Manage NFS over RDMA

ONTAP 9

NetApp
January 25, 2022

Table of Contents

Manage NFS over RDMA

NFS over RDMA

NFS over RDMA utilizes RDMA adapters, allowing data to be copied directly between storage system memory and host system memory, circumventing CPU interruptions and overhead.

NFS over RDMA configurations are designed for customers with latency sensitive or high-bandwidth workloads such as machine learning and analytics. NVIDIA has extended NFS over RDMA to enable GPU Direct Storage (GDS). GDS further accelerates GPU-enabled workloads by bypassing the CPU and main memory altogether using RDMA to transfer data between the storage system and GPU memory directly.

In ONTAP 9.10.1, this configuration is only supported for the NFSv4.0 protocol when used with the Mellanox CX-5 or CX-6 adapter, which provides support for RDMA using version 2 of the RoCE protocol. GDS is only supported using NVIDIA Tesla- and Ampere-family GPUs with Mellanox NIC cards and MOFED software. NFS over RDMA support is limited to node-local traffic only. Standard FlexVols or FlexGroups where all constituents are on the same node are supported and must be accessed from a LIF on the same node.

Requirements

- Storages systems must be running ONTAP 9.10.1
- Both nodes in the HA pair must be the same version
- Storage system controllers must have RDMA support (currently A400, A700, and A800)
- Storage appliance configured with RDMA-supported hardware (e.g. Mellanox CX-5 or CX-6)
- Data LIFs must be configured to support RDMA.
- Clients must be using Mellanox RDMA-capable NIC cards and Mellanox OFED (MOFED) network software.

Next Steps

- Configure NICs for NFS over RDMA
- Configure LIFs for NFS over RDMA
- NFS settings for NFS over RDMA

Further reading

- RFC 7530: NFS Version 4 Protocol
- RFC 8166: Remote Direct Memory Access Transport for Remote Procedure Call Version 1
- RFC 8167: Bidirectional Remote Procedure Call on RPC-over-RDMA Transports
- RFC 8267: NFS Upper-Layer Binding to RPC-over-RDMA version 1

Configure NICs for NFS over RDMA

NFS over RDMA requires NIC configuration for both the client system and storage platform.

Storage platform configuration

An X1148 RDMA adapter needs to be installed on the server. If you are using an HA configuration, you must have a corresponding X1148 adapter on the failover partner so RDMA service can continue during failover. The NIC must be ROCE capable.

Beginning with ONTAP 9.10.1, you can view a list of RDMA offload protocols with the command: `network port show -rdma-protocols roce`

Client system configuration

Clients must be using Mellanox RDMA-capable NIC cards (e.g. X1148) and Mellanox OFED network software. Consult Mellanox documentation for supported models and versions. Although the client and server can be directly connected, the use of switches is recommended due to improved failover performance with a switch.

The client, server, and any switches, and all ports on switches must be configured using Jumbo frames. Also ensure that priority flow-control is in effect on any switches.

Once this configuration is confirmed, you can mount the NFS.

Steps

1. Check if RDMA access is enabled on the NFS server with the command:

   `vserver nfs show-vserver` *`vserver_name`*

   By default, `-rdma` should be enabled. If it is not, enable RDMA access on the NFS server:

   `vserver nfs modify -vserver` *`vserver_name`* `-rdma enabled`

2. Mount the client via NFSv4.0 over RDMA:
   a. The input for the proto parameter depends on the server IP protocol version. If it is v4, use `proto=rdma`. If it is v6, use `proto=rdma6`.
   b. Specify the NFS target port as `port=20049` instead of the standard port of 2049:

      `mount -o vers=4,minorversion=0,proto=rdma,port=20049` *`Server_IP_address:/volume_path mount_point`*

3. **OPTIONAL**: If you need to unmount the client, run the command `unmount` *`mount_path`*

Configure LIFs for NFS over RDMA

To utilize NFS over RDMA, you must configure your LIFs to be RDMA compatible. Both the LIF and its failover pair must be RDMA capable.

Beginning with ONTAP 9.10.1, there is a new rdma_protocols parameter on the network interface commands where you can configure support, along with setting an appropriate service policy.

Procedure to create a new LIF

1. Create a LIF:

   `network interface create -vserver` *`vserver_name`* `-lif` *`lif_name`* `-service-policy`

```
service_policy_name -home-node node_name -home-port port_name {-address
IP_address -netmask netmask_value | -subnet-name subnet_name} -firewall-policy
policy_name -auto-revert {true|false} -rdma-protocols roce
```

a. The service policy must be either default-data-files or a custom policy that includes the data-nfs network interface service.

b. The `-rdma-protocols` parameter accepts a list, which is by default empty. When `roce` is added as a value, the LIF can only be configured on ports supporting RoCE offload, affecting bot LIF migration and failover.

Procedure to modify a LIF

1. You can check the status of your LIFs with the `network interface show` command. The service policy must include the data-nfs network interface service. The `-rdma-protocols` list should include `roce`. If either of these conditions are untrue, modify the LIF.
2. To modify the LIF, run:

```
network interface modify vserver vserver_name -lif lif_name -service-policy
service_policy_name -home-node node_name -home-port port_name {-address
IP_address -netmask netmask_value | -subnet-name subnet_name} -firewall-policy
policy_name -auto-revert {true|false} -rdma-protocols roce
```

Modifying a LIF to require a particular offload protocol when the LIF is not currently assigned to a port that supports that protocol will produce an error.

More Information

Create a LIF

What LIFs are

Modify the NFS configuration

In most cases, you will not need to modify the configuration of the NFS-enabled vserver for NFS over RDMA.

If you are, however, dealing with issues related to Mellanox chips and LIF migration, you should increase the NFSv4 locking grace period. By default, the grace period is set to 45 seconds. Beginning with ONTAP 9.10.1, the grace period can go as high as 180 seconds.

Steps

1. Set the privilege level to advanced:

```
set -privilege advanced
```

2. Enter the following command:

```
vserver nfs modify -vserver vserver_name -v4-grace-seconds number_of_seconds
```

For more about this task, see Specifying the NFSv4 locking grace period.