Red Hat Enterprise Linux 8

Managing storage devices

Deploying and configuring single-node storage in Red Hat Enterprise Linux 8

Last Updated: 2020-04-28

Red Hat Enterprise Linux 8 Managing storage devices

Deploying and configuring single-node storage in Red Hat Enterprise Linux 8

Legal Notice

Abstract

This documentation collection provides instructions on how to effectively manage storage devices in Red Hat Enterprise Linux 8.

Table of Contents

PROVIDING FEEDBACK ON RED HAT DOCUMENTATION

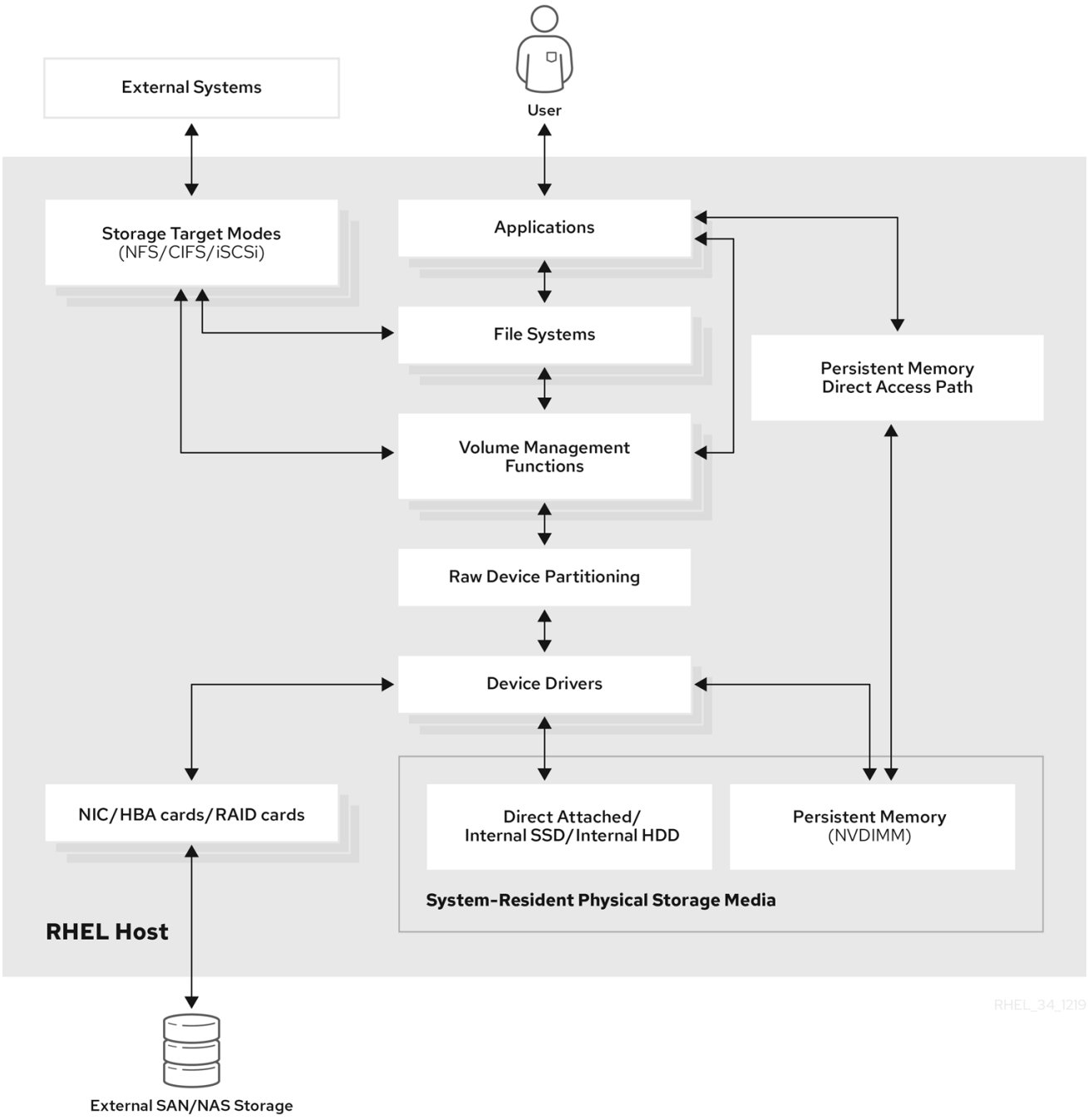
We appreciate your input on our documentation. Please let us know how we could make it better. To do so:

- For simple comments on specific passages:
  1. Make sure you are viewing the documentation in the *Multi-page HTML* format. In addition, ensure you see the **Feedback** button in the upper right corner of the document.
  2. Use your mouse cursor to highlight the part of text that you want to comment on.
  3. Click the **Add Feedback** pop-up that appears below the highlighted text.
  4. Follow the displayed instructions.
- For submitting more complex feedback, create a Bugzilla ticket:
  1. Go to the Bugzilla website.
  2. As the Component, use **Documentation**.
  3. Fill in the **Description** field with your suggestion for improvement. Include a link to the relevant part(s) of documentation.
  4. Click **Submit Bug**.

CHAPTER 1. OVERVIEW OF AVAILABLE STORAGE OPTIONS

This chapter describes the storage types that are available in Red Hat Enterprise Linux 8. Red Hat Enterprise Linux offers a variety of options for managing local storage, and for attaching to the remote storage. Figure 1.1, "High Level Red Hat Enterprise Linux Storage Diagram" describes the different storage options:

Figure 1.1. High Level Red Hat Enterprise Linux Storage Diagram



1.1. LOCAL STORAGE OPTIONS

Following are the local storage options available in Red Hat Enterprise Linux 8:

- Basic Disk Administration:
  Using parted and fdisk, you can create, modify, delete, and view the partitions. Following are the partitioning layout standards:

    - Master Boot Record (MBR): It is used with BIOS-based computers. You can create primary, extended, and logical partitions.

    - GUID Partition Table (GPT): It uses Globally Unique identifier (GUID) and provides unique disk and partition GUID.
      To encrypt the partition, you can use Linux Unified Key Setup-on-disk-format (LUKS). To encrypt the partition, select the option during the installation and the prompt displays to enter the passphrase. This passphrase unlocks the encryption key.

- Storage Consumption Options:

    - Non-Volatile Dual In-line Memory Modules (NVDIMM) Management: It is a combination of memory and storage. You can enable and manage various types of storage on NVDIMM devices connected to your system.

    - Block Storage Management: Data is stored in the form of blocks where each block has a unique identifier.

    - File Storage: Data is stored at file level on the local system. These data can be accessed locally using XFS (default) or ext4, and over a network by using NFS and SMB.

- Creating and Managing Logical Volumes:

    - Logical Volume Manager (LVM): It creates logical devices from physical devices. Logical volume (LV) is a combination of the physical volumes (PV) and volume groups (VG). Configuring LVM include:

        - Creating PV from the hard drives.

        - Creating VG from the PV.

        - Creating LV from the VG assigning mount points to the LV.

