



Red Hat Enterprise Linux 8

Configuring InfiniBand and RDMA networks

Configuring and managing high-speed network protocols and RDMA hardware

Red Hat Enterprise Linux 8 Configuring InfiniBand and RDMA networks

Configuring and managing high-speed network protocols and RDMA hardware

Legal Notice

Copyright © 2023 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

<http://creativecommons.org/licenses/by-sa/3.0/>

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux[®] is the registered trademark of Linus Torvalds in the United States and other countries.

Java[®] is a registered trademark of Oracle and/or its affiliates.

XFS[®] is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL[®] is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js[®] is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack[®] Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

Abstract

You can configure and manage Remote Directory Memory Access (RDMA) networks and InfiniBand hardware at an enterprise level by using various protocols. These include RDMA over Converged Ethernet (RoCE), the software implementation of RoCE (Soft-RoCE), the IP networks protocol such as iWARP, the software implementation of iWARP (Soft-iWARP), and the Network File System over RDMA (NFSoverRDMA) protocol as a native support on RDMA-supported hardware. For low-latency and high-throughput connections, you can configure IP over InfiniBand (IPoIB) and Open Subnet Manager (OpenSM).

Table of Contents

MAKING OPEN SOURCE MORE INCLUSIVE	3
PROVIDING FEEDBACK ON RED HAT DOCUMENTATION	4
CHAPTER 1. UNDERSTANDING INFINIBAND AND RDMA	5
CHAPTER 2. CONFIGURING ROCE	6
2.1. OVERVIEW OF ROCE PROTOCOL VERSIONS	6
2.2. TEMPORARILY CHANGING THE DEFAULT ROCE VERSION	6
2.3. CONFIGURING SOFT-ROCE	7
CHAPTER 3. CONFIGURING SOFT-IWARP	9
3.1. OVERVIEW OF IWARP AND SOFT-IWARP	9
3.2. CONFIGURING SOFT-IWARP	9
CHAPTER 4. CONFIGURING THE CORE RDMA SUBSYSTEM	11
4.1. RENAMING IPOIB DEVICES	11
4.2. INCREASING THE AMOUNT OF MEMORY THAT USERS ARE ALLOWED TO PIN IN THE SYSTEM	11
4.3. CONFIGURING THE RDMA SERVICE	12
4.4. ENABLING NFS OVER RDMA (NFSORDMA)	13
CHAPTER 5. CONFIGURING AN INFINIBAND SUBNET MANAGER	15
5.1. INSTALLING THE OPENSMB SUBNET MANAGER	15
5.2. CONFIGURING OPENSMB USING THE SIMPLE METHOD	15
5.3. CONFIGURING OPENSMB BY EDITING THE OPENSMB.CONF FILE	16
5.4. CONFIGURING MULTIPLE OPENSMB INSTANCES	17
5.5. CREATING A PARTITION CONFIGURATION	18
CHAPTER 6. CONFIGURING IPOIB	20
6.1. THE IPOIB COMMUNICATION MODES	20
6.2. UNDERSTANDING IPOIB HARDWARE ADDRESSES	20
6.3. CONFIGURING AN IPOIB CONNECTION USING NMCLI COMMANDS	21
6.4. CONFIGURING AN IPOIB CONNECTION BY USING THE NETWORK RHEL SYSTEM ROLE	21
6.5. CONFIGURING AN IPOIB CONNECTION USING NM-CONNECTION-EDITOR	23
CHAPTER 7. TESTING INFINIBAND NETWORKS	26
7.1. TESTING EARLY INFINIBAND RDMA OPERATIONS	26
7.2. TESTING AN IPOIB USING THE PING UTILITY	28
7.3. TESTING AN RDMA NETWORK USING QPERF AFTER IPOIB IS CONFIGURED	28

MAKING OPEN SOURCE MORE INCLUSIVE

Red Hat is committed to replacing problematic language in our code, documentation, and web properties. We are beginning with these four terms: master, slave, blacklist, and whitelist. Because of the enormity of this endeavor, these changes will be implemented gradually over several upcoming releases. For more details, see [our CTO Chris Wright's message](#).

PROVIDING FEEDBACK ON RED HAT DOCUMENTATION

We appreciate your feedback on our documentation. Let us know how we can improve it.

Submitting comments on specific passages

1. View the documentation in the **Multi-page HTML** format and ensure that you see the **Feedback** button in the upper right corner after the page fully loads.
2. Use your cursor to highlight the part of the text that you want to comment on.
3. Click the **Add Feedback** button that appears near the highlighted text.
4. Add your feedback and click **Submit**.

Submitting feedback through Jira (account required)

1. Log in to the [Jira](#) website.
2. Click **Create** in the top navigation bar
3. Enter a descriptive title in the **Summary** field.
4. Enter your suggestion for improvement in the **Description** field. Include links to the relevant parts of the documentation.
5. Click **Create** at the bottom of the dialogue.

CHAPTER 1. UNDERSTANDING INFINIBAND AND RDMA

InfiniBand refers to two distinct things:

- The physical link-layer protocol for InfiniBand networks
- The InfiniBand Verbs API, an implementation of the remote direct memory access (RDMA) technology

RDMA provides access between the main memory of two computers without involving an operating system, cache, or storage. Using RDMA, data transfers with high-throughput, low-latency, and low CPU utilization.

In a typical IP data transfer, when an application on one machine sends data to an application on another machine, the following actions happen on the receiving end:

1. The kernel must receive the data.
2. The kernel must determine that the data belongs to the application.
3. The kernel wakes up the application.
4. The kernel waits for the application to perform a system call into the kernel.
5. The application copies the data from the internal memory space of the kernel into the buffer provided by the application.

