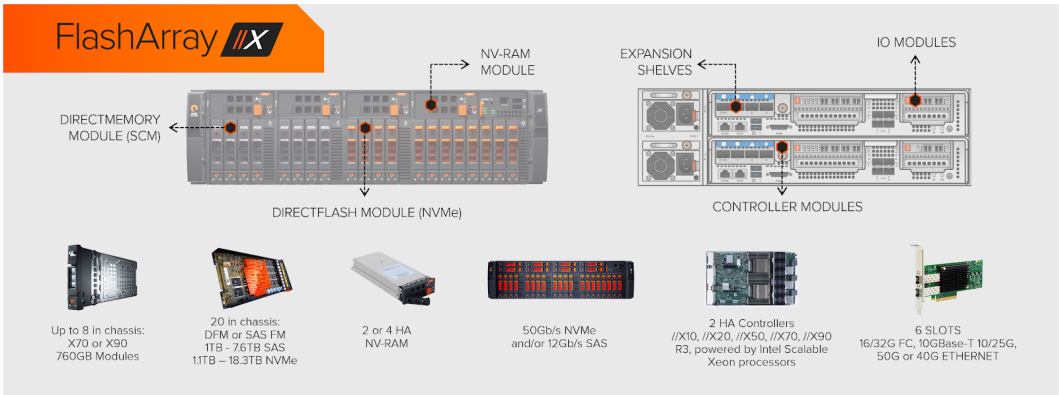


Figure 2. FlashArray//X Hardware Components

The storage controllers are the central components of the FlashArray architecture. They run the Purity software stack to provide storage and storage services. The hardware and software are purpose-built for flash storage and take full advantage of the NVMe specifications. The internal storage components—NVRAM and DirectFlash Modules—connect via PCIe to the processors so that the software can communicate using native NVMe to provide the enhanced performance benefits of the protocol. You can extend these efficiencies to the hosts using NVMe-oF capable IO ports. Figure 3 highlights the internal architecture.

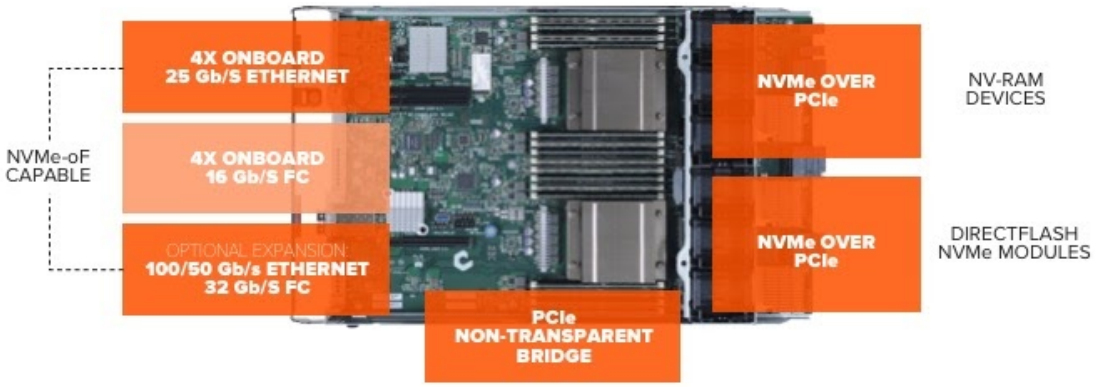


Figure 3. Native NVMe Controller Architecture

For these tests, we used a Pure FlashArray//X with 100G NVMe/RoCE adapters. FlashArray™ supports a variety of industry-standard features including IEEE DCBX, LLDP, PFC, and ECN to enable lossless connectivity. When RoCE services are enabled for a FlashArray, the congestion control features are automatically enabled to minimize the configuration required on the FlashArray. The complementary configuration for the network components is described in the appendices.

Mellanox Host Adapters

The testing for this paper uses Mellanox ConnectX-5 adapters on the initiators and array. These adapter cards introduce new acceleration engines to maximize performance on storage platforms. With sub-700 nanosecond latency, a very high message rate, and NVMe-oF offloads, it provides the highest performance compared to its predecessor versions.

These host adapters also support PFC, ECN, and DSCP, which are used to configure a lossless environment for NVMe/RoCe. All these features are industry standards and may be found on other adapters. In this paper, each adapter port is referred to as a network interface card (NIC).

Arista System Architecture

The Arista 7050X3 Series are Trident3 chipset-based switches, offering latency as low as 800ns and an intelligent fully shared packet buffer of up to 32MB for superior burst absorption. Comprehensive support for 10/25/40/50/100GbE speeds coupled with Arista EOS ensures that the 7050X3 delivers the flexibility and features for big data, cloud, and virtualized network environments. It accommodates myriad applications and east-west traffic patterns found in modern data centers. The 7050SX3-48YC12 offers an overall throughput of 4.8Tbps, while the 7050CX3-32S offers 6.4Tbps. With low latency and no oversubscription, the 7050SX3-48YC12 is optimized for high-performance server and storage deployments. The 7050X3 Series switches use dynamic thresholds to allocate packet memory based on traffic class, queue depth, and network QoS policies, ensuring a fair allocation to all ports of both lossy and lossless classes.

Additionally, with support for features such as PFC, DCBX, and RoCEv2, the Arista 7050X3 enables lossless Ethernet for storage applications. For traffic classes requiring lossless frame delivery, a fixed amount of buffer (~16% of total buffer capacity) is set aside to absorb any in-flight packets that arrive after flow control (such as PFC/PAUSE) is issued. The lossless buffer is a shared pool across all ingress ports with a defined minimum and maximum buffer space for each port.

Network Architecture for NVMe-oF

Traditionally, IP-enabled storage networks like iSCSI have been deployed with redundancy at the leaf layer using Ethernet bonding capability to provide multipaths to clients. This means they use the transport to provide active-active paths and handle failover scenarios.

An active-active solution like multi-chassis link aggregation (MLAG) is deployed to provide redundancy at the leaf layer. MLAG on the leaf layer handles data traffic well. Such traffic is well-suited to land on any NIC on the client. With NVMe-oF using RoCEv2 over a Layer 3 network, multipathing is provided via Linux's multipath driver and NVMe configuration on the initiator. Each NIC on an NVMe-oF client is assigned a unique IP address. RoCEv2 uses queue pairs where the initiator connects to the IPs of the target devices. Once the queue pairs are established, clients can perform RDMA send, receive, read, and write operations. Because the queue pairs are linked with the IP addresses on the NICs, a packet destined to one IP address cannot land on the other NIC of the same client. Therefore, MLAG is not a suitable high-availability solution for NVMe/RoCE. Each NIC connects to multiple IP addresses on the Pure FlashArray to provide multiple paths across the various network fabric topologies.

