NVMe-oF Host Configuration for RHEL 8.4 with ONTAP

SAN Host

Ranu Kundu
September 13, 2021

Table of Contents

NVMe-oF Host Configuration for RHEL 8.4 with ONTAP

Supportability

NVMe over Fabrics or NVMe-oF (including NVMe/FC and other transports) is supported with RHEL 8.4 with ANA (Asymmetric Namespace Access). ANA is the ALUA equivalent in the NVMe-oF environment, and is currently implemented with in-kernel NVMe Multipath. The details for enabling NVMe-oF with in-kernel NVMe Multipath using ANA on RHEL 8.4 and ONTAP as the target has been documented here.

Features

- Starting RHEL 8.2, `nvme-fc auto-connect` scripts are included in the native `nvme-cli` package. You can rely on these native auto-connect scripts instead of having to install the external vendor provided outbox auto-connect scripts.
- Starting RHEL 8.2, a native `udev` rule is already provided as part of the `nvme-cli` package which enables round-robin load balancing for NVMe multipath. You need not manually create this rule any more (as was done in RHEL 8.1).
- Starting RHEL 8.2, both NVMe and SCSI traffic can be run on the same co-existent host. In fact, that is expected to be the commonly deployed host config for customers. Therefore, for SCSI, you may configure `dm-multipath` as usual for SCSI LUNs resulting in `mpath` devices, whereas NVMe multipath may be used to configure NVMe-oF multipath devices on the host.
- Starting RHEL 8.2, the NetApp plugin in the native `nvme-cli` package is capable of displaying ONTAP details as well for ONTAP namespaces.

Known limitations

For RHEL 8.4, in-kernel NVMe multipath remains disabled by default. Therefore, you need to enable it manually.

Configuration requirements

Refer to the NetApp Interoperability Matrix for accurate details regarding supported configurations.

Enabling in-kernel NVMe Multipath

1. Install RHEL 8.4 GA on the server. After the installation is complete, verify that you are running the specified RHEL 8.4 GA kernel. See the NetApp Interoperability Matrix for the most current list of supported versions.
2. After the installation is complete, verify that you are running the specified RHEL 8.4 kernel. See the NetApp Interoperability Matrix for the most current list of supported versions.

   Example:

```
# uname -r
4.18.0-305.el8.x86_64
```

3. Install the `nvme-cli` package:

   Example:

```
# rpm -qa|grep nvme-cli
nvme-cli-1.12-3.el8.x86_64
```

4. Enable in-kernel NVMe multipath:

```
# grubby --args=nvme_core.multipath=Y --update-kernel /boot/vmlinuz-
4.18.0-305.el8.x86_64
```

5. On the host, check the host NQN string at `/etc/nvme/hostnqn` and verify that it matches the host NQN string for the corresponding subsystem on the ONTAP array. Example:

```
# cat /etc/nvme/hostnqn
nqn.2014-08.org.nvmexpress:uuid:9ed5b327-b9fc-4cf5-97b3-1b5d986345d1
::> vserver nvme subsystem host show -vserver vs_fcnvme_141
Vserver     Subsystem       Host NQN
----------- ---------------
------------------------------------------------------------
vs_fcnvme_14 nvme_141_1     nqn.2014-08.org.nvmexpress:uuid:9ed5b327-
b9fc-4cf5-97b3-1b5d986345d1
```

> If the host NQN strings do not match, you should use the `vserver modify` command to update the host NQN string on your corresponding ONTAP NVMe subsystem to match the host NQN string `/etc/nvme/hostnqn` on the host.

6. Reboot the host.

If you intend to run both NVMe & SCSI co-existent traffic on the same host, it is recommended to use in-kernel NVMe multipath for ONTAP namespaces and dm-multipath for ONTAP LUNs respectively. This means that the ONTAP namespaces should be excluded from dm-multipath to prevent dm-multipath from claiming these namespace devices. This can be done by adding the enable_foreign setting to the `/etc/multipath.conf` file:

```
# cat /etc/multipath.conf
defaults {
        enable_foreign     NONE
}
```

7. Restart the multipathd daemon by running a `systemctl restart multipathd` command to allow the new setting to take effect.