    - Virtual Data Optimizer (VDO): It is used for data reduction by using deduplication, compression, and thin provisioning. Using LV below VDO helps in:

        - Extending of VDO volume

        - Spanning VDO volume over multiple devices

- Local File System:

    - XFS

    - Ext4

    - Stratis: It is available as a Technology Preview. Stratis is a hybrid user-and-kernel local storage management system that supports advanced storage features.

1.2. REMOTE STORAGE OPTIONS

Following are the remote storage options available in Red Hat Enterprise Linux 8:

- Storage Connectivity Options:

    - iSCSI: RHEL 8 uses the targetcli tool to add, remove, view, and monitor iSCSI storage interconnects.

    - Fibre Channel (FC): Red Hat Enterprise Linux 8 provides the following native Fibre Channel drivers:
        - **lpfc**
        - **qla2xxx**
        - **Zfcp**
    - Non-volatile Memory Express (**NVMe**) is an interface which allows host software utility to communicate with solid state drives. Use the following types of fabric transport to configure NVMe over fabrics:
        - NVMe over fabrics using Remote Direct Memory Access (RDMA).
        - NVMe over fabrics using Fibre Channel (FC)
    - Device mapper multipathing (DM Multipath) allows you to configure multiple I/O paths between server nodes and storage arrays into a single device. These I/O paths are physical SAN connections that can include separate cables, switches, and controllers.

- Network File system:
    - NFS
    - SMB

1.3. GFS2 FILE SYSTEM CLUSTERED SOLUTION

The Red Hat Global File System 2 (GFS2) file system is a 64-bit symmetric cluster file system which provides a shared name space and manages coherency between multiple nodes sharing a common block device. A GFS2 file system is intended to provide a feature set which is as close as possible to a local file system, while at the same time enforcing full cluster coherency between nodes. To achieve this, the nodes employ a cluster-wide locking scheme for file system resources. This locking scheme uses communication protocols such as TCP/IP to exchange locking information.

In a few cases, the Linux file system API does not allow the clustered nature of GFS2 to be totally transparent; for example, programs using POSIX locks in GFS2 should avoid using the **GETLK** function since, in a clustered environment, the process ID may be for a different node in the cluster. In most cases however, the functionality of a GFS2 file system is identical to that of a local file system.

The Red Hat Enterprise Linux (RHEL) Resilient Storage Add-On provides GFS2, and it depends on the RHEL High Availability Add-On to provide the cluster management required by GFS2.

The **gfs2.ko** kernel module implements the GFS2 file system and is loaded on GFS2 cluster nodes.

To get the best performance from GFS2, it is important to take into account the performance considerations which stem from the underlying design. Just like a local file system, GFS2 relies on the page cache in order to improve performance by local caching of frequently used data. In order to maintain coherency across the nodes in the cluster, cache control is provided by the *glock* state machine.

For more information on GFS2 file systems, see the Configuring GFS2 file systems documentation.

1.4. GLUSTERED SOLUTIONS

This section provides an overview of the glustered options such as Red Hat Gluster Storage (RHGS) or Red Hat Ceph Storage (RHCS).

1.4.1. Red Hat Gluster Storage option

The Red Hat Gluster Storage (RHGS) is a software-defined storage platform. It aggregates disk storage resources from multiple servers into a single global namespace. GlusterFS is an open source distributed file system. It is suitable for cloud and hybrid solutions.

GlusterFS consists of different types of volume, which are the base for GlusterFS and provide different requirements. Volume is a collection of the bricks, which are the storage space themselves.

The following are the types of the GlusterFS volume:

- **Distributed GlusterFS volume** is the default volume. In this case each file is stored in one brick and can not be shared between different bricks.
- **Replicated GlusterFS volume** type maintains the replicas of data. In this case, if one brick fails, the user can still access the data.
- **Distributed replicated GlusterFS volume** is a hybrid volume which distributes replicas over a large number of systems. It is suitable for the environment where the requirements to scale storage and high-reliability are critical.

For more information on RHGS, see the Red Hat gluster storage administration guide.

1.4.2. Red Hat Ceph Storage option

Red Hat Ceph Storage (RHCS) is a scalable, open, software-defined storage platform that combines the most stable version of the Ceph storage system with a Ceph management platform, deployment utilities, and support services.

Red Hat Ceph Storage is designed for cloud infrastructure and web-scale object storage. Red Hat Ceph Storage clusters consist of the following types of nodes:

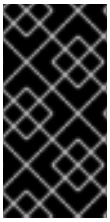
Red Hat Ceph Storage Ansible administration node

This type of node acts as the traditional Ceph Administration node did for previous versions of Red Hat Ceph Storage. This type of node provides the following functions:

- Centralized storage cluster management
- The Ceph configuration files and keys
- Optionally, local repositories for installing Ceph on nodes that cannot access the Internet for security reasons

Monitor nodes

Each monitor node runs the monitor daemon (**ceph-mon**), which maintains a master copy of the cluster map. The cluster map includes the cluster topology. A client connecting to the Ceph cluster retrieves the current copy of the cluster map from the monitor which enables the client to read from and write data to the cluster.



IMPORTANT

Ceph can run with one monitor; however, to ensure high availability in a production cluster, Red Hat will only support deployments with at least three monitor nodes. Red Hat recommends deploying a total of 5 Ceph Monitors for storage clusters exceeding 750 OSDs.

OSD nodes

Each Object Storage Device (OSD) node runs the Ceph OSD daemon (**ceph-osd**), which interacts with logical disks attached to the node. Ceph stores data on these OSD nodes.

Ceph can run with very few OSD nodes, which the default is three, but production clusters realize better performance beginning at modest scales, for example 50 OSDs in a storage cluster. Ideally, a Ceph cluster has multiple OSD nodes, allowing isolated failure domains by creating the CRUSH map.

MDS nodes

Each Metadata Server (MDS) node runs the MDS daemon (**ceph-mds**), which manages metadata related to files stored on the Ceph File System (CephFS). The MDS daemon also coordinates access to the shared cluster.

Object Gateway node

Ceph Object Gateway node runs the Ceph RADOS Gateway daemon (**ceph-radosgw**), and is an object storage interface built on top of **librados** to provide applications with a RESTful gateway to Ceph Storage Clusters. The Ceph Object Gateway supports two interfaces:

S3

Provides object storage functionality with an interface that is compatible with a large subset of the Amazon S3 RESTful API.

Swift

Provides object storage functionality with an interface that is compatible with a large subset of the OpenStack Swift API.

For more information on RHCS, see the Red Hat Ceph Storage documentation.

CHAPTER 2. GETTING STARTED WITH PARTITIONS

As a system administrator, you can use the following procedures to create, delete, and modify various types of disk partitions.

For an overview of the advantages and disadvantages to using partitions on block devices, see the following KBase article: https://access.redhat.com/solutions/163853.

2.1. VIEWING THE PARTITION TABLE

As a system administrator, you can display the partition table of a block device to see the partition layout and details about individual partitions.

2.1.1. Viewing the partition table with parted

This procedure describes how to view the partition table on a block device using the **parted** utility.

Procedure

1. Start the interactive **parted** shell:

```
# parted block-device
```

   - Replace *block-device* with the path to the device you want to examine: for example, **/dev/sda**.