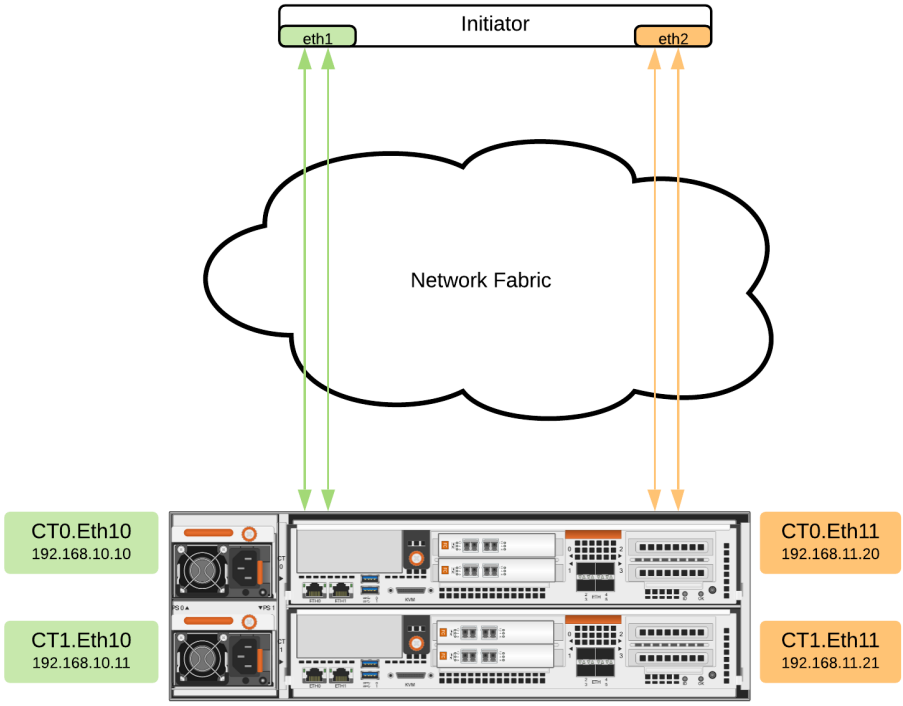


Figure 4. Multiple connections/paths are established by the initiator

Using BGP EVPN as the control-plane of the fabric helps ease configuration on leaf and spine layers, as well as eases rack scalability with minimal configuration touchpoints. VXLAN is used as the data-plane of the fabric, which provides flexibility to extend VLANs across multiple racks. With edge ports on the leaf layer continuing to act as trunk or access switchports, as in traditional SANs, the Class of Service (CoS) value can be mapped to a DSCP value on the leaf's incoming interface to preserve the transmit-queue throughout the fabric and help maintain a lossless fabric. (The configuration details are listed in the Appendix.) Also, support for DSCP-based PFC on the 7050X3 Series switches ensures that this lossless behavior is honored in the VXLAN fabric.

Test Topology

For this white paper, we tested NVMe performance across three different topologies:

- Direct-connect—initiators directly connected to the storage array
- Single-hop—initiators and array both connect to a single Arista switch
- Traditional two-tier leaf-spine architecture

The Direct-connect topology provides a baseline of expected performance. This allows us to compare the impact of the additional switch hops in the single-hop and traditional network designs. Flexible I/O tester is used as a traffic generator to measure latency and bandwidth.

NVMe/RoCE requires a lossless topology. To accomplish this, the initiator, network, and array must have a mechanism for flow control. Lossless transport is accomplished by using PFC and/or Explicit Congestion Notification (ECN). Differential Services Code Point (DSCP) is enabled on the Mellanox ConnectX-5 NICs on initiators and array across all topologies to mark and queue the NVMe/RoCE traffic. All endpoints are configured to send NVMe/RoCE traffic in transmit-queue 3.

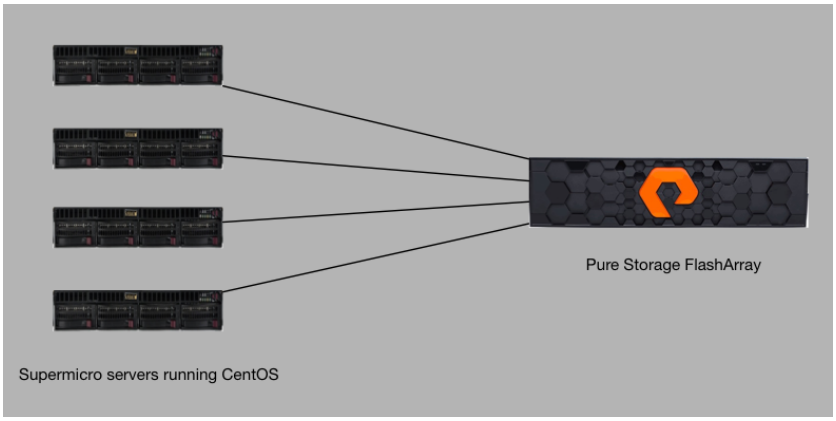


Figure 5. Direct connect test topology

In the direct-connect topology (Figure 5), baseline performance is measured for latency and bandwidth. All four initiators and array are configured for lossless transport using PFC and ECN, with DSCP configurations responsible for sending all traffic in transmit-queue 3. Both 25G and 100G speeds are tested.

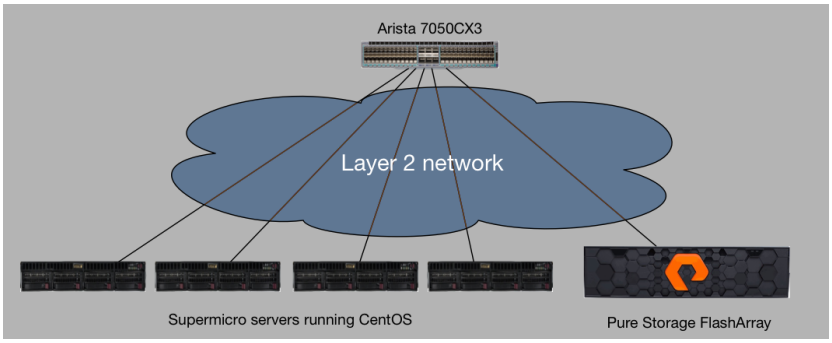


Figure 6. Single-hop test topology

In the single-hop topology (Figure 6), performance is measured for latency and bandwidth. All four initiators and array are configured for lossless transport using PFC and ECN, with DSCP configurations responsible for sending all traffic in transmit-queue 3. The Arista 7050X3 switch is configured with PFC for lossless transport in transmit-queue 3, with all L2 links. For this topology, only PFC is used as the congestion control mechanism. Both 25G and 100G speeds are tested.