This process means that most network traffic is copied across the main memory of the system if the host adapter uses direct memory access (DMA) or otherwise at least twice. Additionally, the computer executes some context switches to switch between the kernel and application. These context switches can cause a higher CPU load with high traffic rates while slowing down the other tasks.

Unlike traditional IP communication, RDMA communication bypasses the kernel intervention in the communication process. This reduces the CPU overhead. The RDMA protocol enables the host adapter to decide after a packet enters the network which application should receive it and where to store it in the memory space of that application. Instead of sending the packet for processing to the kernel and copying it into the memory of the user application, the host adapter directly places the packet contents in the application buffer. This process requires a separate API, the InfiniBand Verbs API, and applications need to implement the InfiniBand Verbs API to use RDMA.

Red Hat Enterprise Linux supports both the InfiniBand hardware and the InfiniBand Verbs API. Additionally, it supports the following technologies to use the InfiniBand Verbs API on non-InfiniBand hardware:

- Internet Wide Area RDMA Protocol (iWARP): A network protocol that implements RDMA over IP networks
- RDMA over Converged Ethernet (RoCE), which is also known as InfiniBand over Ethernet (IBoE): A network protocol that implements RDMA over Ethernet networks

Additional resources

- [Configuring RoCE](#)

CHAPTER 2. CONFIGURING ROCE

Remote Direct Memory Access (RDMA) provides remote execution for Direct Memory Access (DMA). RDMA over Converged Ethernet (RoCE) is a network protocol that utilizes RDMA over an Ethernet network. For configuration, RoCE requires specific hardware and some of the hardware vendors are Mellanox, Broadcom, and QLogic.

2.1. OVERVIEW OF ROCE PROTOCOL VERSIONS

RoCE is a network protocol that enables remote direct memory access (RDMA) over Ethernet.

The following are the different RoCE versions:

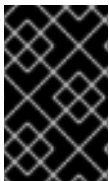
RoCE v1

The RoCE version 1 protocol is an Ethernet link layer protocol with ethertype **0x8915** that enables the communication between any two hosts in the same Ethernet broadcast domain.

RoCE v2

The RoCE version 2 protocol exists on the top of either the UDP over IPv4 or the UDP over IPv6 protocol. For RoCE v2, the UDP destination port number is **4791**.

The RDMA_CM sets up a reliable connection between a client and a server for transferring data. RDMA_CM provides an RDMA transport-neutral interface for establishing connections. The communication uses a specific RDMA device and message-based data transfers.



IMPORTANT

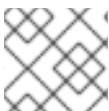
Using different versions like RoCE v2 on the client and RoCE v1 on the server is not supported. In such a case, configure both the server and client to communicate over RoCE v1.

Additional resources

- [Temporarily changing the default RoCE version](#)

2.2. TEMPORARILY CHANGING THE DEFAULT ROCE VERSION

Using the RoCE v2 protocol on the client and RoCE v1 on the server is not supported. If the hardware in your server supports RoCE v1 only, configure your clients for RoCE v1 to communicate with the server. For example, you can configure a client that uses the **mlx5_0** driver for the Mellanox ConnectX-5 InfiniBand device that only supports RoCE v1.



NOTE

Changes described here will remain effective until you reboot the host.

Prerequisites

- The client uses an InfiniBand device with RoCE v2 protocol.
- The server uses an InfiniBand device that only supports RoCE v1.

Procedure

1. Create the `/sys/kernel/config/rdma_cm/mlx5_0/` directory:

```
# mkdir /sys/kernel/config/rdma_cm/mlx5_0/
```

2. Display the default RoCE mode:

```
# cat /sys/kernel/config/rdma_cm/mlx5_0/ports/1/default_roce_mode
```

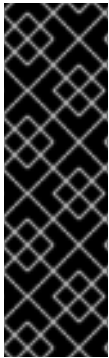
```
RoCE v2
```

3. Change the default RoCE mode to version 1:

```
# echo "IB/RoCE v1" > /sys/kernel/config/rdma_cm/mlx5_0/ports/1/default_roce_mode
```

2.3. CONFIGURING SOFT-ROCE

Soft-RoCE is a software implementation of remote direct memory access (RDMA) over Ethernet, which is also called RXE. Use Soft-RoCE on hosts without RoCE host channel adapters (HCA).



IMPORTANT

The Soft-RoCE feature is provided as a Technology Preview only. Technology Preview features are not supported with Red Hat production Service Level Agreements (SLAs), might not be functionally complete, and Red Hat does not recommend using them for production. These previews provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process.

See [Technology Preview Features Support Scope](#) on the Red Hat Customer Portal for information about the support scope for Technology Preview features.

Prerequisites

- An Ethernet adapter is installed

Procedure

1. Install the **iproute**, **libibverbs**, **libibverbs-utils**, and **infiniband-diags** packages:

```
# yum install iproute libibverbs libibverbs-utils infiniband-diags
```

2. Display the RDMA links:

```
# rdma link show
```

3. Load the **rdma_rxe** kernel module and add a new **rxe** device named **rxe0** that uses the **enp0s1** interface:

```
# rdma link add rxe0 type rxe netdev enp1s0
```

Verification

1. View the state of all RDMA links:

rdma link show

```
link rxe0/1 state ACTIVE physical_state LINK_UP netdev enp1s0
```

2. List the available RDMA devices:

ibv_devices

device	node GUID
-----	-----
rxex0	505400ffed5e0fb

3. You can use the **ibstat** utility to display a detailed status:

ibstat rxex0

```
CA 'rxex0'
CA type:
Number of ports: 1
Firmware version:
Hardware version:
Node GUID: 0x505400ffed5e0fb
System image GUID: 0x0000000000000000
Port 1:
  State: Active
  Physical state: LinkUp
  Rate: 100
  Base lid: 0
  LMC: 0
  SM lid: 0
  Capability mask: 0x00890000
  Port GUID: 0x505400ffed5e0fb
  Link layer: Ethernet
```

CHAPTER 3. CONFIGURING SOFT-IWARP

Remote Direct Memory Access (RDMA) uses several libraries and protocols over an Ethernet such as iWARP, Soft-iWARP for performance improvement and aided programming interface.

