Setting up NFS/RDMA

Author: NetApp and Open Grid Computing (May 29, 2008)

> **Warning**
> This document is probably obsolete.

Overview

This document describes how to install and setup the Linux NFS/RDMA client and server software.

The NFS/RDMA client was first included in Linux 2.6.24. The NFS/RDMA server was first included in the following release, Linux 2.6.25.

In our testing, we have obtained excellent performance results (full 10Gbit wire bandwidth at minimal client CPU) under many workloads. The code passes the full Connectathon test suite and operates over both Infiniband and iWARP RDMA adapters.

Getting Help

If you get stuck, you can ask questions on the nfs-rdma-devel@lists.sourceforge.net mailing list.

Installation

These instructions are a step by step guide to building a machine for use with NFS/RDMA.

- Install an RDMA device

  Any device supported by the drivers in drivers/infiniband/hw is acceptable.

  Testing has been performed using several Mellanox-based IB cards, the Ammasso AMS1100 iWARP adapter, and the Chelsio cxgb3 iWARP adapter.

- Install a Linux distribution and tools

  The first kernel release to contain both the NFS/RDMA client and server was Linux 2.6.25 Therefore, a distribution compatible with this and subsequent Linux kernel release should be installed.

  The procedures described in this document have been tested with distributions from Red Hat's Fedora Project (http://fedora.redhat.com/).

- Install nfs-utils-1.1.2 or greater on the client

  An NFS/RDMA mount point can be obtained by using the mount.nfs command in nfs-utils-1.1.2 or greater (nfs-utils-1.1.1 was the first nfs-utils version with support for NFS/RDMA mounts, but for various reasons we recommend using nfs-utils-1.1.2 or greater). To see which version of mount.nfs you are using, type:

  ```
  $ /sbin/mount.nfs -V
  ```

  If the version is less than 1.1.2 or the command does not exist, you should install the latest version of nfs-utils.

  Download the latest package from: https://www.kernel.org/pub/linux/utils/nfs

  Uncompress the package and follow the installation instructions.

  If you will not need the idmapper and gssd executables (you do not need these to create an NFS/RDMA enabled mount command), the installation process can be simplified by disabling these features when running configure:

  ```
  $ ./configure --disable-gss --disable-nfsv4
  ```

  To build nfs-utils you will need the tcp_wrappers package installed. For more information on this see the package's README and INSTALL files.

  After building the nfs-utils package, there will be a mount.nfs binary in the utils/mount directory. This binary can be used to initiate NFS v2, v3, or v4 mounts. To initiate a v4 mount, the binary must be called mount.nfs4. The standard technique is to create a symlink called mount.nfs4 to mount.nfs.

  This mount.nfs binary should be installed at /sbin/mount.nfs as follows:

  ```
  $ sudo cp utils/mount/mount.nfs /sbin/mount.nfs
  ```

  In this location, mount.nfs will be invoked automatically for NFS mounts by the system mount command.

> **Note**
>
> mount.nfs and therefore nfs-utils-1.1.2 or greater is only needed on the NFS client machine. You do not need this specific version of nfs-utils on the server. Furthermore, only the mount.nfs command from nfs-utils-1.1.2 is needed on the client.

- Install a Linux kernel with NFS/RDMA

  The NFS/RDMA client and server are both included in the mainline Linux kernel version 2.6.25 and later. This and other versions of the Linux kernel can be found at: https://www.kernel.org/pub/linux/kernel/

  Download the sources and place them in an appropriate location.

- Configure the RDMA stack

  Make sure your kernel configuration has RDMA support enabled. Under Device Drivers -> InfiniBand support, update the kernel configuration to enable InfiniBand support [NOTE: the option name is misleading. Enabling InfiniBand support is required for all RDMA devices (IB, iWARP, etc.)].

  Enable the appropriate IB HCA support (mlx4, mthca, ehca, ipath, etc.) or iWARP adapter support (amso, cxgb3, etc.).

  If you are using InfiniBand, be sure to enable IP-over-InfiniBand support.

- Configure the NFS client and server

  Your kernel configuration must also have NFS file system support and/or NFS server support enabled. These and other NFS related configuration options can be found under File Systems -> Network File Systems.