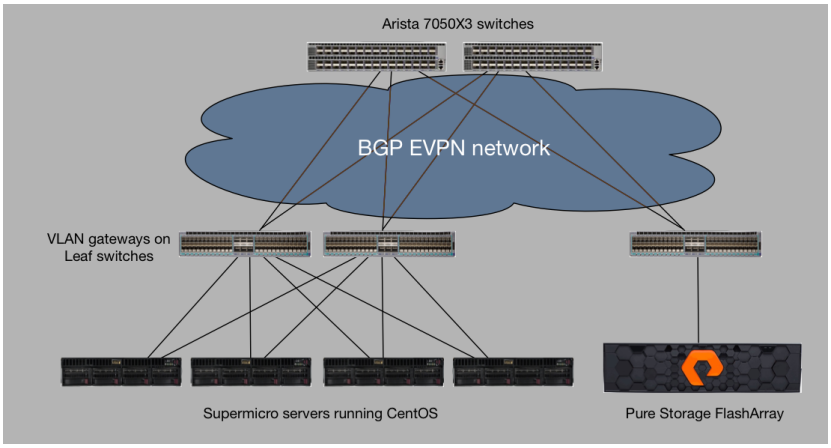


Figure 7. Leaf-spine test topology

In the two-tier leaf-spine topology, performance is measured for latency and bandwidth. All switches run BGP EVPN, with Arista 7050SX3 as leaves configured as EVPN L3 VTEPs with Arista 7050CX3 as spines configured as EVPN route-servers. The edge ports on leaves are configured as L2 access ports with appropriate SVIs acting as gateways. Initiators and array communicate via VXLAN routing. The FlashArray and all four initiators are configured for lossless transport with PFC and ECN, with DSCP configurations responsible for sending all traffic in transmit-queue 3.

Test Methodology

The testing setup is divided into the three topologies. For each setup, tests are performed at 25G and 100G NIC speeds. For testing at 25G speed, only leaf-to-endpoints port speeds are changed to 25G, while speed on the leaf-to-spine links is maintained at 100G. The tests and test results are designed to demonstrate the impact of the network topology and the congestion-control mechanisms on the storage traffic and as such will focus on the metrics that best demonstrate network latency and bandwidth. The main focus of these tests is to measure read and write I/O performance across different topologies, NIC speeds, and congestion-control mechanisms with the Arista 7050X3 Series switches.

In the leaf-spine topology, all Arista switches are tested across the following combinations of congestion-control mechanisms:

- PFC alone enabled on all Arista switches
- ECN alone enabled on all Arista switches
- ECN enabled on all Arista switches and PFC enabled only on leaf switch ports facing the endpoints.

To set a baseline for the effects of the topology on the performance, we need to measure the latency. In performance testing, a variety of factors can impact latency, including the block size, the busyness of the initiator, congestion on the network, or the busyness of the array. To minimize the impact of anything other than the network, we choose a small block

size and test a single initiator connecting to the array with a single-threaded IO job. For latency measurements across topologies and protocols, we use a block size of 4k with a single-threaded I/O job and a single client job at a time with multiple runs per client.

The total latency of all the client runs is averaged. Due to the nature of the FlashArray services, read latencies tend to be more variable due to mechanisms like data reduction, which can speed up read latency in some instances. The most consistent measure of latency will be the writes, so we focus on that value to identify the impact across different topologies.

A low block size provides the best way to measure the latency impact of the network, but it doesn't provide a large amount of stress on the overall network bandwidth. Measuring the impact of the flow control mechanisms across the topologies requires a larger block size.

For bandwidth measurement across topologies and protocols, we use a block size of 512k. We measure bandwidth with multi-threaded I/O jobs to maximize throughput. In contrast to latency measurements using read, throughput measurements will see the highest values on read. The displayed bandwidth values represent the cumulative performance of all four initiators running the same I/O job simultaneously.

NVMe/RoCE vs. iSCSI

Consider NVMe/RoCE in a leaf-spine environment if you want to modernize the storage protocol converting from an iSCSI deployment. To demonstrate the advantages of NVMe/RoCE over iSCSI storage networks in a leaf-spine topology, the test methodology also includes a comparison of iSCSI vs. NVMe/RoCE on the Arista 7050X3 Series switches with the EVPN Leaf-Spine topology. Extensive testing with iSCSI isn't included as we've kept the focus of the white paper to NVMe-oF. Performance numbers of the iSCSI network are compared with an NVMe/RoCE network at the end of the results section.

Test Results

The key metrics for storage performance are IOPS, latency, and throughput. These three metrics are tightly coupled. For example, for read and write operations of constant block size, IOPS and throughput are directly correlated. That means for operations of the same block size, if you change throughput by 10%, IOPS will also change by 10%. The IOPS and throughput are affected by available bandwidth, the performance of the target and initiator, and latency.

In these tests, we are controlling the constant available bandwidth and performance of the target and array. This allows us to focus on the latency aspect of the network. To accomplish this, there are two latency aspects that we need to measure:

- The latency introduced by the network architecture
- The difference in throughput caused by the introduction of latency using various congestion control mechanisms.

This is accomplished by performing two sets of tests.

- The first set focuses on the minimal latency measured between the initiator (server) and the target (array) for a set of tests. This will demonstrate the additional latency introduced by the network architecture (such as the hardware, topology, and traffic protocols) by measuring the average latency of a 4K I/O work profile.

- The second set focuses on the maximum throughput measured between the initiator and target for a set of tests. These tests will demonstrate the effect of the various flow control mechanisms by measuring the average throughput of a 512K I/O work profile.

Set 1. Latency Testing

When reviewing the latency numbers, we will focus our attention on the 4K write numbers for the various topologies. Figure 8 notes the baseline numbers compared to the single hop and leaf-spine (three-hop) shows a 10µs latency increase.