2. View the partition table:

```
(parted) print
```

3. Optionally, use the following command to switch to another device you want to examine next:

```
(parted) select block-device
```

Additional resources

- The **parted(8)** man page.

2.1.2. Example output of parted print

This section provides an example output of the **print** command in the **parted** shell and describes fields in the output.

Example 2.1. Output of the print command

```
Model: ATA SAMSUNG MZNLN256 (scsi)
Disk /dev/sda: 256GB
Sector size (logical/physical): 512B/512B
Partition Table: msdos
Disk Flags:

Number  Start   End     Size    Type     File system  Flags
```

```
1     1049kB  269MB   268MB   primary   xfs         boot
2     269MB   34.6GB  34.4GB  primary
3     34.6GB  45.4GB  10.7GB  primary
4     45.4GB  256GB   211GB   extended
5     45.4GB  256GB   211GB   logical
```

Following is a description of the fields:

Model: ATA SAMSUNG MZNLN256 (scsi)

The disk type, manufacturer, model number, and interface.

Disk /dev/sda: 256GB

The file path to the block device and the storage capacity.

Partition Table: msdos

The disk label type.

Number

The partition number. For example, the partition with minor number 1 corresponds to **/dev/sda1**.

Start and **End**

The location on the device where the partition starts and ends.

Type

Valid types are metadata, free, primary, extended, or logical.

File system

The file system type. If the **File system** field of a device shows no value, this means that its file system type is unknown. The **parted** utility cannot recognize the file system on encrypted devices.

Flags

Lists the flags set for the partition. Available flags are **boot**, **root**, **swap**, **hidden**, **raid**, **lvm**, or **lba**.

2.2. CREATING A PARTITION TABLE ON A DISK

As a system administrator, you can format a block device with different types of partition tables to enable using partitions on the device.

WARNING

Formatting a block device with a partition table deletes all data stored on the device.

2.2.1. Considerations before modifying partitions on a disk

This section lists key points to consider before creating, removing, or resizing partitions.

NOTE

This section does not cover the DASD partition table, which is specific to the IBM Z architecture. For information on DASD, see:

- Configuring a Linux instance on IBM Z
- The What you should know about DASD article at the IBM Knowledge Center

The maximum number of partitions
The number of partitions on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, you can have either:
  - Up to four primary partitions, or
  - Up to three primary partitions, one extended partition, and multiple logical partitions within the extended.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum number of partitions is 128. While the GPT specification allows for more partitions by growing the area reserved for the partition table, common practice used by the **parted** utility is to limit it to enough area for 128 partitions.

The maximum size of a partition
The size of a partition on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, the maximum size is 2TiB.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum size is 8ZiB.

If you want to create a partition larger than 2TiB, the disk must be formatted with GPT.

Size alignment
The **parted** utility enables you to specify partition size using multiple different suffixes:

MiB, GiB, or TiB

Size expressed in powers of 2.

- The starting point of the partition is aligned to the exact sector specified by size.
- The ending point is aligned to the specified size minus 1 sector.

MB, GB, or TB

Size expressed in powers of 10.
The starting and ending point is aligned within one half of the specified unit: for example, ±500KB when using the MB suffix.

2.2.2. Comparison of partition table types

This section compares the properties of different types of partition tables that you can create on a block device.

Table 2.1. Partition table types

Partition table	Maximum number of partitions	Maximum partition size
Master Boot Record (MBR)	4 primary, or 3 primary and 12 logical inside an extended partition	2TiB
GUID Partition Table (GPT)	128	8ZiB

2.2.3. Creating a partition table on a disk with parted

This procedure describes how to format a block device with a partition table using the **parted** utility.

Procedure

1. Start the interactive **parted** shell:

```
# parted block-device
```

   - Replace *block-device* with the path to the device where you want to create a partition table: for example, /**dev**/**sda**.

2. Determine if there already is a partition table on the device:

```
(parted) print
```

   If the device already contains partitions, they will be deleted in the next steps.

3. Create the new partition table:

```
(parted) mklabel table-type
```

   - Replace *table-type* with with the intended partition table type:
     - **msdos** for MBR
     - **gpt** for GPT

   Example 2.2. Creating a GPT table

   For example, to create a GPT table on the disk, use:

```
(parted) mklabel gpt
```

   The changes start taking place as soon as you enter this command, so review it before executing it.

4. View the partition table to confirm that the partition table exists:

```
(parted) print
```

5. Exit the **parted** shell:

```
(parted) quit
```

Additional resources

- The **parted(8)** man page.

Next steps

- Create partitions on the device. See Section 2.3, "Creating a partition" for details.

2.3. CREATING A PARTITION

As a system administrator, you can create new partitions on a disk.

2.3.1. Considerations before modifying partitions on a disk

This section lists key points to consider before creating, removing, or resizing partitions.

NOTE

This section does not cover the DASD partition table, which is specific to the IBM Z architecture. For information on DASD, see:

- Configuring a Linux instance on IBM Z
- The What you should know about DASD article at the IBM Knowledge Center

The maximum number of partitions
The number of partitions on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, you can have either:
  - Up to four primary partitions, or
  - Up to three primary partitions, one extended partition, and multiple logical partitions within the extended.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum number of partitions is 128. While the GPT specification allows for more partitions by growing the area reserved for the partition table, common practice used by the **parted** utility is to limit it to enough area for 128 partitions.

The maximum size of a partition
The size of a partition on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, the maximum size is 2TiB.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum size is 8ZiB.

If you want to create a partition larger than 2TiB, the disk must be formatted with GPT.

Size alignment
The **parted** utility enables you to specify partition size using multiple different suffixes:

MiB, GiB, or TiB

Size expressed in powers of 2.

- The starting point of the partition is aligned to the exact sector specified by size.
- The ending point is aligned to the specified size minus 1 sector.

MB, GB, or TB

Size expressed in powers of 10.
The starting and ending point is aligned within one half of the specified unit: for example, ±500KB when using the MB suffix.

2.3.2. Partition types

This section describes different attributes that specify the type of a partition.

Partition types or flags
The partition type, or flag, is used by a running system only rarely. However, the partition type matters to on-the-fly generators, such as **systemd-gpt-auto-generator**, which use the partition type to, for example, automatically identify and mount devices.

- The **parted** utility provides some control of partition types by mapping the partition type to *flags*. The parted utility can handle only certain partition types: for example LVM, swap, or RAID.
- The **fdisk** utility supports the full range of partition types by specifying hexadecimal codes.

Partition file system type
The **parted** utility optionally accepts a file system type argument when creating a partition. The value is used to:

- Set the partition flags on MBR, or
- Set the partition UUID type on GPT. For example, the **swap**, **fat**, or **hfs** file system types set different GUIDs. The default value is the Linux Data GUID.

The argument does not modify the file system on the partition in any way. It only differentiates between the supported flags or GUIDs.

The following file system types are supported:

- **xfs**
- **ext2**
- **ext3**
- **ext4**
- **fat16**
- **fat32**
- **hfs**
- **hfs+**

- **linux-swap**
- **ntfs**
- **reiserfs**

2.3.3. Creating a partition with parted

This procedure describes how to create a new partition on a block device using the **parted** utility.