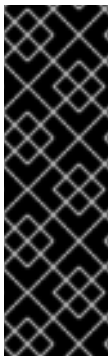
3.1. OVERVIEW OF IWARP AND SOFT-IWARP

Remote direct memory access (RDMA) uses the Internet Wide-area RDMA Protocol (iWARP) over Ethernet for converged and low latency data transmission over TCP. Using standard Ethernet switches and the TCP/IP stack, iWARP routes traffic across the IP subnets. This provides flexibility to efficiently use the existing infrastructure. In Red Hat Enterprise Linux, multiple providers implement iWARP in their hardware network interface cards. For example, **cxgb4**, **irdma**, **qedr** etc.

Soft-iWARP (siw) is a software-based iWARP kernel driver and user library for Linux. It is a software-based RDMA device that provides a programming interface to RDMA hardware when attached to network interface cards. It provides an easy way to test and validate the RDMA environment.

3.2. CONFIGURING SOFT-IWARP

Soft-iWARP (siw) implements the Internet Wide-area RDMA Protocol (iWARP) Remote direct memory access (RDMA) transport over the Linux TCP/IP network stack. It enables a system with a standard Ethernet adapter to interoperate with an iWARP adapter or with another system running the Soft-iWARP driver or a host with the hardware that supports iWARP.



IMPORTANT

The Soft-iWARP feature is provided as a Technology Preview only. Technology Preview features are not supported with Red Hat production Service Level Agreements (SLAs), might not be functionally complete, and Red Hat does not recommend using them for production. These previews provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process.

See [Technology Preview Features Support Scope](#) on the Red Hat Customer Portal for information about the support scope for Technology Preview features.

To configure Soft-iWARP, you can use this procedure in a script to run automatically when the system boots.

Prerequisites

- An Ethernet adapter is installed

Procedure

1. Install the **iproute**, **libibverbs**, **libibverbs-utils**, and **infiniband-diags** packages:

```
# yum install iproute libibverbs libibverbs-utils infiniband-diags
```

2. Display the RDMA links:

```
# rdma link show
```

3. Load the **siw** kernel module:

```
# modprobe siw
```

4. Add a new **siw** device named **siw0** that uses the **enp0s1** interface:

```
# rdma link add siw0 type siw netdev enp0s1
```

Verification

1. View the state of all RDMA links:

```
# rdma link show
```

```
link siw0/1 state ACTIVE physical_state LINK_UP netdev enp0s1
```

2. List the available RDMA devices:

```
# ibv_devices
```

device	node GUID
-----	-----
siw0	0250b6fffea19d61

3. You can use the **ibv_devinfo** utility to display a detailed status:

```
# ibv_devinfo siw0
```

```
hca_id:      siw0
transport:   iWARP (1)
fw_ver:      0.0.0
node_guid:    0250:b6ff:fea1:9d61
sys_image_guid: 0250:b6ff:fea1:9d61
vendor_id:    0x626d74
vendor_part_id: 1
hw_ver:      0x0
phys_port_cnt: 1
  port:      1
    state:    PORT_ACTIVE (4)
    max_mtu:  1024 (3)
    active_mtu: 1024 (3)
    sm_lid:    0
    port_lid:  0
    port_lmc:  0x00
    link_layer: Ethernet
```

CHAPTER 4. CONFIGURING THE CORE RDMA SUBSYSTEM

The **rdma** service configuration manages the network protocols and communication standards such as InfiniBand, iWARP, and RoCE.

4.1. RENAMING IPOIB DEVICES

By default, the kernel names Internet Protocol over InfiniBand (IPoIB) devices, for example, **ib0**, **ib1**, and so on. To avoid conflicts, Red Hat recommends creating a rule in the **udev** device manager to create persistent and meaningful names such as **mlx4_ib0**.

Prerequisites

- You've installed an InfiniBand device.

Procedure

1. Display the hardware address of the device **ib0**:

```
# ip link show ib0
8: ib0: >BROADCAST,MULTICAST,UP,LOWER_UP< mtu 65520 qdisc pfifo_fast state UP
mode DEFAULT qlen 256
    link/infiniband 80:00:02:00:fe:80:00:00:00:00:00:00:00:02:c9:03:00:31:78:f2 brd
    00:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:00:ff:ff:ff:ff
```

The last eight bytes of the address are required to create a **udev** rule in the next step.

2. To configure a rule that renames the device with the **00:02:c9:03:00:31:78:f2** hardware address to **mlx4_ib0**, edit the **/etc/udev/rules.d/70-persistent-ipoib.rules** file and add an **ACTION** rule:

```
ACTION=="add", SUBSYSTEM=="net", DRIVERS=="*", ATTR{type}=="32",
ATTR{address}=="?*00:02:c9:03:00:31:78:f2", NAME="mlx4_ib0"
```

3. Reboot the host:

```
# reboot
```

Additional resources

- **udev(7)** man page
- [Understanding IPoIB hardware addresses](#)

4.2. INCREASING THE AMOUNT OF MEMORY THAT USERS ARE ALLOWED TO PIN IN THE SYSTEM

Remote direct memory access (RDMA) operations require the pinning of physical memory. As a consequence, the kernel is not allowed to write memory into the swap space. If a user pins too much memory, the system can run out of memory, and the kernel terminates processes to free up more memory. Therefore, memory pinning is a privileged operation.

If non-root users need to run large RDMA applications, it is necessary to increase the amount of memory to maintain pages in primary memory pinned all the time.