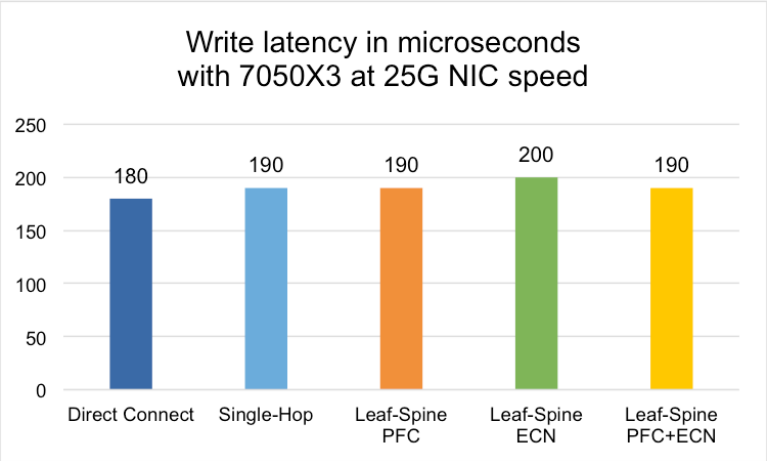


Figure 8. Write latency across different topologies and congestion-control mechanisms

For comparison in Figure 9, you can see the read latency shows similar results, but the variation is wider (40µs higher for the leaf-spine). This is due to the variation caused by read-services components like data reduction. These differences are typically a result of some data being rebuilt from metadata cache instead of reading directly from the drives. This causes some of the reads to perform better than others, causing inconsistent results.

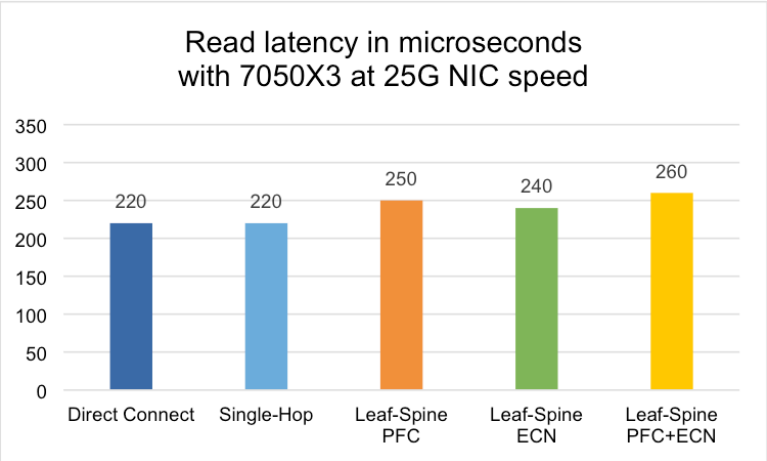


Figure 9. Read latency across different topologies and congestion-control mechanisms

When we compare the same write tests for 100Gb/s (Figure 10), we get consistent results on the order of tens of microseconds across the single-hop topology and the leaf-spine topology (three hops). This is expected since the NIC speed shouldn't impact latency for this test scenario.

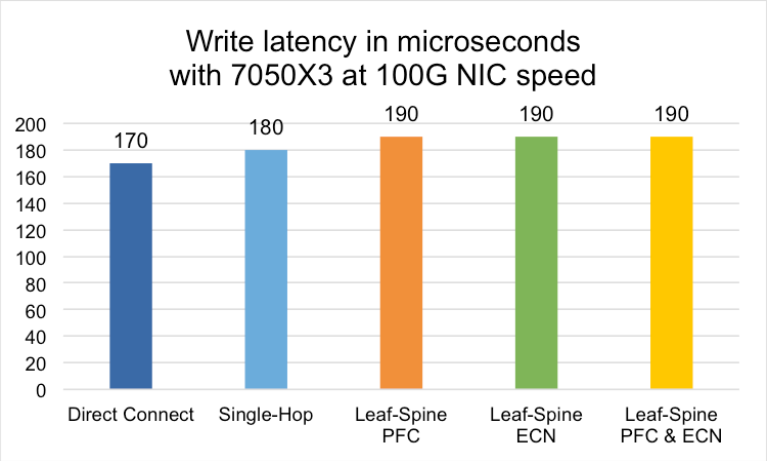


Figure 10. Write latency across different topologies and congestion-control mechanisms

Figure 11 shows similar overall results to the 25Gb/s read tests, but the variation in the latencies are more pronounced (30µs lower for single-hop and 70µs higher for leaf-spine) than shown in Figure 9.

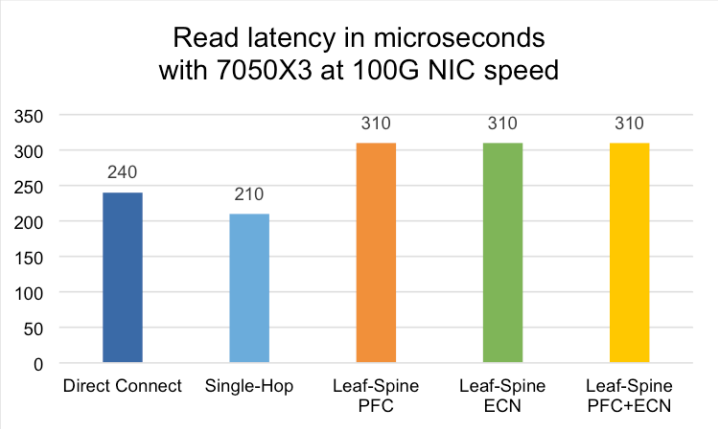


Figure 11. Read latency across different topologies and congestion-control mechanisms

Figures 9 and 11 demonstrate why we would choose the write latency as the test measurement instead of read latency. The variation in the read test indicates that there was acceleration for the reads in some IO, which can create inconsistency in the results without controlling those factors. However, the write path for FlashArray provides a consistent acceleration method because it uses NVRAM technology and therefore provides the most consistent latency measurements.

The write latency in these tests demonstrates that the network topology adds some latency to the storage traffic, but that it is on the order of tens of microseconds.

When compared to the overall latency, which includes those service latencies, these tests show that the latency differences introduced by the network are negligible compared to the latency measurements in a direct connection between the initiator and the array.

Set 2. Throughput Testing

The purpose of the throughput testing is to understand how the various congestion mechanisms (PFC, ECN, or PFC+ECN) impact performance and to determine whether one is superior. Throughput testing uses a 512K block size with multiple initiators and is designed to push the array and/or initiator to the point that it's overwhelmed and needs a congestion mechanism to limit traffic. These tests stress only the array and the initiators, and not the network fabric because there is ample bandwidth in each of the network topologies.