Configuring NVMe/FC

Broadcom/Emulex

1. Verify that you are using the supported adapter. See the NetApp Interoperability Matrix for the most current list of supported adapters.

```
# cat /sys/class/scsi_host/host*/modelname
LPe32002-M2
LPe32002-M2
# cat /sys/class/scsi_host/host*/modeldesc
Emulex LightPulse LPe32002-M2 2-Port 32Gb Fibre Channel Adapter
Emulex LightPulse LPe32002-M2 2-Port 32Gb Fibre Channel Adapter
```

2. Verify that you are using the recommended Broadcom lpfc firmware and inbox driver. See the NetApp Interoperability Matrix for the most current list of supported adapter driver and firmware versions.

```
# cat /sys/class/scsi_host/host*/fwrev
12.8.340.8, sli-4:2:c
12.8.340.8, sli-4:2:c
# cat /sys/module/lpfc/version
0:12.8.0.5
```

3. Verify that lpfc_enable_fc4_type is set to 3

```
# cat /sys/module/lpfc/parameters/lpfc_enable_fc4_type
3
```

4. Verify that the initiator ports are up and running, and you are able to see the target LIFs.

```
# cat /sys/class/fc_host/host*/port_name
0x100000109b1c1204
0x100000109b1c1205
# cat /sys/class/fc_host/host*/port_state
Online
Online
# cat /sys/class/scsi_host/host*/nvme_info
NVME Initiator Enabled
XRI Dist lpfc0 Total 6144 IO 5894 ELS 250
NVME LPORT lpfc0 WWPN x100000109b1c1204 WWNN x200000109b1c1204 DID
x011d00 ONLINE
NVME RPORT WWPN x203800a098dfdd91 WWNN x203700a098dfdd91 DID x010c07
TARGET DISCSRVC ONLINE
NVME RPORT WWPN x203900a098dfdd91 WWNN x203700a098dfdd91 DID x011507
TARGET DISCSRVC ONLINE
NVME Statistics
LS: Xmt 0000000f78 Cmpl 0000000f78 Abort 00000000
LS XMIT: Err 00000000 CMPL: xb 00000000 Err 00000000
Total FCP Cmpl 000000002fe29bba Issue 000000002fe29bc4 OutIO
000000000000000a
abort 00001bc7 noxri 00000000 nondlp 00000000 qdepth 00000000 wqerr
00000000 err 00000000
FCP CMPL: xb 00001e15 Err 0000d906
NVME Initiator Enabled
XRI Dist lpfc1 Total 6144 IO 5894 ELS 250
NVME LPORT lpfc1 WWPN x100000109b1c1205 WWNN x200000109b1c1205 DID
x011900 ONLINE
NVME RPORT WWPN x203d00a098dfdd91 WWNN x203700a098dfdd91 DID x010007
TARGET DISCSRVC ONLINE
NVME RPORT WWPN x203a00a098dfdd91 WWNN x203700a098dfdd91 DID x012a07
TARGET DISCSRVC ONLINE
NVME Statistics
LS: Xmt 0000000fa8 Cmpl 0000000fa8 Abort 00000000
LS XMIT: Err 00000000 CMPL: xb 00000000 Err 00000000
Total FCP Cmpl 000000002e14f170 Issue 000000002e14f17a OutIO
000000000000000a
abort 000016bb noxri 00000000 nondlp 00000000 qdepth 00000000 wqerr
00000000 err 00000000
FCP CMPL: xb 00001f50 Err 0000d9f8
```

Enabling 1MB I/O size (Optional)

ONTAP reports an MDTS (Max Data Transfer Size) of 8 in the Identify Controller data which means the maximum I/O request size should be up to 1 MB. However, to issue I/O requests of size 1 MB for the Broadcom NVMe/FC host, the lpfc parameter `lpfc_sg_seg_cnt` should also be bumped up to 256 from the

default value of 64. Use the following instructions to do so:

1. Append the value `256` in the respective `modprobe lpfc.conf` file:

```
# cat /etc/modprobe.d/lpfc.conf
options lpfc lpfc_sg_seg_cnt=256
```

2. Run a `dracut -f` command, and reboot the host.
3. After reboot, verify that the above setting has been applied by checking the corresponding `sysfs` value:

```
# cat /sys/module/lpfc/parameters/lpfc_sg_seg_cnt
256
```

Now the Broadcom FC-NVMe host should be able to send up to 1MB I/O requests on the ONTAP namespace devices.

Marvell/QLogic

The native inbox `qla2xxx` driver included in the RHEL 8.4 GA kernel has the latest upstream fixes which are essential for ONTAP support.