Procedure

- As the **root** user, create the file **/etc/security/limits.conf** with the following contents:

```
@rdma soft memlock unlimited
@rdma hard memlock unlimited
```

Verification

- Log in as a member of the **rdma** group after editing the **/etc/security/limits.conf** file.
Note that Red Hat Enterprise Linux applies updated **ulimit** settings when the user logs in.
- Use the **ulimit -l** command to display the limit:

```
$ ulimit -l
unlimited
```

If the command returns **unlimited**, the user can pin an unlimited amount of memory.

Additional resources

- limits.conf(5)** man page

4.3. CONFIGURING THE RDMA SERVICE

With the Remote Direct Memory Access (RDMA) protocol, you can transfer data between the RDMA enabled systems over the network by using the main memory. The RDMA protocol provides low latency and high throughput. To manage supported network protocols and communication standards, you need to configure the **rdma** service. This configuration includes high speed network protocols such as RoCE and iWARP, and communication standards such as Soft-RoCE and Soft-iWARP. When Red Hat Enterprise Linux detects InfiniBand, iWARP, or RoCE devices and their configuration files residing at the **/etc/rdma/modules/*** directory, the **udev** device manager instructs **systemd** to start the **rdma** service. Configuration of modules in the **/etc/rdma/modules/rdma.conf** file remains persistent after reboot. You need to restart the **rdma-load-modules@rdma.service** configuration service to apply changes.

Procedure

- Edit the **/etc/rdma/modules/rdma.conf** file and uncomment the modules that you want to enable:

```
# These modules are loaded by the system if any RDMA devices is installed

# iSCSI over RDMA client support
ib_iser

# iSCSI over RDMA target support
ib_isert

# SCSI RDMA Protocol target driver
ib_srpt

# User access to RDMA verbs (supports libibverbs)
ib_uverbs
```



```
# User access to RDMA connection management (supports librdmacm)
rdma_ucm

# RDS over RDMA support
# rds_rdma

# NFS over RDMA client support
xprtrdma

# NFS over RDMA server support
svcrdma
```

- Restart the service to make the changes effective:

```
# systemctl restart <rdma-load-modules@rdma.service>
```

Verification

- After a reboot, check the service status:

```
# systemctl status <rdma-load-modules@rdma.service>
```

4.4. ENABLING NFS OVER RDMA (NFSORDMA)

In Red Hat Enterprise Linux 8, Remote direct memory access (RDMA) service on RDMA-capable hardware provides Network File System (NFS) protocol support for high-speed file transfer over the network.

Procedure

- Install the **rdma-core** package:

```
# yum install rdma-core
```

- Verify the lines with **xprtrdma** and **svcrdma** are not commented out in the **/etc/rdma/modules/rdma.conf** file:

```
# NFS over RDMA client support
xprtrdma
# NFS over RDMA server support
svcrdma
```

- On the NFS server, create directory **/mnt/nfsordma** and export it to **/etc/exports**:

```
# mkdir /mnt/nfsordma
# echo "/mnt/nfsordma *(fsid=0,rw,async,insecure,no_root_squash)" >> /etc/exports
```

- On the NFS client, mount the nfs-share with server IP address, for example, **172.31.0.186**:

```
# mount -o rdma,port=20049 172.31.0.186:/mnt/nfs-share /mnt/nfs
```

- Restart the **nfs-server** service:

```
# systemctl restart nfs-server
```

Additional resources

- [The RFC 5667 standard](#)

CHAPTER 5. CONFIGURING AN INFINIBAND SUBNET MANAGER

All InfiniBand networks must have a subnet manager running for the network to function. This is true even if two machines are connected directly with no switch involved.

It is possible to have more than one subnet manager. In that case, one acts as a master and another subnet manager acts as a slave that will take over in case the master subnet manager fails.

Most InfiniBand switches contain an embedded subnet manager. However, if you need a more up-to-date subnet manager or if you require more control, use the **OpenSM** subnet manager provided by Red Hat Enterprise Linux.

5.1. INSTALLING THE OPENSMTM SUBNET MANAGER

OpenSM is a subnet manager and administrator that follows the InfiniBand specifications to initialize InfiniBand hardware where at least one instance of **OpenSM** service always runs.

Procedure

1. Install the **opensmtm** package:

```
# yum install opensmtm
```

2. Configure OpenSM in case the default installation does not match your environment. With only one InfiniBand port, the host acts as the master subnet manager that does not require any custom changes. The default configuration works without any modification.
3. Enable and start the **opensmtm** service:

```
# systemctl enable --now opensmtm
```

Additional resources

- **opensmtm(8)** man page

5.2. CONFIGURING OPENSMTM USING THE SIMPLE METHOD

OpenSM is an InfiniBand specification-based subnet manager and administrator that configures the InfiniBand fabric, a network topology to interconnect the InfiniBand nodes.

Prerequisites

- One or more InfiniBand ports are installed on the server

Procedure

1. Obtain the GUIDs for the ports using the **ibstat** utility:

```
# ibstat -d mlx4_0
```

```
CA 'mlx4_0'
```

```

CA type: MT4099
Number of ports: 2
Firmware version: 2.42.5000
Hardware version: 1
Node GUID: 0xf4521403007be130
System image GUID: 0xf4521403007be133
Port 1:
  State: Active
  Physical state: LinkUp
  Rate: 56
  Base lid: 3
  LMC: 0
  SM lid: 1
  Capability mask: 0x02594868
  Port GUID: 0xf4521403007be131
  Link layer: InfiniBand
Port 2:
  State: Down
  Physical state: Disabled
  Rate: 10
  Base lid: 0
  LMC: 0
  SM lid: 0
  Capability mask: 0x04010000
  Port GUID: 0xf65214ffe7be132
  Link layer: Ethernet

```

**NOTE**

Some InfiniBand adapters use the same GUID for the node, system, and port.