In each of these tests, we achieved 12GB/s to 13GB/s read throughput and a write throughput close to 4GB/s to 4.5GB/s. These tests are performed using a 512K block size and all four initiators performing I/O jobs simultaneously.

Because the tests were designed to push the limits on the end devices (initiator and target) to demonstrate the congestion mechanisms, the results show no observable throughput benefit of 100GB/s over 25GB/s. If the network was a point of congestion, it would trigger the same congestion mechanism and impact the throughput of the tests. It stands to reason that more bandwidth can result in more overall throughput. For most designs, it makes sense to maximize the connectivity speed on the target connections to handle the throughput of multiple host connections.

As in the latency tests, the write throughput will have the least variation and provide the best indication of the impact of the congestion mechanism on the throughput. The Direct Connect topology yields the best performance. Implementation of PFC/ECN in leaf-spine topology brings the throughput down because of the lossless transport mechanisms kicking in. The end points or switches can generate PAUSE/ECN frames, resulting in the sender pausing its transmissions for while the receiver services existing IO in the queue.

Figures 12 through 15 show the bandwidth performance across the different topologies and congestion-control mechanisms at 100Gb/s and 25Gb/s speeds. The results are very consistent across topologies and congestion mechanisms. Read jobs show the best throughput performance because the write throughput is typically limited by the performance of the target (i.e. writes require more resources than reads). In all of the below cases, congestion is created on the endpoints, triggering transmission of PAUSE/ECN frames. The switches react by buffering data packets and forwarding the congestion control frames toward the sender. The results show that addition of switches in the network path does not affect the throughput significantly compared to the direct-connect topology.

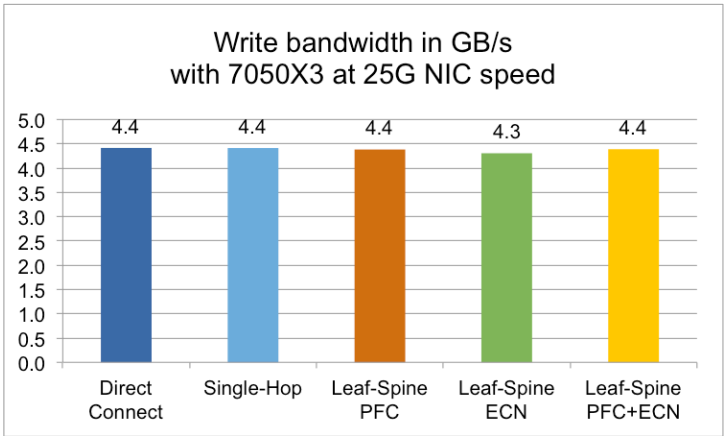


Figure 12. Write bandwidth across different topologies and congestion-control mechanisms

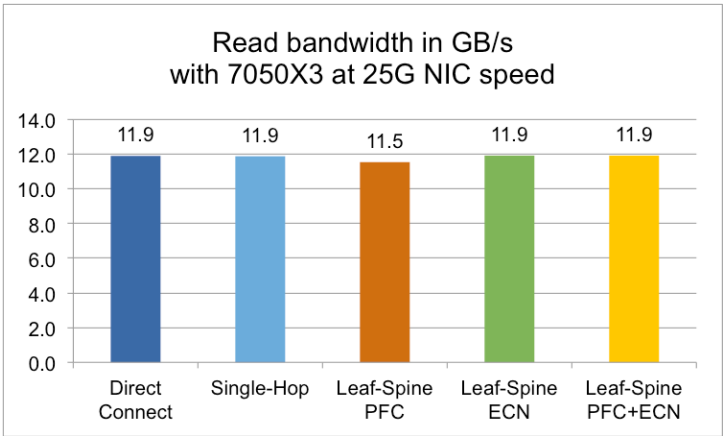


Figure 13. Read bandwidth across different topologies and congestion-control mechanisms

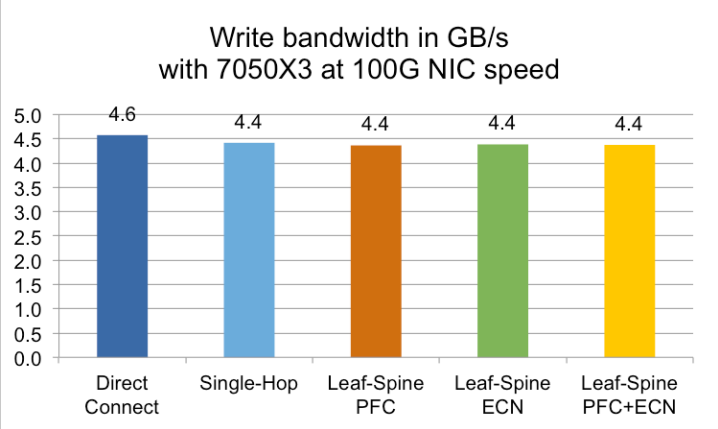


Figure 14. Write bandwidth across different topologies and congestion-control mechanisms

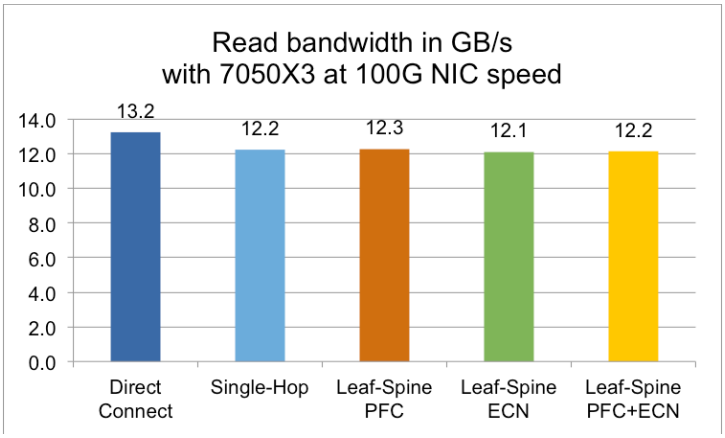


Figure 15. Read bandwidth across different topologies and congestion-control mechanisms

Many NVMe/RoCE deployments have been on isolated, single-hop networks. This testing suggests that introducing NVMe/RoCe traffic across a well-designed leaf-spine data center network will have no considerable performance impact.

iSCSI comparison

There has long been a comparison of Fibre Channel storage protocols to Ethernet storage protocols. This has always been in the context of iSCSI vs. FC-SCSI. In general, iSCSI performance has been considered sub-standard or “second rate” when compared to FC-SCSI. A lot of the performance differences have been blamed on the network. This had an impact early on, with lower speed and highly oversubscribed networks.

