Configuring NVMeoF RoCE For SUSE 15

For a long time, storage has been outclassed when compared to the leaps and bounds compute performance has grown over the last ten years. With the adoption of the NVMe (non-volatile memory express) as a standard for accessing flash storage, this is no longer true. We can now exploit the levels of parallelism available in modern NVMe devices to achieve lower latency and greater performance.

With the launch of DirectFlash™ Fabric earlier in 2019, FlashArray//X™ is now capable of delivering the low latencies and performance gains for shared storage environments. Prior to any implementation of NVM Express over fabrics (NVMe-oF), those wishing to benefit from NVMe storage would need to use direct-attached storage. This is not always ideal as many applications and organizations depend on centralized storage with data services in order to reduce costs and complexity, and increase efficiency.

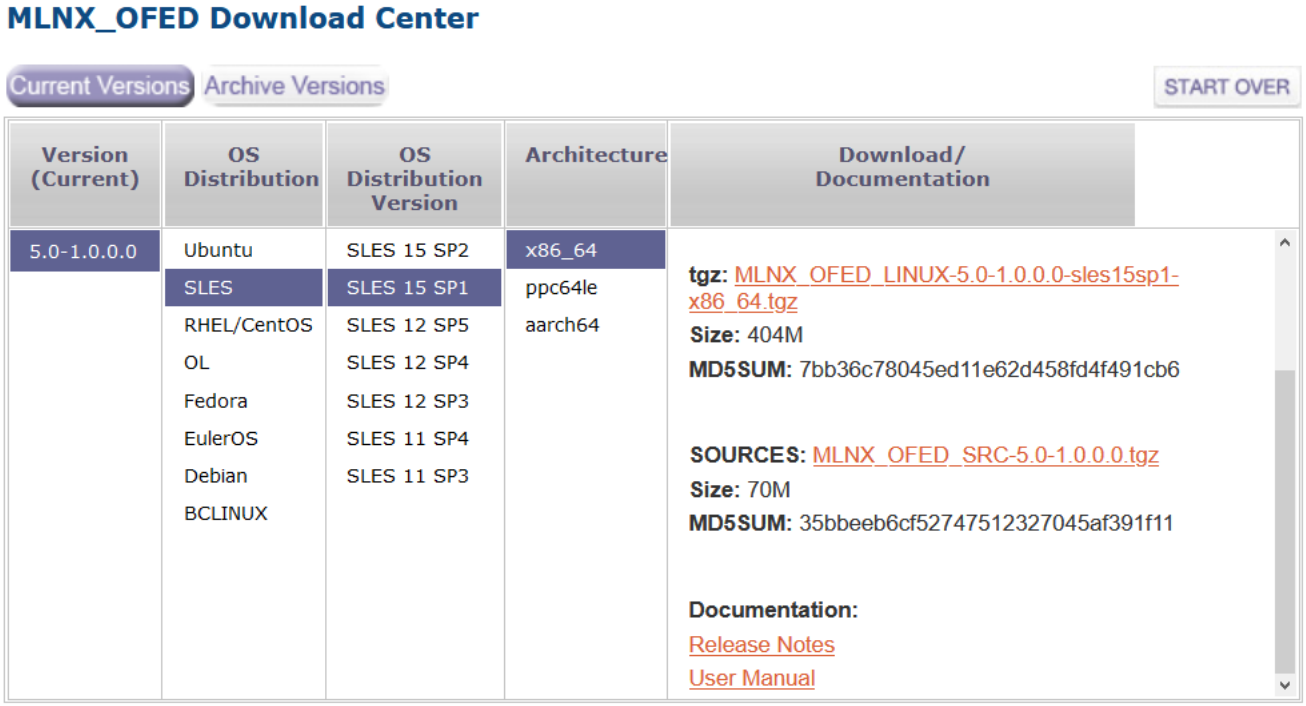
The purpose of this blog post is to provide the steps required to implement NVMe-oF using RDMA over Converged Ethernet (RoCE) for SUSE Enterprise Linux (SLES) 15 and subsequent releases.

An important item to note is that RoCE requires a lossless network, requiring global pause flow control or PFC to be configured on the network for smooth operation.

All of the below steps are implemented using Mellanox Connect-X4 adapters.

System and software requirements

- SUSE 15 SP1 or higher.
- A Mellanox Connect-X4 or higher adapter installed in the system.
- The Development Tools Module should be added in Extension and Module Selection.
- The latest Mellanox OFED SRC package for SUSE needs to be downloaded and built for the priority flow control(PFC) quality of service (QoS) tools. This was the package used in the below steps. NVMe/RoCE works with the inbox drivers and Mellanox OFED does not need to be installed.



Step 1. Install the following packages using the zypper package manager on the host.

- rpm-build
- nvme-cli

Step 2. Configure multipathing on the host.

- Ensure Native NVMe multipathing is turned off by appending “nvme-core.multipath=N” to the optional kernel parameters in /boot/grub2/grub.cfg (reboot required)
- Add the following device definition to the multipath.conf file :

[crayon-642795b0d719f042937108/]

Step 3. Build the Mallanox OFED package to get access to the QOS tool on the host

- Decompress the downloaded Mellanox OFED package

[crayon-642795b0d71aa493658097/]

- In the decompressed folder install the source rpm

[crayon-642795b0d71ad978915592/]

- Build the Mellanox kernel specification to get access to the Mellanox QOS utility