Prerequisites

- There is a partition table on the disk. For details on how to format the disk, see Section 2.2, "Creating a partition table on a disk".
- If the partition you want to create is larger than 2TiB, the disk must be formatted with the GUID Partition Table (GPT).

Procedure

1. Start the interactive **parted** shell:

```
# parted block-device
```

   - Replace *block-device* with the path to the device where you want to create a partition: for example, /**dev**/**sda**.

2. View the current partition table to determine if there is enough free space:

```
(parted) print
```

   - If there is not enough free space, you can resize an existing partition. For more information, see Section 2.5, "Resizing a partition".
   - From the partition table, determine:
     - The start and end points of the new partition
     - On MBR, what partition type it should be.

3. Create the new partition:

```
(parted) mkpart part-type name fs-type start end
```

   - Replace *part-type* with with **primary**, **logical**, or **extended** based on what you decided from the partition table. This applies only to the MBR partition table.
   - Replace *name* with an arbitrary partition name. This is required for GPT partition tables.
   - Replace *fs-type* with any one of **xfs**, **ext2**, **ext3**, **ext4**, **fat16**, **fat32**, **hfs**, **hfs+**, **linux-swap**, **ntfs**, or **reiserfs**. The *fs-type* parameter is optional. Note that **parted** does not create the file system on the partition.

- Replace *start* and *end* with the sizes that determine the starting and ending points of the partition, counting from the beginning of the disk. You can use size suffixes, such as **512MiB**, **20GiB**, or **1.5TiB**. The default size megabytes.

Example 2.3. Creating a small primary partition

For example, to create a primary partition from 1024MiB until 2048MiB on an MBR table, use:

```
(parted) mkpart primary 1024MiB 2048MiB
```

The changes start taking place as soon as you enter this command, so review it before executing it.

4. View the partition table to confirm that the created partition is in the partition table with the correct partition type, file system type, and size:

```
(parted) print
```

5. Exit the **parted** shell:

```
(parted) quit
```

6. Use the following command to wait for the system to register the new device node:

```
# udevadm settle
```

7. Verify that the kernel recognizes the new partition:

```
# cat /proc/partitions
```

Additional resources

- The **parted(8)** man page.

2.3.4. Setting a partition type with fdisk

This procedure describes how to set a partition type, or flag, using the **fdisk** utility.

Prerequisites

- There is a partition on the disk.

Procedure

1. Start the interactive **fdisk** shell:

```
# fdisk block-device
```

- Replace *block-device* with the path to the device where you want to set a partition type: for example, /**dev**/**sda**.

2. View the current partition table to determine the minor partition number:

```
Command (m for help): print
```

You can see the current partition type in the **Type** column and its corresponding type ID in the **Id** column.

3. Enter the partition type command and select a partition using its minor number:

```
Command (m for help): type
Partition number (1,2,3 default 3): 2
```

4. Optionally, list the available hexadecimal codes:

```
Hex code (type L to list all codes): L
```

5. Set the partition type:

```
Hex code (type L to list all codes): 8e
```

6. Write your changes and exit the **fdisk** shell:

```
Command (m for help): write
The partition table has been altered.
Syncing disks.
```

7. Verify your changes:

```
# fdisk --list block-device
```

2.4. REMOVING A PARTITION

As a system administrator, you can remove a disk partition that is no longer used to free up disk space.

WARNING

Removing a partition deletes all data stored on the partition.

2.4.1. Considerations before modifying partitions on a disk

This section lists key points to consider before creating, removing, or resizing partitions.

NOTE

This section does not cover the DASD partition table, which is specific to the IBM Z architecture. For information on DASD, see:

- Configuring a Linux instance on IBM Z
- The What you should know about DASD article at the IBM Knowledge Center

The maximum number of partitions
The number of partitions on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, you can have either:
  - Up to four primary partitions, or
  - Up to three primary partitions, one extended partition, and multiple logical partitions within the extended.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum number of partitions is 128. While the GPT specification allows for more partitions by growing the area reserved for the partition table, common practice used by the **parted** utility is to limit it to enough area for 128 partitions.

The maximum size of a partition
The size of a partition on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, the maximum size is 2TiB.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum size is 8ZiB.

If you want to create a partition larger than 2TiB, the disk must be formatted with GPT.

Size alignment
The **parted** utility enables you to specify partition size using multiple different suffixes:

MiB, GiB, or TiB

Size expressed in powers of 2.

- The starting point of the partition is aligned to the exact sector specified by size.
- The ending point is aligned to the specified size minus 1 sector.

MB, GB, or TB

Size expressed in powers of 10.
The starting and ending point is aligned within one half of the specified unit: for example, ±500KB when using the MB suffix.

2.4.2. Removing a partition with parted

This procedure describes how to remove a disk partition using the **parted** utility.

Procedure

1. Start the interactive **parted** shell:

```
# parted block-device
```

   - Replace *block-device* with the path to the device where you want to remove a partition: for example, /**dev**/**sda**.

2. View the current partition table to determine the minor number of the partition to remove:

```
(parted) print
```

3. Remove the partition:

```
(parted) rm minor-number
```

   - Replace *minor-number* with the minor number of the partition you want to remove: for example, **3**.

   The changes start taking place as soon as you enter this command, so review it before executing it.

4. Confirm that the partition is removed from the partition table:

```
(parted) print
```

5. Exit the **parted** shell:

```
(parted) quit
```

6. Verify that the kernel knows the partition is removed:

```
# cat /proc/partitions
```

7. Remove the partition from the /**etc**/**fstab** file if it is present. Find the line that declares the removed partition, and remove it from the file.

8. Regenerate mount units so that your system registers the new /**etc**/**fstab** configuration:

```
# systemctl daemon-reload
```

9. If you have deleted a swap partition or removed pieces of LVM, remove all references to the partition from the kernel command line in the /**etc**/**default**/**grub** file and regenerate GRUB configuration:

   - On a BIOS-based system:

```
# grub2-mkconfig --output=/etc/grub2.cfg
```

   - On a UEFI-based system:

```
# grub2-mkconfig --output=/etc/grub2-efi.cfg
```

10. To register the changes in the early boot system, rebuild the **initramfs** file system:

```
# dracut --force --verbose
```

Additional resources

- The **parted(8)** man page

2.5. RESIZING A PARTITION

As a system administrator, you can extend a partition to utilize unused disk space, or shrink a partition to use its capacity for different purposes.

2.5.1. Considerations before modifying partitions on a disk

This section lists key points to consider before creating, removing, or resizing partitions.

NOTE

This section does not cover the DASD partition table, which is specific to the IBM Z architecture. For information on DASD, see:

- Configuring a Linux instance on IBM Z
- The What you should know about DASD article at the IBM Knowledge Center

The maximum number of partitions
The number of partitions on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, you can have either:
  - Up to four primary partitions, or
  - Up to three primary partitions, one extended partition, and multiple logical partitions within the extended.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum number of partitions is 128. While the GPT specification allows for more partitions by growing the area reserved for the partition table, common practice used by the **parted** utility is to limit it to enough area for 128 partitions.