2. Edit the `/etc/sysconfig/opensm` file and set the GUIDs in the **GUIDS** parameter:

```
GUIDS="GUID_1 GUID_2"
```

3. You can set the **PRIORITY** parameter if multiple subnet managers are available in your subnet. For example:

```
PRIORITY=15
```

Additional resources

- `/etc/sysconfig/opensm`

5.3. CONFIGURING OPENSMTM BY EDITING THE OPENSMTM.CONF FILE

OpenSM performance depends on the number of available InfiniBand ports on the device. You can customize it by editing `/etc/rdma/opensm.conf` file.

Prerequisites

- Only one InfiniBand port is installed on the server.

Procedure

1. Edit the **/etc/rdma/opensm.conf** file and customize the settings to match your environment. After updating an **opensm** package, the **yum** utility overrides the **/etc/rdma/opensm.conf** and creates a copy which is the new OpenSM configuration file **/etc/rdma/opensm.conf.rpmnew**. So, you can compare the previous and new files to identify changes and incorporate them manually in file **opensm.conf**.
2. Restart the **opensm** service:

```
# systemctl restart opensm
```

5.4. CONFIGURING MULTIPLE OPENSMT INSTANCES

OpenSM is an InfiniBand constrained subnet manager and administrator. To provide high performance, **OpenSM** uses switched fabric network topology that interconnects with InfiniBand network nodes.

Prerequisites

- One or more InfiniBand ports are installed on the server.

Procedure

1. Copy the **/etc/rdma/opensm.conf** file to **/etc/rdma/opensm.conf.orig** file:

```
# cp /etc/rdma/opensm.conf /etc/rdma/opensm.conf.orig
```

When you install an updated **opensm** package, the **yum** utility overrides the **/etc/rdma/opensm.conf**. With the copy created in this step, compare the previous and new files to identify changes and incorporate them manually in the instance-specific **opensm.conf** files.

2. Create a copy of the **/etc/rdma/opensm.conf** file:

```
# cp /etc/rdma/opensm.conf /etc/rdma/opensm.conf.1
```

For each instance, create and append a unique and continuous number to a copy of the configuration file.

After updating the **opensm** package, the **yum** utility stores the new OpenSM configuration file as **/etc/rdma/opensm.conf.rpmnew**. Compare this file with your customized **/etc/rdma/opensm.conf.*** files, and manually incorporate the changes.

3. Edit the copy you created in the previous step, and customize the settings for the instance to match your environment. For example, set the **guid**, **subnet_prefix**, and **logdir** parameters.
4. Optionally, create a **partitions.conf** file with a unique name specifically for this subnet and reference that file in the **partition_config_file** parameter in the corresponding copy of the **opensm.conf** file.
5. Repeat the previous steps for each instance you want to create.
6. Start the **opensm** service:

```
# systemctl start opensm
```

The **opensm** service automatically starts a unique instance for each **opensm.conf.*** file in the **/etc/rdma/** directory. If multiple **opensm.conf.*** files exist, the service ignores settings in the **/etc/sysconfig/opensm** file as well as in the base **/etc/rdma/opensm.conf** file.

5.5. CREATING A PARTITION CONFIGURATION

Partitions enable administrators to create subnets on InfiniBand similar to Ethernet VLANs.



IMPORTANT

If you define a partition with a specific speed such as 40 Gbps, all hosts within this partition must support this speed minimum. If a host does not meet the speed requirements, it can't join the partition. Therefore, set the speed of a partition to the lowest speed supported by any host with permission to join the partition.

Prerequisites

- One or more InfiniBand ports are installed on the server

Procedure

1. Edit the **/etc/rdma/partitions.conf** file to configure the partitions as follows:



NOTE

All fabrics must contain the **0x7fff** partition, and all switches and all hosts must belong to that fabric.

Add the following content to the file to create the **0x7fff** default partition at a reduced speed of **10** Gbps, and a partition **0x0002** with a speed of **40** Gbps:

```
# For reference:
# IPv4 IANA reserved multicast addresses:
# http://www.iana.org/assignments/multicast-addresses/multicast-addresses.txt
# IPv6 IANA reserved multicast addresses:
# http://www.iana.org/assignments/ipv6-multicast-addresses/ipv6-multicast-addresses.xml
#
# mtu =
# 1 = 256
# 2 = 512
# 3 = 1024
# 4 = 2048
# 5 = 4096
#
# rate =
# 2 = 2.5 GBit/s
# 3 = 10 GBit/s
# 4 = 30 GBit/s
# 5 = 5 GBit/s
# 6 = 20 GBit/s
# 7 = 40 GBit/s
# 8 = 60 GBit/s
# 9 = 80 GBit/s
# 10 = 120 GBit/s
```

Default=0x7fff, rate=3, mtu=4, scope=2, defmember=full:

ALL, ALL_SWITCHES=full;

Default=0x7fff, ipoib, rate=3, mtu=4, scope=2:

```
mgid=ff12:401b::ffff:ffff # IPv4 Broadcast address
mgid=ff12:401b::1         # IPv4 All Hosts group
mgid=ff12:401b::2         # IPv4 All Routers group
mgid=ff12:401b::16        # IPv4 IGMP group
mgid=ff12:401b::fb        # IPv4 mDNS group
mgid=ff12:401b::fc        # IPv4 Multicast Link Local Name Resolution group
mgid=ff12:401b::101       # IPv4 NTP group
mgid=ff12:401b::202       # IPv4 Sun RPC
mgid=ff12:601b::1         # IPv6 All Hosts group
mgid=ff12:601b::2         # IPv6 All Routers group
mgid=ff12:601b::16        # IPv6 MLDv2-capable Routers group
mgid=ff12:601b::fb        # IPv6 mDNS group
mgid=ff12:601b::101       # IPv6 NTP group
mgid=ff12:601b::202       # IPv6 Sun RPC group
mgid=ff12:601b::1:3       # IPv6 Multicast Link Local Name Resolution group
ALL=full, ALL_SWITCHES=full;
```

ib0_2=0x0002, rate=7, mtu=4, scope=2, defmember=full:

ALL, ALL_SWITCHES=full;

ib0_2=0x0002, ipoib, rate=7, mtu=4, scope=2:

```
mgid=ff12:401b::ffff:ffff # IPv4 Broadcast address
mgid=ff12:401b::1         # IPv4 All Hosts group
mgid=ff12:401b::2         # IPv4 All Routers group
mgid=ff12:401b::16        # IPv4 IGMP group
mgid=ff12:401b::fb        # IPv4 mDNS group
mgid=ff12:401b::fc        # IPv4 Multicast Link Local Name Resolution group
mgid=ff12:401b::101       # IPv4 NTP group
mgid=ff12:401b::202       # IPv4 Sun RPC
mgid=ff12:601b::1         # IPv6 All Hosts group
mgid=ff12:601b::2         # IPv6 All Routers group
mgid=ff12:601b::16        # IPv6 MLDv2-capable Routers group
mgid=ff12:601b::fb        # IPv6 mDNS group
mgid=ff12:601b::101       # IPv6 NTP group
mgid=ff12:601b::202       # IPv6 Sun RPC group
mgid=ff12:601b::1:3       # IPv6 Multicast Link Local Name Resolution group
ALL=full, ALL_SWITCHES=full;
```

CHAPTER 6. CONFIGURING IPOIB

By default, InfiniBand does not use the internet protocol (IP) for communication. However, IP over InfiniBand (IPoB) provides an IP network emulation layer on top of InfiniBand remote direct memory access (RDMA) networks. This allows existing unmodified applications to transmit data over InfiniBand networks, but the performance is lower than if the application would use RDMA natively.



NOTE

The Mellanox devices, starting from ConnectX-4 and above, on RHEL 8 and later use Enhanced IPoB mode by default (datagram only). Connected mode is not supported on these devices.

6.1. THE IPOIB COMMUNICATION MODES

An IPoB device is configurable in either **Datagram** or **Connected** mode. The difference is the type of queue pair the IPoB layer attempts to open with the machine at the other end of the communication:

- In the **Datagram** mode, the system opens an unreliable, disconnected queue pair. This mode does not support packages larger than Maximum Transmission Unit (MTU) of the InfiniBand link layer. During transmission of data, the IPoB layer adds a 4-byte IPoB header on top of the IP packet. As a result, the IPoB MTU is 4 bytes less than the InfiniBand link-layer MTU. As **2048** is a common InfiniBand link-layer MTU, the common IPoB device MTU in **Datagram** mode is **2044**.
- In the **Connected** mode, the system opens a reliable, connected queue pair. This mode allows messages larger than the InfiniBand link-layer MTU. The host adapter handles packet segmentation and reassembly. As a result, in the **Connected** mode, the messages sent from InfiniBand adapters have no size limits. However, there are limited IP packets due to the **data** field and TCP/IP **header** field. For this reason, the IPoB MTU in the **Connected** mode is **65520** bytes.

The **Connected** mode has a higher performance but consumes more kernel memory.

Though a system is configured to use the **Connected** mode, a system still sends multicast traffic using the **Datagram** mode because InfiniBand switches and fabric cannot pass multicast traffic in the **Connected** mode. Also, when the host is not configured to use the **Connected** mode, the system falls back to the **Datagram** mode.

While running an application that sends multicast data up to MTU on the interface, configures the interface in **Datagram** mode or configure the application to cap the send size of a packet that will fit in datagram-sized packets.

6.2. UNDERSTANDING IPOIB HARDWARE ADDRESSES

IPoB devices have a **20** byte hardware address that consists of the following parts:

- The first 4 bytes are flags and queue pair numbers
- The next 8 bytes are the subnet prefix
The default subnet prefix is **0xfe:80:00:00:00:00:00:00**. After the device connects to the subnet manager, the device changes this prefix to match with the configured subnet manager.
- The last 8 bytes are the Globally Unique Identifier (GUID) of the InfiniBand port that attaches to the IPoB device

**NOTE**

As the first 12 bytes can change, don't use them in the **udev** device manager rules.

6.3. CONFIGURING AN IPOIB CONNECTION USING NMCLI COMMANDS

The **nmcli** command-line utility controls the NetworkManager and reports network status using CLI.

Prerequisites

- An InfiniBand device is installed on the server
- The corresponding kernel module is loaded

Procedure

1. Create the InfiniBand connection to use the **mlx4_ib0** interface in the **Connected** transport mode and the maximum MTU of **65520** bytes:

```
# nmcli connection add type infiniband con-name mlx4_ib0 ifname mlx4_ib0 transport-mode Connected mtu 65520
```

2. You can also set **0x8002** as a **P_Key** interface of the **mlx4_ib0** connection:

```
# nmcli connection modify mlx4_ib0 infiniband.p-key 0x8002
```

3. To configure the IPv4 settings set a static IPv4 address, network mask, default gateway, and DNS server of the **mlx4_ib0** connection:

```
# nmcli connection modify mlx4_ib0 ipv4.addresses 192.0.2.1/24
# nmcli connection modify mlx4_ib0 ipv4.gateway 192.0.2.254
# nmcli connection modify mlx4_ib0 ipv4.dns 192.0.2.253
# nmcli connection modify mlx4_ib0 ipv4.method manual
```