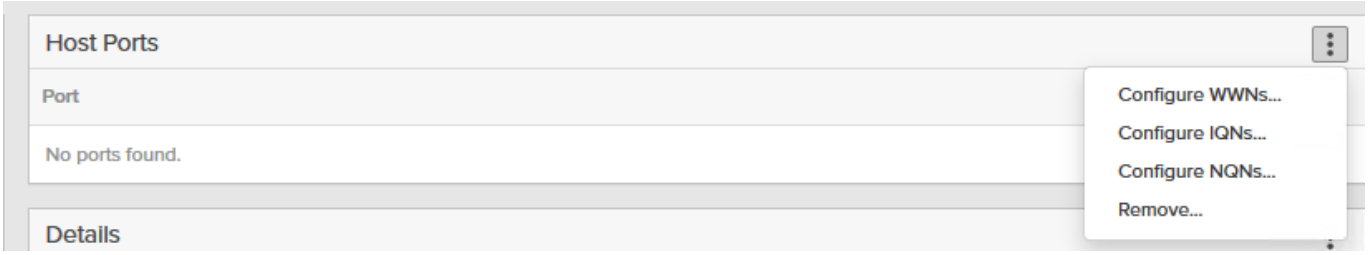
[crayon-642795b0d71ae638403171/]

- Once built use the mlnx_qos tool to set the correct PFC queue and DSCP trust state for each mellanox port used for NVME-oF through RoCE in the system (the ports on our system were named eth6 and eth7)
  [crayon-642795b0d71b0145821892/]

Step 4. Set the TOS for RoCE ports

Run the following command loop to set the TOS for all RDMA interfaces to 106:
[crayon-642795b0d71b2519722372/]
Step 5. Generate and get the NVMe qualified name on the host and then configure and connect some volumes to it in the Pure Storage Web GUI.

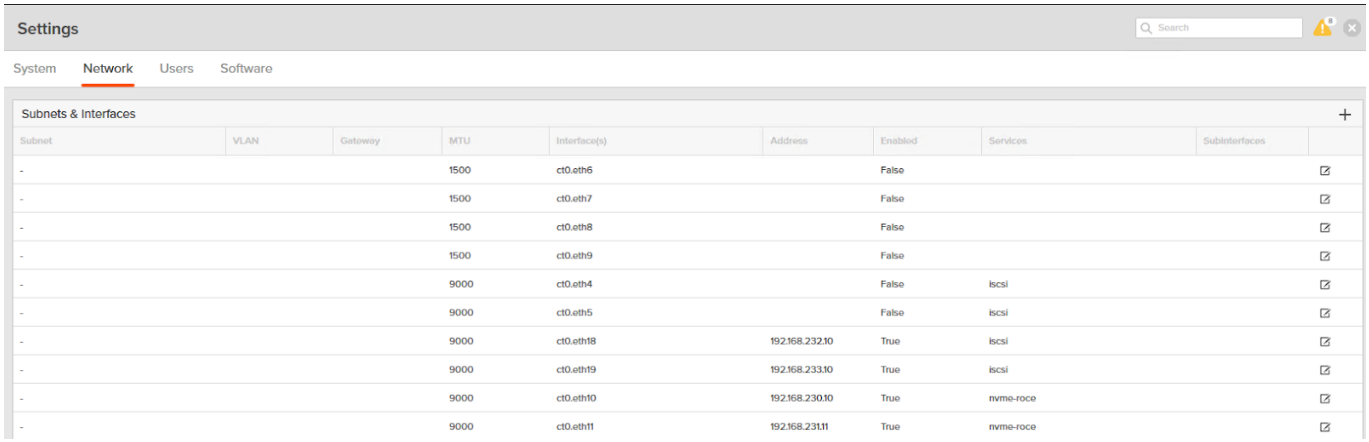
Run the following command to generate the NVMe qualified name(NQN) and retrieve it for later use. An NQN serves the same purpose as an internet qualified name (IQN) for iSCSI or world wide name(WWN) for Fiber channel.
[crayon-642795b0d71b3656700298/]
Navigate to the Storage View and in the hosts tab create a host. Once this host is created navigate to its management view and in the hosts ports section select the three vertical ellipses and select “Configure NQNs…”



Using the output from cat /etc /nvme/hostnqn copy this value into the dialog and press Add.

Connect the required volumes to this host.

Note the NVMe-roce ports and IP addresses to connect to in the Settings view under the network tab. NVMeoF RoCE support and service configuration on the FlashArray//X needs to be completed by Pure Support.



Step 6. Load the required NVMe kernel modules and connect the FlashArray volumes using RoCE.

First load nvme-core and nvme-rdma :
[crayon-642795b0d71b5125121121/]
Then discover the NQN for the NVMeoF target at the NVMe-roce ports noted in the FlashArray GUI.
[crayon-642795b0d71b9211830566/]
Take note of the subnqn in the returned text as this is used to :
[crayon-642795b0d71bb156486173/]
For each port to connect to on the FlashArray run the following to connect to all volumes for the relevant host via multiple paths:
[crayon-642795b0d71bc985633549/]
Ensure device-mapper multipath is enabled and check the devices which have been returned to it :
[crayon-642795b0d71be651082390/]
The devices connected will show up as below if configured correctly:
[crayon-642795b0d71bf709551924/]
Step 7. Set the best practice parameters for the NVMe-oF connected devices as set out in thisknowledge base article.

Create the file “/etc/udev/rules.d/90-pure.rules” and add the following lines before saving the file and running “udevadm control –reload-rules” :
[crayon-642795b0d71c0479738222/]
And that is it! Now the devices can be mounted and use the same as any other, with the added benefit of lower latency, comprehensive data services and management tools offered by FlashArray™.

Additional Resources:

FlashArray Product Features – NVMe

SUSE 15 Storage Administration Guide – NVMe-oF

Working with Source RPMs in SUSE

Blog Post : Pure brings hyperscale Architecture to the enterprise