The maximum size of a partition
The size of a partition on a device is limited by the type of the partition table:

- On a device formatted with the **Master Boot Record (MBR)** partition table, the maximum size is 2TiB.
- On a device formatted with the **GUID Partition Table (GPT)**, the maximum size is 8ZiB.

If you want to create a partition larger than 2TiB, the disk must be formatted with GPT.

Size alignment
The **parted** utility enables you to specify partition size using multiple different suffixes:

MiB, GiB, or TiB

Size expressed in powers of 2.

- The starting point of the partition is aligned to the exact sector specified by size.
- The ending point is aligned to the specified size minus 1 sector.

MB, GB, or TB

Size expressed in powers of 10.
The starting and ending point is aligned within one half of the specified unit: for example, ±500KB when using the MB suffix.

2.5.2. Resizing a partition with parted

This procedure resizes a disk partition using the **parted** utility.

Prerequisites

- If you want to shrink a partition, back up the data that are stored on it.

WARNING

Shrinking a partition might result in data loss on the partition.

- If you want to resize a partition to be larger than 2TiB, the disk must be formatted with the GUID Partition Table (GPT). For details on how to format the disk, see Section 2.2, "Creating a partition table on a disk".

Procedure

1. If you want to shrink the partition, shrink the file system on it first so that it is not larger than the resized partition. Note that XFS does not support shrinking.
2. Start the interactive **parted** shell:

```
# parted block-device
```

   - Replace *block-device* with the path to the device where you want to resize a partition: for example, /**dev**/**sda**.

3. View the current partition table:

```
(parted) print
```

   From the partition table, determine:

   - The minor number of the partition
   - The location of the existing partition and its new ending point after resizing

4. Resize the partition:

```
(parted) resizepart minor-number new-end
```

- Replace *minor-number* with the minor number of the partition that you are resizing: for example, **3**.
- Replace *new-end* with the size that determines the new ending point of the resized partition, counting from the beginning of the disk. You can use size suffixes, such as **512MiB**, **20GiB**, or **1.5TiB**. The default size megabytes.

Example 2.4. Extending a partition

For example, to extend a partition located at the beginning of the disk to be 2GiB in size, use:

```
(parted) resizepart 1 2GiB
```

The changes start taking place as soon as you enter this command, so review it before executing it.

5. View the partition table to confirm that the resized partition is in the partition table with the correct size:

```
(parted) print
```

6. Exit the **parted** shell:

```
(parted) quit
```

7. Verify that the kernel recognizes the new partition:

```
# cat /proc/partitions
```

8. If you extended the partition, extend the file system on it as well. See (reference) for details.

Additional resources

- The **parted(8)** man page.

CHAPTER 3. OVERVIEW OF PERSISTENT NAMING ATTRIBUTES

As a system administrator, you need to refer to storage volumes using persistent naming attributes to build storage setups that are reliable over multiple system boots.

3.1. DISADVANTAGES OF NON-PERSISTENT NAMING ATTRIBUTES

Red Hat Enterprise Linux provides a number of ways to identify storage devices. It is important to use the correct option to identify each device when used in order to avoid inadvertently accessing the wrong device, particularly when installing to or reformatting drives.

Traditionally, non-persistent names in the form of /**dev**/**sd***(major number)(minor number)* are used on Linux to refer to storage devices. The major and minor number range and associated **sd** names are allocated for each device when it is detected. This means that the association between the major and minor number range and associated **sd** names can change if the order of device detection changes.

Such a change in the ordering might occur in the following situations:

- The parallelization of the system boot process detects storage devices in a different order with each system boot.
- A disk fails to power up or respond to the SCSI controller. This results in it not being detected by the normal device probe. The disk is not accessible to the system and subsequent devices will have their major and minor number range, including the associated **sd** names shifted down. For example, if a disk normally referred to as **sdb** is not detected, a disk that is normally referred to as **sdc** would instead appear as **sdb**.
- A SCSI controller (host bus adapter, or HBA) fails to initialize, causing all disks connected to that HBA to not be detected. Any disks connected to subsequently probed HBAs are assigned different major and minor number ranges, and different associated **sd** names.
- The order of driver initialization changes if different types of HBAs are present in the system. This causes the disks connected to those HBAs to be detected in a different order. This might also occur if HBAs are moved to different PCI slots on the system.
- Disks connected to the system with Fibre Channel, iSCSI, or FCoE adapters might be inaccessible at the time the storage devices are probed, due to a storage array or intervening switch being powered off, for example. This might occur when a system reboots after a power failure, if the storage array takes longer to come online than the system take to boot. Although some Fibre Channel drivers support a mechanism to specify a persistent SCSI target ID to WWPN mapping, this does not cause the major and minor number ranges, and the associated **sd** names to be reserved; it only provides consistent SCSI target ID numbers.

These reasons make it undesirable to use the major and minor number range or the associated **sd** names when referring to devices, such as in the **/etc/fstab** file. There is the possibility that the wrong device will be mounted and data corruption might result.

Occasionally, however, it is still necessary to refer to the **sd** names even when another mechanism is used, such as when errors are reported by a device. This is because the Linux kernel uses **sd** names (and also SCSI host/channel/target/LUN tuples) in kernel messages regarding the device.

3.2. FILE SYSTEM AND DEVICE IDENTIFIERS

This sections explains the difference between persistent attributes identifying file systems and block devices.

File system identifiers

File system identifiers are tied to a particular file system created on a block device. The identifier is also stored as part of the file system. If you copy the file system to a different device, it still carries the same file system identifier. On the other hand, if you rewrite the device, such as by formatting it with the **mkfs** utility, the device loses the attribute.

File system identifiers include:

- Unique identifier (UUID)
- Label

Device identifiers

Device identifiers are tied to a block device: for example, a disk or a partition. If you rewrite the device, such as by formatting it with the **mkfs** utility, the device keeps the attribute, because it is not stored in the file system.

Device identifiers include:

- World Wide Identifier (WWID)
- Partition UUID
- Serial number

Recommendations

- Some file systems, such as logical volumes, span multiple devices. Red Hat recommends accessing these file systems using file system identifiers rather than device identifiers.

3.3. DEVICE NAMES MANAGED BY THE UDEV MECHANISM IN /DEV/DISK/

This section lists different kinds of persistent naming attributes that the **udev** service provides in the /**dev**/**disk**/ directory.

The **udev** mechanism is used for all types of devices in Linux, not just for storage devices. In the case of storage devices, Red Hat Enterprise Linux contains **udev** rules that create symbolic links in the /**dev**/**disk**/ directory. This enables you to refer to storage devices by:

- Their content
- A unique identifier
- Their serial number.

Although **udev** naming attributes are persistent, in that they do not change on their own across system reboots, some are also configurable.

3.3.1. File system identifiers

The UUID attribute in /dev/disk/by-uuid/

Entries in this directory provide a symbolic name that refers to the storage device by a **unique identifier** (UUID) in the content (that is, the data) stored on the device. For example:

```
/dev/disk/by-uuid/3e6be9de-8139-11d1-9106-a43f08d823a6
```

You can use the UUID to refer to the device in the **/etc/fstab** file using the following syntax:

```
UUID=3e6be9de-8139-11d1-9106-a43f08d823a6
```

You can configure the UUID attribute when creating a file system, and you can also change it later on.