The advances in the leaf-spine architecture—such as non-blocking designs, network overlays, and increase in port speeds—have all but eliminated the network as the bottleneck. The testing data in this paper comparing NVME/RoCE over direct connect vs. leaf-spine topologies provides evidence that a well-designed network will not affect performance.

As further evidence of NVMe/RoCE’s benefits as an Ethernet-based storage protocol, we ran the 4K latency tests and the 512K throughput tests for iSCSI on the leaf-spine topology.

The following graphs show the performance of iSCSI across the same network using the same initiators and array, and just changing the transport. Figures 16 and 17 show the single-threaded write and read latency of 4k I/O at 100Gb/s and 25Gb/s NIC speeds for NVMe/RoCE vs iSCSI. The graphs show a ~40% decrease in the 4k write latency for NVMe/RoCE and ~25-30% decrease in read latency. The decreases are due to efficiencies of the protocol stack end-to-end as well as offload from the CPU of the host and initiator. As seen before, the NIC speed has no impact on latency.

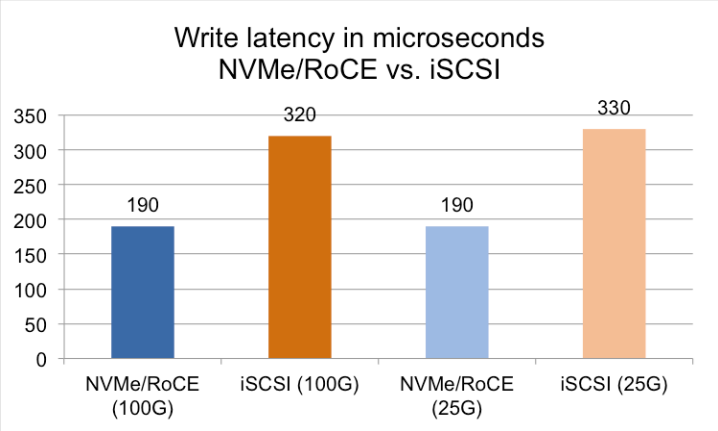


Figure 16. Write latency comparison between NVMe/RoCE and iSCSI at different NIC speeds using Arista 7050X3 switches

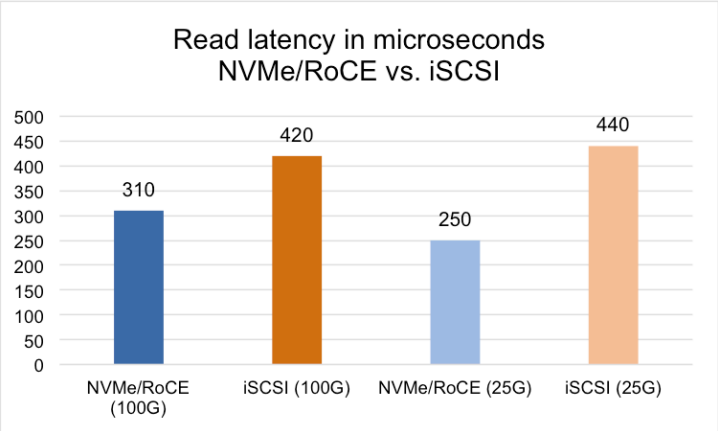


Figure 17. Read latency comparison between NVMe/RoCE and iSCSI at different NIC speeds using Arista 7050X3 switches

Figures 18 and 19 show the multi-threaded write and read bandwidth of 512k I/O at 100Gb/s and 25Gb/s NIC speeds for NVMe/RoCE vs iSCSI. Figure 19 indicates that the NVMe/RoCE write throughput has an improvement of about 5% over iSCSI. Figure 19 shows that the performance difference of the 512k read IO is closer to 20%. The write throughput difference is muted by the fact that the limit of the workload profile is ~4.4GB/s.

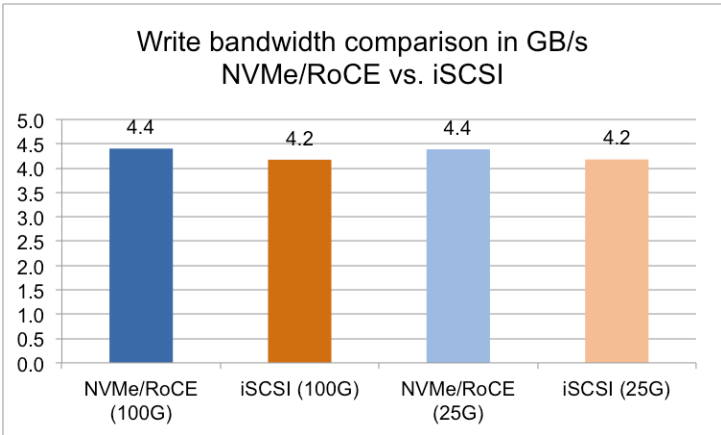


Figure 18. Write bandwidth comparison between NVMe/RoCE and iSCSI at different NIC speeds using Arista 7050X3 switches

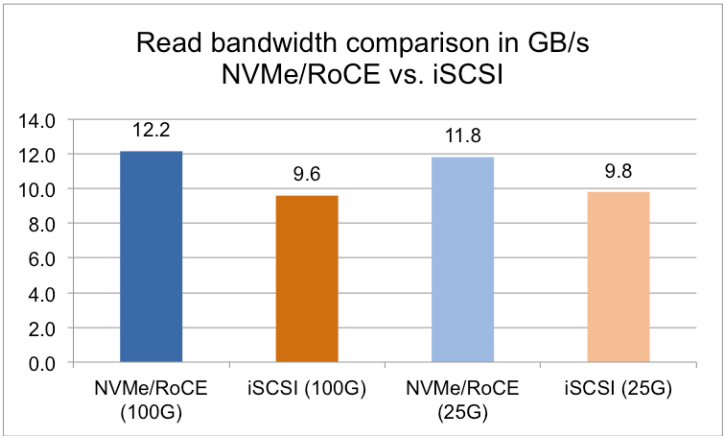


Figure 19. Write bandwidth comparison between NVMe/RoCE and iSCSI at different NIC speeds using Arista 7050X3 switches

The iSCSI to NVMe/RoCE comparison provides additional evidence that the network is not the primary reason that traditional ethernet enabled storage has been less performant than more efficient protocols.