4. To configure the IPv6 settings set a static IPv6 address, network mask, default gateway, and DNS server of the **mlx4_ib0** connection:

```
# nmcli connection modify mlx4_ib0 ipv6.addresses 2001:db8:1::1/32
# nmcli connection modify mlx4_ib0 ipv6.gateway 2001:db8:1::fffe
# nmcli connection modify mlx4_ib0 ipv6.dns 2001:db8:1::fffd
# nmcli connection modify mlx4_ib0 ipv6.method manual
```

5. To activate the **mlx4_ib0** connection:

```
# nmcli connection up mlx4_ib0
```

6.4. CONFIGURING AN IPOIB CONNECTION BY USING THE NETWORK RHEL SYSTEM ROLE

You can use the **network** RHEL System Role to remotely create NetworkManager connection profiles for IP over InfiniBand (IPoIB) devices. For example, remotely add an InfiniBand connection for the **mlx4_ib0** interface with the following settings by running an Ansible playbook:

- An IPoIB device - **mlx4_ib0.8002**
- A partition key **p_key** - **0x8002**
- A static **IPv4** address - **192.0.2.1** with a **/24** subnet mask
- A static **IPv6** address - **2001:db8:1::1** with a **/64** subnet mask

Perform this procedure on the Ansible control node.

Prerequisites

- [You have prepared the control node and the managed nodes](#)
- You logged in to the control node as a user who can run playbooks on the managed nodes.
- The account you use to connect to the managed nodes has **sudo** permissions on them.
- The managed nodes or groups of managed nodes on which you want to run this playbook are listed in the Ansible inventory file.
- An InfiniBand device named **mlx4_ib0** is installed in the managed nodes.
- The managed nodes use NetworkManager to configure the network.

Procedure

1. Create a playbook file, for example **~/IPoIB.yml**, with the following content:

```
---
- name: Configure the network
  hosts: managed-node-01.example.com
  tasks:
    - name: Configure IPoIB
      include_role:
        name: rhel-system-roles.network

  vars:
    network_connections:

    # InfiniBand connection mlx4_ib0
    - name: mlx4_ib0
      interface_name: mlx4_ib0
      type: infiniband

    # IPoIB device mlx4_ib0.8002 on top of mlx4_ib0
    - name: mlx4_ib0.8002
      type: infiniband
      autoconnect: yes
      infiniband:
        p_key: 0x8002
        transport_mode: datagram
```

```
parent: mlx4_ib0
ip:
  address:
    - 192.0.2.1/24
    - 2001:db8:1::1/64
state: up
```

If you set a **p_key** parameter as in this example, do not set an **interface_name** parameter on the IPoIB device.

2. Run the playbook:

```
# ansible-playbook ~/IPoIB.yml
```

Verification

1. On the **managed-node-01.example.com** host, display the IP settings of the **mlx4_ib0.8002** device:

```
# ip address show mlx4_ib0.8002
...
inet 192.0.2.1/24 brd 192.0.2.255 scope global noprefixroute ib0.8002
    valid_lft forever preferred_lft forever
inet6 2001:db8:1::1/64 scope link tentative noprefixroute
    valid_lft forever preferred_lft forever
```

2. Display the partition key (P_Key) of the **mlx4_ib0.8002** device:

```
# cat /sys/class/net/mlx4_ib0.8002/pkey
0x8002
```

3. Display the mode of the **mlx4_ib0.8002** device:

```
# cat /sys/class/net/mlx4_ib0.8002/mode
datagram
```

Additional resources

- [/usr/share/ansible/roles/rhel-system-roles.network/README.md](#) file

6.5. CONFIGURING AN IPOIB CONNECTION USING NM-CONNECTION-EDITOR

The **nmcli-connection-editor** application configures and manages network connections stored by NetworkManager using the management console.

Prerequisites

- An InfiniBand device is installed on the server.
- Corresponding kernel module is loaded
- The **nm-connection-editor** package is installed.

Procedure

1. Enter the command:

```
$ nm-connection-editor
```

2. Click the **+** button to add a new connection.
3. Select the **InfiniBand** connection type and click **Create**.
4. On the **InfiniBand** tab:
 - a. Change the connection name if you want to.
 - b. Select the transport mode.
 - c. Select the device.
 - d. Set an MTU if needed.
5. On the **IPv4 Settings** tab, configure the IPv4 settings. For example, set a static IPv4 address, network mask, default gateway, and DNS server:

The screenshot shows the 'nm-connection-editor' window titled 'Editing mlx4_ib0'. The 'Connection name' field is set to 'mlx4_ib0'. The 'IPv4 Settings' tab is selected, showing the 'Method' set to 'Manual'. Under 'Addresses', a table lists one address: 192.0.2.1 with a netmask of 24 and a gateway of 192.0.2.254. The 'DNS servers' field is set to 192.0.2.253. 'Add' and 'Delete' buttons are next to the address table.

Address	Netmask	Gateway
192.0.2.1	24	192.0.2.254

6. On the **IPv6 Settings** tab, configure the IPv6 settings. For example, set a static IPv6 address, network mask, default gateway, and DNS server:

Editing **mlx4_ib0**

Connection name: **mlx4_ib0**

General InfiniBand Proxy IPv4 Settings **IPv6 Settings**

Method: **Manual**

Addresses

Address	Prefix	Gateway
2001:db8::1	32	2001:db8::fffe

Add
Delete

DNS servers: **2001:db8::fffd**

7. Click **Save** to save the team connection.
8. Close **nm-connection-editor**.
9. You can set a **P_Key** interface. As this setting is not available in **nm-connection-editor**, you must set this parameter on the command line.
For example, to set **0x8002** as **P_Key** interface of the **mlx4_ib0** connection:

```
# nmcli connection modify mlx4_ib0 infiniband.p-key 0x8002
```

CHAPTER 7. TESTING INFINIBAND NETWORKS

7.1. TESTING EARLY INFINIBAND RDMA OPERATIONS

InfiniBand provides low latency and high performance for Remote Direct Memory Access (RDMA).