The Label attribute in /dev/disk/by-label/

Entries in this directory provide a symbolic name that refers to the storage device by a **label** in the content (that is, the data) stored on the device.

For example:

```
/dev/disk/by-label/Boot
```

You can use the label to refer to the device in the **/etc/fstab** file using the following syntax:

```
LABEL=Boot
```

You can configure the Label attribute when creating a file system, and you can also change it later on.

3.3.2. Device identifiers

The WWID attribute in /dev/disk/by-id/

The World Wide Identifier (WWID) is a persistent, **system-independent identifier** that the SCSI Standard requires from all SCSI devices. The WWID identifier is guaranteed to be unique for every storage device, and independent of the path that is used to access the device. The identifier is a property of the device but is not stored in the content (that is, the data) on the devices.

This identifier can be obtained by issuing a SCSI Inquiry to retrieve the Device Identification Vital Product Data (page **0x83**) or Unit Serial Number (page **0x80**).

Red Hat Enterprise Linux automatically maintains the proper mapping from the WWID-based device name to a current **/dev/sd** name on that system. Applications can use the **/dev/disk/by-id/** name to reference the data on the disk, even if the path to the device changes, and even when accessing the device from different systems.

Example 3.1. WWID mappings

WWID symlink	Non-persistent device	Note
/dev/disk/by-id/scsi-3600508b400105e210000900000490000	**/dev/sda**	A device with a page **0x83** identifier
/dev/disk/by-id/scsi-SSEAGATE_ST373453LW_3HW1RHM6	**/dev/sdb**	A device with a page **0x80** identifier

WWID symlink	Non-persistent device	Note
/dev/disk/by-id/ata-SAMSUNG_MZNLN256HMHQ-000L7_S2WDNX0J336519-part3	**/dev/sdc3**	A disk partition

In addition to these persistent names provided by the system, you can also use **udev** rules to implement persistent names of your own, mapped to the WWID of the storage.

The Partition UUID attribute in /dev/disk/by-partuuid

The Partition UUID (PARTUUID) attribute identifies partitions as defined by GPT partition table.

Example 3.2. Partition UUID mappings

PARTUUID symlink	Non-persistent device
/dev/disk/by-partuuid/4cd1448a-01	**/dev/sda1**
/dev/disk/by-partuuid/4cd1448a-02	**/dev/sda2**
/dev/disk/by-partuuid/4cd1448a-03	**/dev/sda3**

The Path attribute in /dev/disk/by-path/

This attribute provides a symbolic name that refers to the storage device by the **hardware path** used to access the device.

WARNING

The Path attribute is unreliable, and Red Hat does not recommend using it.

3.4. THE WORLD WIDE IDENTIFIER WITH DM MULTIPATH

This section describes the mapping between the World Wide Identifier (WWID) and non-persistent device names in a Device Mapper Multipath configuration.

If there are multiple paths from a system to a device, DM Multipath uses the WWID to detect this. DM Multipath then presents a single "pseudo-device" in the **/dev/mapper/wwid** directory, such as **/dev/mapper/3600508b400105df70000e00000ac0000**.

The command **multipath -l** shows the mapping to the non-persistent identifiers:

- ***Host*:*Channel*:*Target*:*LUN***
- /**dev**/**sd** name
- ***major*:*minor*** number

Example 3.3. WWID mappings in a multipath configuration

An example output of the **multipath -l** command:

```
3600508b400105df70000e00000ac0000 dm-2 vendor,product
[size=20G][features=1 queue_if_no_path][hwhandler=0][rw]
\_ round-robin 0 [prio=0][active]
 \_ 5:0:1:1 sdc 8:32  [active][undef]
 \_ 6:0:1:1 sdg 8:96  [active][undef]
\_ round-robin 0 [prio=0][enabled]
 \_ 5:0:0:1 sdb 8:16  [active][undef]
 \_ 6:0:0:1 sdf 8:80  [active][undef]
```

DM Multipath automatically maintains the proper mapping of each WWID-based device name to its corresponding /**dev**/**sd** name on the system. These names are persistent across path changes, and they are consistent when accessing the device from different systems.

When the **user_friendly_names** feature of DM Multipath is used, the WWID is mapped to a name of the form /**dev**/**mapper**/**mpath*N***. By default, this mapping is maintained in the file /**etc**/**multipath**/**bindings**. These **mpath*N*** names are persistent as long as that file is maintained.

IMPORTANT

If you use **user_friendly_names**, then additional steps are required to obtain consistent names in a cluster.

3.5. LIMITATIONS OF THE UDEV DEVICE NAMING CONVENTION

The following are some limitations of the **udev** naming convention:

- It is possible that the device might not be accessible at the time the query is performed because the **udev** mechanism might rely on the ability to query the storage device when the **udev** rules are processed for a **udev** event. This is more likely to occur with Fibre Channel, iSCSI or FCoE storage devices when the device is not located in the server chassis.
- The kernel might send **udev** events at any time, causing the rules to be processed and possibly causing the /**dev**/**disk**/**by-***/ links to be removed if the device is not accessible.
- There might be a delay between when the **udev** event is generated and when it is processed, such as when a large number of devices are detected and the user-space **udevd** service takes some amount of time to process the rules for each one. This might cause a delay between when the kernel detects the device and when the /**dev**/**disk**/**by-***/ names are available.
- External programs such as **blkid** invoked by the rules might open the device for a brief period of time, making the device inaccessible for other uses.

3.6. LISTING PERSISTENT NAMING ATTRIBUTES

This procedure describes how to find out the persistent naming attributes of non-persistent storage devices.

Procedure

- To list the UUID and Label attributes, use the **lsblk** utility:

```
$ lsblk --fs storage-device
```

For example:

Example 3.4. Viewing the UUID and Label of a file system

```
$ lsblk --fs /dev/sda1

NAME FSTYPE LABEL UUID                                 MOUNTPOINT
sda1 xfs    Boot  afa5d5e3-9050-48c3-acc1-bb30095f3dc4 /boot
```

- To list the PARTUUID attribute, use the **lsblk** utility with the **--output +PARTUUID** option:

```
$ lsblk --output +PARTUUID
```

For example:

Example 3.5. Viewing the PARTUUID attribute of a partition

```
$ lsblk --output +PARTUUID /dev/sda1

NAME MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT PARTUUID
sda1   8:1    0  512M  0 part /boot      4cd1448a-01
```

- To list the WWID attribute, examine the targets of symbolic links in the **/dev/disk/by-id/** directory. For example:

Example 3.6. Viewing the WWID of all storage devices on the system

```
$ file /dev/disk/by-id/*

/dev/disk/by-id/ata-QEMU_HARDDISK_QM00001
symbolic link to ../../sda
/dev/disk/by-id/ata-QEMU_HARDDISK_QM00001-part1
symbolic link to ../../sda1
/dev/disk/by-id/ata-QEMU_HARDDISK_QM00001-part2
symbolic link to ../../sda2
/dev/disk/by-id/dm-name-rhel_rhel8-root
symbolic link to ../../dm-0
/dev/disk/by-id/dm-name-rhel_rhel8-swap
symbolic link to ../../dm-1
/dev/disk/by-id/dm-uuid-LVM-
QIWtEHtXGobe5bewlIUDivKOz5ofkgFhP0RMFsNyySVihqEl2cWWbR7MjXJolD6g
symbolic link to ../../dm-1
/dev/disk/by-id/dm-uuid-LVM-
```

```
QIWtEHtXGobe5bewlIUDivKOz5ofkgFhXqH2M45hD2H9nAf2qfWSrlRLhzfMyOKd
symbolic link to ../../dm-0
/dev/disk/by-id/lvm-pv-uuid-atlr2Y-vuMo-ueoH-CpMG-4JuH-AhEF-wu4QQm
symbolic link to ../../sda2
```

3.7. MODIFYING PERSISTENT NAMING ATTRIBUTES

This procedure describes how to change the UUID or Label persistent naming attribute of a file system.