Conclusion

NVMe/RoCE improves performance for storage workloads that you can extend to existing and new storage deployments with Arista switches and Pure Storage FlashArray//X. The leaf-spine design with BGP EVPN VXLAN provides ease of configuration and scalability advantages to extend Layer 2 or Layer 3 domains across multiple compute and storage racks. You can choose the congestion-control mechanism of your choice with this architecture because it provides similar performance for read and write workloads across PFC, ECN, and PFC+ECN scenarios.

Appendix

Data traffic in multiple queues and congestion on Arista switches is not part of this document. Congestion control mechanisms are enabled on Arista switches, but the switches do not actively experience congestion.

Arista Configuration Overview

This section describes the required configuration on Arista 7050X3 switches used for testing in this white paper. ECN thresholds listed below are for the tested scenarios and will vary depending on the traffic profile in deployment scenarios.

QoS, ECN, PFC Configuration: To maintain QoS markings set by the initiator and the array, following QoS configuration is used on the Leaf Arista switches:

```
policy-map type quality-of-service pm-RoCE
  class cm-RoCE
      set dscp 24
   !
   class class-default
```

This policy-map is applied to the ingress interfaces on the Leaf switches (7050X3) with the following ECN thresholds:

```
interface Ethernet51/1
   switchport access vlan 111
   service-policy type qos input pm-RoCE
   !
   tx-queue 3
      random-detect ecn minimum-threshold 256 kbytes maximum-threshold 512 kbytes max-
mark-probability 100 weight 0
      random-detect ecn count
```

For testing PFC, the following PFC configuration is used on all Arista switches on all active interfaces:

```
interface Ethernet51/1
   priority-flow-control on
   priority-flow-control priority 3 no-drop
```

In VXLAN environments, to copy ECN bits from the inner to the outer header and vice versa, the following configuration is needed.

```
interface Vxlan1
   vxlan qos ecn propagation
```

Show commands from Arista switches: The output below is from the 7050X3 leaf switch (connected to the array) during a write job. Ethernet 51/1 to 54/1 are client-facing ports, while Ethernet 49/1 to 50/1 are uplinks to spine layer switches. PFC counters show pause frames being propagated through the fabric. The LANZ output shows the buffers getting exercised on the same leaf:

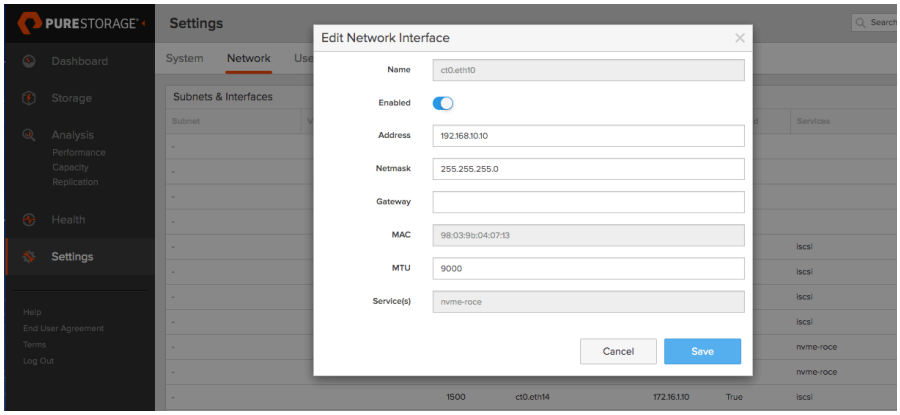
```
Leaf3-7050X3# show priority-flow-control counters | nz
Port                 RxPfc          TxPfc
Et49/1                   0          37706
Et50/1                   0          46060
Et51/1              317226              0
Et52/1              325121              0
Et53/1              316348              0
Et54/1              321240              0

Leaf3-7050X3#show queue-monitor length

Report generated at 2020-04-17 17:54:16
S-Start, U-Update, E-End, TC-Traffic Class
Segment size for S, U and E congestion records is 256 bytes
* Max queue length during period of congestion
+ Period of congestion exceeded counter
-------------------------------------------------------------------------------------------
Type    Time                  Interface      Congestion     Queue
                              (TC)           duration       length
                                             (usecs)        (segments)
-------------------------------------------------------------------------------------------
 E   0:00:00.02961 ago        Et52/1(3)      4379           732*
 S   0:00:00.03291 ago        Et54/1(3)      N/A            680
 E   0:00:00.03376 ago        Et54/1(3)      940            1114*
 S   0:00:00.03399 ago        Et52/1(3)      N/A            732
 S   0:00:00.03470 ago        Et54/1(3)      N/A            1114
 S   0:00:00.03476 ago        Et53/1(3)      N/A            595
 E   0:00:00.03635 ago        Et51/1(3)      286            527*
 S   0:00:00.03663 ago        Et51/1(3)      N/A            527
 E   0:00:00.03766 ago        Et51/1(3)      644            675*
 S   0:00:00.03830 ago        Et51/1(3)      N/A            675
```

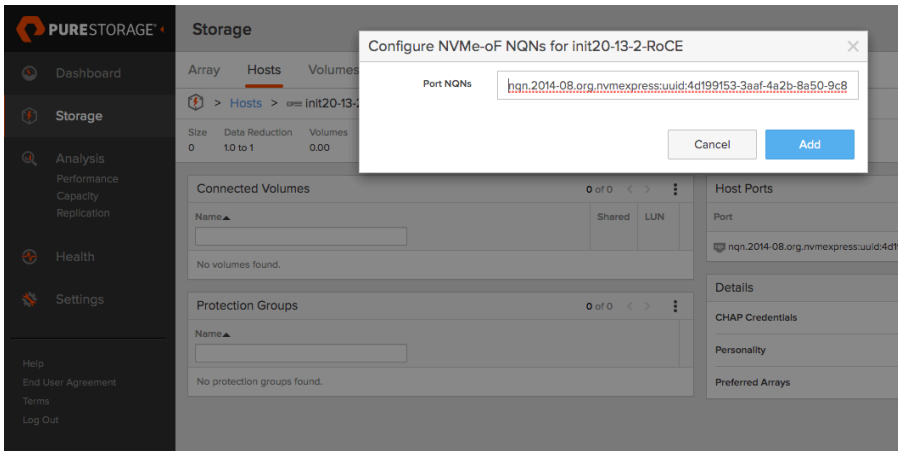
Pure Configuration Overview

This section describes the configuration of NVMe/RoCE on Pure FlashArray//X. NVMe/RoCE services are available on Pure FlashArray //X20R2 and FlashArray//X20R3 and above that have an NVMe/RoCE card installed. NVMe/RoCE services are enabled by contacting support@purestorage.com. Once the services are enabled you will need to set up an IP address, netmask, and MTU for each RoCE interface using either the GUI or the CLI.




```
pureuser@flasharray> purenetwork setattr --address 192.168.10.10 --netmask 255.255.255.0 --mtu 9000 ct0.eth10
Name       Enabled  Subnet  Address        Mask           Gateway  MTU   MAC                Speed       Services  Subinterfaces
ct0.eth10  True     -       192.168.10.10  255.255.255.0  -        9000  98:03:9b:04:07:13  25.00 Gb/s  nvme-roce -
pureuser@flasharray>
```