- Verify that you are running the supported adapter driver and firmware versions using the following command:

```
# cat /sys/class/fc_host/host*/symbolic_name
QLE2742 FW:v9.06.02 DVR:v10.02.00.104-k
QLE2742 FW:v9.06.02 DVR:v10.02.00.104-k
```

- Verify `ql2xnvmeenable` is set which enables the Marvell adapter to function as a NVMe/FC initiator using the following command:

```
# cat /sys/module/qla2xxx/parameters/ql2xnvmeenable
1
```

Validating NVMe-oF

- Verify that in-kernel NVMe multipath is enabled:

```
# cat /sys/module/nvme_core/parameters/multipath
Y
```

- Verify the appropriate NVMe-oF settings, `model` set to `NetApp ONTAP Controller` and `load`

`balancing iopolicy` set to `round-robin`, so that the respective ONTAP namespaces properly reflect on the host:

```
# cat /sys/class/nvme-subsystem/nvme-subsys*/model
NetApp ONTAP Controller
NetApp ONTAP Controller
# cat /sys/class/nvme-subsystem/nvme-subsys*/iopolicy
round-robin
round-robin
```

NVMe/FC

1. Verify that the NVMe/FC ONTAP namespaces properly reflect on the host.

   Example:

```
# nvme list
Node         SN                   Model                    Namespace
Usage
-------      -----------------    -------------            ---------
------------
/dev/nvme0n1 814vWBNRwf9HAAAAAAAB  NetApp ONTAP Controller   1
85.90 GB / 85.90 GB

Format       FW Rev
------       -------
4 KiB + 0 B  FFFFFFFF
```

2. Verify that the controller state of each path is live and has proper ANA status.

   Example:

```
# nvme list-subsys /dev/nvme0n1
nvme-subsys0 - NQN=nqn.1992-
08.com.netapp:sn.5f5f2c4aa73b11e9967e00a098df41bd:subsystem.nvme_141_1
\
+- nvme0 fc traddr=nn-0x203700a098dfdd91:pn-0x203800a098dfdd91
host_traddr=nn-0x200000109b1c1204:pn-0x100000109b1c1204 live
inaccessible
+- nvme1 fc traddr=nn-0x203700a098dfdd91:pn-0x203900a098dfdd91
host_traddr=nn-0x200000109b1c1204:pn-0x100000109b1c1204 live
inaccessible
+- nvme2 fc traddr=nn-0x203700a098dfdd91:pn-0x203a00a098dfdd91
host_traddr=nn-0x200000109b1c1205:pn-0x100000109b1c1205 live optimized
+- nvme3 fc traddr=nn-0x203700a098dfdd91:pn-0x203d00a098dfdd91
host_traddr=nn-0x200000109b1c1205:pn-0x100000109b1c1205 live optimized
```

3. Verify the NetApp plug-in displays proper values for each ONTAP namespace device.

   Example:

```
# nvme netapp ontapdevices -o column
Device        Vserver         Namespace Path
NSID
--------     -----------      --------------------------
-------
/dev/nvme0n1  vs_fcnvme_141   /vol/fcnvme_141_vol_1_1_0/fcnvme_141_ns
1
UUID                                    Size
-----                                   --------
72b887b1-5fb6-47b8-be0b-33326e2542e2   85.90GB

# nvme netapp ontapdevices -o json
{
"ONTAPdevices" : [
    {
        "Device" : "/dev/nvme0n1",
        "Vserver" : "vs_fcnvme_141",
        "Namespace_Path" : "/vol/fcnvme_141_vol_1_1_0/fcnvme_141_ns",
        "NSID" : 1,
        "UUID" : "72b887b1-5fb6-47b8-be0b-33326e2542e2",
        "Size" : "85.90GB",
        "LBA_Data_Size" : 4096,
        "Namespace_Size" : 20971520
    }
  ]
}
```

Troubleshooting

Before commencing any troubleshooting for any NVMe/FC failures, ensure that you are running a configuration that is compliant to the IMT specifications and then proceed with the next steps to debug any host side issues.

LPFC Verbose Logging

1. You can set the `lpfc_log_verbose` driver setting to any of the following values to log NVMe/FC events:

```
#define LOG_NVME 0x00100000 /* NVME general events. */
#define LOG_NVME_DISC 0x00200000 /* NVME Discovery/Connect events. */
#define LOG_NVME_ABTS 0x00400000 /* NVME ABTS events. */
#define LOG_NVME_IOERR 0x00800000 /* NVME IO Error events. */
```

2. After setting any of these values, run `dracut-f` command to recreate the `initramfs` and reboot the host.

3. After rebooting, verify the settings:

```
# cat /etc/modprobe.d/lpfc.conf
options lpfc lpfc_log_verbose=0xf00083

# cat /sys/module/lpfc/parameters/lpfc_log_verbose
15728771
```

qla2xxx Verbose Logging

There is no similar specific qla2xxx logging for NVMe/FC as for lpfc driver. Therefore, you may set the general qla2xxx logging level using the following steps:

1. Append the `ql2xextended_error_logging=0x1e400000` value to the corresponding `modprobe qla2xxx conf` file.
2. Recreate the `initramfs` by running `dracut -f` command and then reboot the host.
3. After reboot, verify that the verbose logging has been applied as follows:

```
# cat /etc/modprobe.d/qla2xxx.conf
options qla2xxx ql2xnvmeenable=1 ql2xextended_error_logging=0x1e400000
# cat /sys/module/qla2xxx/parameters/ql2xextended_error_logging
507510784
```

Common nvme-cli Errors and Workarounds

The errors displayed by `nvme-cli` during nvme discover, nvme connect or nvme connect-all operations and the workarounds are shown in the following table:

Errors displayed by `nvme-cli`	**Probable cause**	**Workaround**
`Failed to write to /dev/nvme-fabrics: Invalid argument`	Incorrect syntax	Ensure you are using the correct syntax for the above nvme commands.

Errors displayed by `nvme-cli`	Probable cause	Workaround

`Failed to write to /dev/nvme-fabrics: No such file or directory`	Multiple issues could trigger this.
	Passing wrong arguments to the nvme commands is one of the common causes.
		• Ensure you have passed the correct arguments (such as, correct WWNN string, WWPN string, and more) to the commands.
		• If the arguments are correct, but you still see this error, check if the `/sys/class/scsi_host/host*/nvme_info` output is proper, the NVMe initiator showing as `Enabled`, and the NVMe/FC target LIFs properly showing up here under the remote ports sections.
		Example:
		`# cat /sys/class/scsi_host/host*/nvme_info`
		`NVME Initiator Enabled`
		`NVME LPORT lpfc0 WWPN x10000090fae0ec9d WWNN x20000090fae0ec9d DID x012000 ONLINE`
		`NVME RPORT WWPN x200b00a098c80f09 WWNN x200a00a098c80f09 DID x010601 TARGET DISCSRVC ONLINE`
		`NVME Statistics`
		`LS: Xmt 0000000000000006 Cmpl 0000000000000006`
		`FCP: Rd 0000000000000071 Wr 0000000000000005 IO 0000000000000031 Cmpl 00000000000000a6 Outstanding 0000000000000001`
		`NVME Initiator Enabled`
		`NVME LPORT lpfc1 WWPN x10000090fae0ec9e WWNN x20000090fae0ec9e DID x012400 ONLINE`
		`NVME RPORT WWPN x200900a098c80f09 WWNN x200800a098c80f09 DID x010301 TARGET DISCSRVC ONLINE`
		`NVME Statistics`
		`LS: Xmt 0000000000000006 Cmpl 0000000000000006`
		`FCP: Rd 0000000000000073 Wr 0000000000000005 IO 0000000000000031 Cmpl 00000000000000a8 Outstanding 0000000000000001`
		• If the target LIFs don’t show up as above in the nvme_info output, check the `/var/log/messages` and `dmesg` output for any suspicious NVMe/FC failures, and report or fix accordingly.

Errors displayed by `nvme-cli`	Probable cause	Workaround
`No discovery log entries to fetch`	Generally seen if the `/etc/nvme/hostnqn` string has not been added to the corresponding subsystem on the NetApp array or an incorrect hostnqn string has been added to the respective subsystem.	Ensure the exact `/etc/nvme/hostnqn` string is added to the corresponding subsystem on the NetApp array (verify through the `vserver nvme subsystem host show` command).
`Failed to write to /dev/nvme-fabrics: Operation already in progress`	Seen if the controller associations or specified operation is already created or in the process of being created. This could happen as part of the auto-connect scripts installed above.	None. For nvme discover, try running this command after some time. For nvme connect and connect-all, run `nvme list` command to verify that the namespace devices are already created and displayed on the host.

When to contact technical support

If you are still facing issues, please collect the following files and command outputs and contact technical support for further triage:

```
cat /sys/class/scsi_host/host*/nvme_info
/var/log/messages
dmesg
nvme discover output as in:
nvme discover --transport=fc --traddr=nn-0x200a00a098c80f09:pn
-0x200b00a098c80f09 --host-traddr=nn-0x20000090fae0ec9d:pn
-0x10000090fae0ec9d
nvme list
nvme list-subsys /dev/nvmeXnY
```