NOTE

Changing **udev** attributes happens in the background and might take a long time. The **udevadm settle** command waits until the change is fully registered, which ensures that your next command will be able to utilize the new attribute correctly.

In the following commands:

- Replace *new-uuid* with the UUID you want to set; for example, **1cdfbc07-1c90-4984-b5ec-f61943f5ea50**. You can generate a UUID using the **uuidgen** command.
- Replace *new-label* with a label; for example, **backup_data**.

Prerequisites

- If you are modifying the attributes of an XFS file system, unmount it first.

Procedure

- To change the UUID or Label attributes of an **XFS** file system, use the **xfs_admin** utility:

```
# xfs_admin -U new-uuid -L new-label storage-device
# udevadm settle
```

- To change the UUID or Label attributes of an **ext4**, **ext3**, or **ext2** file system, use the **tune2fs** utility:

```
# tune2fs -U new-uuid -L new-label storage-device
# udevadm settle
```

- To change the UUID or Label attributes of a swap volume, use the **swaplabel** utility:

```
# swaplabel --uuid new-uuid --label new-label swap-device
# udevadm settle
```

CHAPTER 4. USING NVDIMM PERSISTENT MEMORY STORAGE

As a system administrator, you can enable and manage various types of storage on Non-Volatile Dual In-line Memory Modules (NVDIMM) devices connected to your system.

For installing Red Hat Enterprise Linux 8 on NVDIMM storage, see Installing to an NVDIMM device instead.

4.1. THE NVDIMM PERSISTENT MEMORY TECHNOLOGY

NVDIMM persistent memory, also called storage class memory or **pmem**, is a combination of memory and storage.

NVDIMM combines the durability of storage with the low access latency and the high bandwidth of dynamic RAM (DRAM):

- NVDIMM storage is byte-addressable, so it can be accessed by using the CPU load and store instructions. In addition to the **read()** and **write()** system calls, which are required for accessing traditional block-based storage, NVDIMM also supports direct load and store programming model.
- The performance characteristics of NVDIMM are similar to DRAM with very low access latency, typically in the tens to hundreds of nanoseconds.
- Data stored on NVDIMM are preserved when the power is off, like with storage.
- The direct access (DAX) technology enables applications to memory map storage directly, without going through the system page cache. This frees up DRAM for other purposes.

NVDIMM is beneficial in use cases such as:

Databases

The reduced storage access latency on NVDIMM can dramatically improve database performance.

Rapid restart

Rapid restart is also called the warm cache effect. For example, a file server has none of the file contents in memory after starting. As clients connect and read or write data, that data is cached in the page cache. Eventually, the cache contains mostly hot data. After a reboot, the system must start the process again on traditional storage.
NVDIMM enables an application to keep the warm cache across reboots if the application is designed properly. In this example, there would be no page cache involved: the application would cache data directly in the persistent memory.

Fast write-cache

File servers often do not acknowledge a client's write request until the data is on durable media. Using NVDIMM as a fast write cache enables a file server to acknowledge the write request quickly thanks to the low latency.

4.2. NVDIMM INTERLEAVING AND REGIONS

NVDIMM devices support grouping into interleaved regions.

NVDIMM devices can be grouped into interleave sets in the same way as regular DRAM. An interleave set is similar to a RAID 0 level (stripe) configuration across multiple DIMMs. An Interleave set is also called a *region*.

Interleaving has the following advantages:

- NVDIMM devices benefit from increased performance when they are configured into interleave sets.
- Interleaving can combine multiple smaller NVDIMM devices into a larger logical device.

NVDIMM interleave sets are configured in the system BIOS or UEFI firmware.

Red Hat Enterprise Linux creates one region device for each interleave set.

4.3. NVDIMM NAMESPACES

NVDIMM regions are divided into one or more namespaces. Namespaces enable you to access the device using different methods, based on the type of the namespace.

Some NVDIMM devices do not support multiple namespaces on a region:

- If your NVDIMM device supports labels, you can subdivide the region into namespaces.
- If your NVDIMM device does not support labels, the region can only contain a single namespace. In that case, Red Hat Enterprise Linux creates a default namespace that covers the entire region.

4.4. NVDIMM ACCESS MODES

You can configure NVDIMM namespaces to use either of the following modes:

sector

Presents the storage as a fast block device. This mode is useful for legacy applications that have not been modified to use NVDIMM storage, or for applications that make use of the full I/O stack, including Device Mapper.
A **sector** device can be used in the same way as any other block device on the system. You can create partitions or file systems on it, configure it as part of a software RAID set, or use it as the cache device for **dm-cache**.

Devices in this mode are available at **/dev/pmem*N*s**. See the **blockdev** value listed after creating the namespace.

devdax, or device direct access (DAX)

Enables NVDIMM devices to support direct access programming as described in the Storage Networking Industry Association (SNIA) Non-Volatile Memory (NVM) Programming Model specification. In this mode, I/O bypasses the storage stack of the kernel. Therefore, no Device Mapper drivers can be used.
Device DAX provides raw access to NVDIMM storage by using a DAX character device node. Data on a **devdax** device can be made durable using CPU cache flushing and fencing instructions. Certain databases and virtual machine hypervisors might benefit from this mode. File systems cannot be created on **devdax** devices.

Devices in this mode are available at **/dev/dax*N.M*.** See the **chardev** value listed after creating the namespace.

fsdax, or file system direct access (DAX)

Enables NVDIMM devices to support direct access programming as described in the Storage Networking Industry Association (SNIA) Non-Volatile Memory (NVM) Programming Model specification. In this mode, I/O bypasses the storage stack of the kernel, and many Device Mapper drivers therefore cannot be used.
You can create file systems on file system DAX devices.

Devices in this mode are available at **/dev/pmem*N***. See the **blockdev** value listed after creating the namespace.

IMPORTANT

The file system DAX technology is provided only as a Technology Preview, and is not supported by Red Hat.

raw

Presents a memory disk that does not support DAX. In this mode, namespaces have several limitations and should not be used.
Devices in this mode are available at **/dev/pmem*N***. See the **blockdev** value listed after creating the namespace.

4.5. CREATING A SECTOR NAMESPACE ON AN NVDIMM TO ACT AS A BLOCK DEVICE

You can configure an NVDIMM device in sector mode, which is also called *legacy mode*, to support traditional, block-based storage.