Once you have set up the ports you need to add the NQN to the host configuration on the Array using the GUI or the CLI:




```
pureuser@flasharray> purehost setattr --nqnlist nqn.2014-08.org.nvmexpress:uuid:4d199153-3aaf-4a2b-8a50-9c826390c0c0 init20-13-2-RoCE
Name               WWN  IQN  NQN                                                                  Host Group
init20-13-2-RoCE   -    -    nqn.2014-08.org.nvmexpress:uuid:4d199153-3aaf-4a2b-8a50-9c826390c0c0  -
pureuser@flasharray>
```

Adapter Configuration

The following section outlines the steps required to configure PFC and ECN on the host (initiator) adapters and is part of the initiator documentation guide on the Pure Storage support site. Refer to the complete initiator configuration guide for RHEL/CentOS. To configure the Mellanox adapters you need to install the Mellanox QOS tools on the host and configure the adapter. To install the tools for RHEL/CentOS, use the following steps as outlined in the Mellanox tools sections of the NVMe/RoCE initiator setup guide.

```
[root@init20-13-2 ~]# wget http://www.mellanox.com/downloads/ofed/MLNX_EN-4.3-
1.0.1.0/MLNX_EN_SRC-4.3-1.0.1.0.tgz

[root@init20-13-2 ~]# tar -zxvf MLNX_EN_SRC-4.3-1.0.1.0.tgz

[root@init20-13-2 ~]# rpm -ivh MLNX_EN_SRC-4.3-1.0.1.0/SRPMS/mlnx-ofa_kernel-4.3-
OFED.4.3.1.0.1.1.g8509e41.src.rpm

[root@init20-13-2 ~]# cd /root/rpmbuild/SPECS

[root@init20-13-2 SPECS]# rpmbuild -bp mlnx-ofa_kernel.spec

[root@init20-13-2 SPECS]# ln -s /root/rpmbuild/BUILD/mlnx-ofa_kernel-
4.3/source/ofed_scripts/utils/mlnx_qos /usr/bin/mlnx_qos

[root@init20-13-2 SPECS]#
```

Once you have installed the tools you will need to configure QoS on the Mellanox adapter as outlined in the Configuring IP QoS on the host sections of NVMe/RoCE initiator setup guide.

In the /opt directory create a qos.sh script.

```
[root@init20-13-2 ~]# vi /opt/qos.sh
```

Press the I key to insert and enter the following into the file.

```
#!/bin/bash
#
for f in `ls /sys/class/infiniband`;
do
        echo "setting TOS for IB interface:" $f
        mkdir -p /sys/kernel/config/rdma_cm/$f/ports/1
        echo 106 > /sys/kernel/config/rdma_cm/$f/ports/1/default_roce_tos
done
#
for i in ```lshw | grep Mellanox -A3 | grep 'logical name'|awk '{print $3}'```
do
        echo "setting dscp trust for interface:" $i
        mlnx_qos -i $i --trust=dscp --pfc=0,0,0,1,0,0,0,0
done
~
-- INSERT --
```

Press escape then wq at the : prompt and then enter to save the configuration and exit.

```
:wq
```

Change the permissions of the file to include execute.

```
[root@init20-13-2 ~]# chmod +x /opt/qos.sh
```

Run the script. This sets the TOS for each RoCE interface and uses the mlnx_qos tool to set the PFC as well as the trust DSCP parameters on the RoCE adapters.

```
[root@init20-13-2 ~]# sh /opt/qos.sh
setting TOS for IB interface: mlx5_0
setting TOS for IB interface: mlx5_1
setting dscp trust for interface: eth1
DCBX mode: OS controlled
Priority trust state: dscp
dscp2prio mapping:
prio:0 dscp:07,06,05,04,03,02,01,00,
prio:1 dscp:15,14,13,12,11,10,09,08,
prio:2 dscp:23,22,21,20,19,18,17,16,
prio:3 dscp:31,30,29,28,27,26,25,24,
prio:4 dscp:39,38,37,36,35,34,33,32,
prio:5 dscp:47,46,45,44,43,42,41,40,
prio:6 dscp:55,54,53,52,51,50,49,48,
prio:7 dscp:63,62,61,60,59,58,57,56,
Cable len: 7
PFC configuration:
priority    0   1   2   3   4   5   6   7
enabled     0   0   0   1   0   0   0   0
tc: 0 ratelimit: unlimited, tsa: strict
priority:  0
priority:  1
priority:  2
priority:  3
priority:  4
priority:  5
priority:  6
priority:  7
setting dscp trust for interface: eth2
DCBX mode: OS controlled
Priority trust state: dscp
dscp2prio mapping:
prio:0 dscp:07,06,05,04,03,02,01,00,
prio:1 dscp:15,14,13,12,11,10,09,08,
prio:2 dscp:23,22,21,20,19,18,17,16,
prio:3 dscp:31,30,29,28,27,26,25,24,
prio:4 dscp:39,38,37,36,35,34,33,32,
prio:5 dscp:47,46,45,44,43,42,41,40,
prio:6 dscp:55,54,53,52,51,50,49,48,
prio:7 dscp:63,62,61,60,59,58,57,56,
```

```
Cable len: 7
PFC configuration:
priority    0   1   2   3   4   5   6   7
enabled     0   0   0   1   0   0   0   0
tc: 0 ratelimit: unlimited, tsa: strict
priority:  0
priority:  1
priority:  2
priority:  3
priority:  4
priority:  5
priority:  6
priority:  7
[root@init20-13-2 ~]#
```

Santa Clara—Corporate Headquarters
5453 Great America Parkway
Santa Clara, CA 95054
Tel: 408-547-5500
www.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 & B, 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

Pure Storage, Inc.
650 Castro Street, #400
Mountain View, CA 94041