- Build, install, reboot

  The NFS/RDMA code will be enabled automatically if NFS and RDMA are turned on. The NFS/RDMA client and server are configured via the hidden SUNRPC_XPRT_RDMA config option that depends on SUNRPC and INFINIBAND. The value of SUNRPC_XPRT_RDMA will be:

  1. N if either SUNRPC or INFINIBAND are N, in this case the NFS/RDMA client and server will not be built
  2. M if both SUNRPC and INFINIBAND are on (M or Y) and at least one is M, in this case the NFS/RDMA client and server will be built as modules
  3. Y if both SUNRPC and INFINIBAND are Y, in this case the NFS/RDMA client and server will be built into the kernel

  Therefore, if you have followed the steps above and turned no NFS and RDMA, the NFS/RDMA client and server will be built.

  Build a new kernel, install it, boot it.

Check RDMA and NFS Setup

Before configuring the NFS/RDMA software, it is a good idea to test your new kernel to ensure that the kernel is working correctly. In particular, it is a good idea to verify that the RDMA stack is functioning as expected and standard NFS over TCP/IP and/or UDP/IP is working properly.

- Check RDMA Setup

  If you built the RDMA components as modules, load them at this time. For example, if you are using a Mellanox Tavor/Sinai/Arbel card:

  ```
  $ modprobe ib_mthca
  $ modprobe ib_ipoib
  ```

  If you are using InfiniBand, make sure there is a Subnet Manager (SM) running on the network. If your IB switch has an embedded SM, you can use it. Otherwise, you will need to run an SM, such as OpenSM, on one of your end nodes.

  If an SM is running on your network, you should see the following:

  ```
  $ cat /sys/class/infiniband/driverX/ports/1/state
  4: ACTIVE
  ```

  where driverX is mthca0, ipath5, ehca3, etc.

  To further test the InfiniBand software stack, use IPoIB (this assumes you have two IB hosts named host1 and host2):

  ```
  host1$ ip link set dev ib0 up
  host1$ ip address add dev ib0 a.b.c.x
  host2$ ip link set dev ib0 up
  host2$ ip address add dev ib0 a.b.c.y
  host1$ ping a.b.c.y
  ```

```
host2$ ping a.b.c.x
```

For other device types, follow the appropriate procedures.

- Check NFS Setup

  For the NFS components enabled above (client and/or server), test their functionality over standard Ethernet using TCP/IP or UDP/IP.

NFS/RDMA Setup

We recommend that you use two machines, one to act as the client and one to act as the server.

One time configuration:

- On the server system, configure the /etc/exports file and start the NFS/RDMA server.

  Exports entries with the following formats have been tested:

```
/vol0   192.168.0.47(fsid=0,rw,async,insecure,no_root_squash)
/vol0   192.168.0.0/255.255.255.0(fsid=0,rw,async,insecure,no_root_squash)
```

  The IP address(es) is(are) the client's IPoIB address for an InfiniBand HCA or the client's iWARP address(es) for an RNIC.

> **Note**
>
> The "insecure" option must be used because the NFS/RDMA client does not use a reserved port.

Each time a machine boots:

- Load and configure the RDMA drivers

  For InfiniBand using a Mellanox adapter:

```
$ modprobe ib_mthca
$ modprobe ib_ipoib
$ ip li set dev ib0 up
$ ip addr add dev ib0 a.b.c.d
```

> **Note**
>
> Please use unique addresses for the client and server!

- Start the NFS server

  If the NFS/RDMA server was built as a module (CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the RDMA transport module:

```
$ modprobe svcrdma
```

  Regardless of how the server was built (module or built-in), start the server:

```
$ /etc/init.d/nfs start
```

  or

```
$ service nfs start
```

  Instruct the server to listen on the RDMA transport:

```
$ echo rdma 20049 > /proc/fs/nfsd/portlist
```

- On the client system

  If the NFS/RDMA client was built as a module (CONFIG_SUNRPC_XPRT_RDMA=m in kernel config), load the RDMA client module:

```
$ modprobe xprtrdma.ko
```

  Regardless of how the client was built (module or built-in), use this command to mount the NFS/RDMA server:

```
$ mount -o rdma,port=20049 <IPoIB-server-name-or-address>:/<export> /mnt
```

  To verify that the mount is using RDMA, run "cat /proc/mounts" and check the "proto" field for the given mount.

  Congratulations! You're using NFS/RDMA!