NOTE

Apart from InfiniBand, if you use IP-based devices such as Internet Wide-area Remote Protocol (iWARP) or RDMA over Converged Ethernet (RoCE) or InfiniBand over Ethernet (IBoE) devices, see:

- [Testing an IPoB using the ping utility](#)
- [Testing an RDMA network using qperf after IPoB is configured](#)

Prerequisites

- You've configured the **rdma** service.
- You've installed the **libibverbs-utils** and **infiniband-diags** packages.

Procedure

1. List the available InfiniBand devices:

```
# ibv_devices
```

device	node GUID
-----	-----
mlx4_0	0002c903003178f0
mlx4_1	f4521403007bcba0

2. Display the information of the **mlx4_1** device:

```
# ibv_devinfo -d mlx4_1
```

```
hca_id: mlx4_1
transport:          InfiniBand (0)
fw_ver:             2.30.8000
node_guid:          f452:1403:007b:cba0
sys_image_guid:     f452:1403:007b:cba3
vendor_id:          0x02c9
vendor_part_id:     4099
hw_ver:             0x0
board_id:           MT_1090120019
phys_port_cnt:      2
  port: 1
    state:           PORT_ACTIVE (4)
    max_mtu:         4096 (5)
    active_mtu:      2048 (4)
    sm_lid:           2
    port_lid:         2
    port_lmc:         0x01
```

```

link_layer:      InfiniBand

port: 2
  state:         PORT_ACTIVE (4)
  max_mtu:       4096 (5)
  active_mtu:    4096 (5)
  sm_lid:        0
  port_lid:      0
  port_lmc:      0x00
  link_layer:    Ethernet

```

3. Display the status of the **mlx4_1** device:

```

# ibstat mlx4_1

CA 'mlx4_1'
CA type: MT4099
Number of ports: 2
Firmware version: 2.30.8000
Hardware version: 0
Node GUID: 0xf4521403007bcba0
System image GUID: 0xf4521403007bcba3
Port 1:
  State: Active
  Physical state: LinkUp
  Rate: 56
  Base lid: 2
  LMC: 1
  SM lid: 2
  Capability mask: 0x0251486a
  Port GUID: 0xf4521403007bcba1
  Link layer: InfiniBand
Port 2:
  State: Active
  Physical state: LinkUp
  Rate: 40
  Base lid: 0
  LMC: 0
  SM lid: 0
  Capability mask: 0x04010000
  Port GUID: 0xf65214fffe7bcba2
  Link layer: Ethernet

```

4. The **ibping** utility pings an InfiniBand address and runs as a client/server.

- a. To start server mode on a host, use the **-S** parameter on port number **-P** with **-C** InfiniBand certificate authority (CA) name:

```
# ibping -S -C mlx4_1 -P 1
```

- b. To start client mode on another host, send some packets **-c** on port number **-P** using **-C** InfiniBand certificate authority (CA) name with **-L** Local Identifier (LID):

```
# ibping -c 50 -C mlx4_0 -P 1 -L 2
```

Additional resources

- **ibping(8)** man page

7.2. TESTING AN IPOIB USING THE PING UTILITY

After you configured IP over InfiniBand (IPoIB), use the **ping** utility to send ICMP packets to test the IPoIB connection.

Prerequisites

- The two RDMA hosts are connected in the same InfiniBand fabric with RDMA ports
- The IPoIB interfaces in both hosts are configured with IP addresses within the same subnet

Procedure

- Use the **ping** utility to send five ICMP packets to the remote host's InfiniBand adapter:

```
# ping -c5 192.0.2.1
```

7.3. TESTING AN RDMA NETWORK USING QPERF AFTER IPOIB IS CONFIGURED

The **qperf** utility measures RDMA and IP performance between two nodes in terms of bandwidth, latency, and CPU utilization.

Prerequisites

- You've installed the **qperf** package on both hosts.
- IPoIB is configured on both hosts.

Procedure

1. Start **qperf** on one of the hosts without any options to act as a server:

```
# qperf
```

2. Use the following commands on the client. The commands use port **1** of the **mlx4_0** host channel adapter in the client to connect to IP address **192.0.2.1** assigned to the InfiniBand adapter in the server.

- a. Display the configuration of the host channel adapter:

```
# qperf -v -i mlx4_0:1 192.0.2.1 conf
```

```
conf:
```

```
loc_node  = rdma-dev-01.lab.bos.redhat.com
loc_cpu   = 12 Cores: Mixed CPUs
loc_os    = Linux 4.18.0-187.el8.x86_64
loc_qperf = 0.4.11
rem_node  = rdma-dev-00.lab.bos.redhat.com
```



```
rem_cpu   = 12 Cores: Mixed CPUs
rem_os    = Linux 4.18.0-187.el8.x86_64
rem_qperf = 0.4.11
```

- b. Display the Reliable Connection (RC) streaming two-way bandwidth:

```
# qperf -v -i mlx4_0:1 192.0.2.1 rc_bi_bw
```

```
rc_bi_bw:
  bw          = 10.7 GB/sec
  msg_rate    = 163 K/sec
  loc_id      = mlx4_0
  rem_id      = mlx4_0:1
  loc_cpus_used = 65 % cpus
  rem_cpus_used = 62 % cpus
```

- c. Display the RC streaming one-way bandwidth:

```
# qperf -v -i mlx4_0:1 192.0.2.1 rc_bw
```

```
rc_bw:
  bw          = 6.19 GB/sec
  msg_rate    = 94.4 K/sec
  loc_id      = mlx4_0
  rem_id      = mlx4_0:1
  send_cost   = 63.5 ms/GB
  recv_cost   = 63 ms/GB
  send_cpus_used = 39.5 % cpus
  recv_cpus_used = 39 % cpus
```

Additional resources

- **qperf(1)** man page