You can either:

- reconfigure an existing namespace to sector mode, or
- create a new sector namespace if there is available space.

Prerequisites

- An NVDIMM device is attached to your system.

4.5.1. Installing ndctl

This procedure installs the **ndctl** utility, which is used to configure and monitor NVDIMM devices.

Procedure

- To install the **ndctl** utility, use the following command:

```
# yum install ndctl
```

4.5.2. Reconfiguring an existing NVDIMM namespace to sector mode

This procedure reconfigures an NVDIMM namespace to sector mode for use as a fast block device.

WARNING

Reconfiguring a namespace deletes all data previously stored on the namespace.

Prerequisites

- The **ndctl** utility is installed. See Section 4.5.1, “Installing ndctl”.

Procedure

1. Reconfigure the selected namespace to sector mode:

```
# ndctl create-namespace \
      --force \
      --reconfig=namespace-ID \
      --mode=sector
```

Example 4.1. Reconfiguring namespace1.0 in sector mode

To reconfigure the **namespace1.0** namespace to use **sector** mode:

```
# ndctl create-namespace \
      --force \
      --reconfig=namespace1.0 \
      --mode=sector

{
  "dev":"namespace1.0",
  "mode":"sector",
  "size":"11.99 GiB (12.87 GB)",
  "uuid":"5805480e-90e6-407e-96a4-23e1cde2ed78",
  "raw_uuid":"879d9e9f-fd43-4ed5-b64f-3bcd0781391a",
  "sector_size":4096,
  "blockdev":"pmem1s",
  "numa_node":1
}
```

2. The reconfigured namespace is now available under the **/dev** directory as **/dev/pmem*N*s**.

Additional resources

- The **ndctl-create-namespace(1)** man page

4.5.3. Creating a new NVDIMM namespace in sector mode

This procedure creates a new sector namespace on an NVDIMM device, enabling you to use it as a traditional block device.

Prerequisites

- The **ndctl** utility is installed. See Section 4.5.1, “Installing ndctl”.
- The NVDIMM device supports labels.

Procedure

1. List the **pmem** regions on your system that have available space. In the following example, space is available in the **region5** and **region4** regions:

```
# ndctl list --regions

[
  {
    "dev":"region5",
    "size":270582939648,
    "available_size":270582939648,
    "type":"pmem",
    "iset_id":-7337419320239190016
  },
  {
    "dev":"region4",
    "size":270582939648,
    "available_size":270582939648,
    "type":"pmem",
    "iset_id":-137289417188962304
  }
]
```

2. On any of the available regions, allocate one or more namespaces:

```
# ndctl create-namespace \
      --mode=sector \
      --region=regionN \
      --size=namespace-size
```

Example 4.2. Creating a namespace on a region

The following command creates a 36-GiB sector namespace on **region4**:

```
# ndctl create-namespace \
      --mode=sector \
      --region=region4 \
      --size=36G
```

3. The new namespace is now available under the **/dev** directory as **/dev/pmem*N*s**.

Additional resources

- The **ndctl-create-namespace(1)** man page

4.6. CREATING A DEVICE DAX NAMESPACE ON AN NVDIMM

You can configure an NVDIMM device in device DAX mode to support character storage with direct access capabilities.

You can either:

- reconfigure an existing namespace to device DAX mode, or
- create a new device DAX namespace if there is available space.

Prerequisites

- An NVDIMM device is attached to your system.

4.6.1. NVDIMM in device direct access mode

Device direct access (device DAX, **devdax**) provides a means for applications to directly access storage, without the involvement of a file system. The benefit of device DAX is that it provides a guaranteed fault granularity, which can be configured using the **--align** option of the **ndctl** utility

For the Intel 64 and AMD64 architecture, the following fault granularities are supported:

- 4 KiB
- 2 MiB
- 1 GiB

Device DAX nodes support only the following system calls:

- **open()**
- **close()**
- **mmap()**

The **read()** and **write()** variants are not supported because the device DAX use case is tied to persistent memory programming.

4.6.2. Installing ndctl

This procedure installs the **ndctl** utility, which is used to configure and monitor NVDIMM devices.

Procedure

- To install the **ndctl** utility, use the following command:

```
# yum install ndctl
```

4.6.3. Reconfiguring an existing NVDIMM namespace to device DAX mode

This procedure reconfigures a namespace on an NVDIMM device to device DAX mode, and enables you to store data on the namespace.

WARNING

Reconfiguring a namespace deletes all data previously stored on the namespace.

Prerequisites

- The **ndctl** utility is installed. See Section 4.6.2, “Installing ndctl”.

Procedure

1. List all namespaces on your system:

```
# ndctl list --namespaces --idle

[
  {
    "dev":"namespace1.0",
    "mode":"raw",
    "size":34359738368,
    "state":"disabled",
    "numa_node":1
  },
  {
    "dev":"namespace0.0",
    "mode":"raw",
    "size":34359738368,
    "state":"disabled",
    "numa_node":0
  }
]
```

2. Reconfigure any namespace:

```
# ndctl create-namespace \
      --force \
      --mode=devdax \
      --reconfig=namespace-ID
```

Example 4.3. Reconfiguring a namespace as device DAX

The following command reconfigures **namespace0.0** for data storage that supports DAX. It is aligned to a 2-MiB fault granularity to ensure that the operating system faults in 2-MiB pages at a time:

```
# ndctl create-namespace \
      --force \
      --mode=devdax \
      --align=2M \
      --reconfig=namespace0.0
```

3. The namespace is now available at the /**dev**/**dax*N.M*** path.

Additional resources

- The **ndctl-create-namespace(1)** man page

4.6.4. Creating a new NVDIMM namespace in device DAX mode

This procedure creates a new device DAX namespace on an NVDIMM device, enabling you to store data on the namespace.

Prerequisites

- The **ndctl** utility is installed. See Section 4.6.2, “Installing ndctl”.
- The NVDIMM device supports labels.

Procedure

1. List the **pmem** regions on your system that have available space. In the following example, space is available in the **region5** and **region4** regions:

```
# ndctl list --regions

[
  {
    "dev":"region5",
    "size":270582939648,
    "available_size":270582939648,
    "type":"pmem",
    "iset_id":-7337419320239190016
  },
  {
    "dev":"region4",
    "size":270582939648,
    "available_size":270582939648,
    "type":"pmem",
    "iset_id":-137289417188962304
  }
]
```

2. On any of the available regions, allocate one or more namespaces:

```
# ndctl create-namespace \
      --mode=devdax \
      --region=regionN \
      --size=namespace-size
```

Example 4.4. Creating a namespace on a region

The following command creates a 36-GiB device DAX namespace on **region4**. It is aligned to a 2-MiB fault granularity to ensure that the operating system faults in 2-MiB pages at a time:

```
# ndctl create-namespace \
```

```
        --mode=devdax \
        --region=region4 \
        --align=2M \
        --size=36G

{
  "dev":"namespace1.2",
  "mode":"devdax",
  "map":"dev",
  "size":"35.44 GiB (38.05 GB)",
  "uuid":"5ae01b9c-1ebf-4fb6-bc0c-6085f73d31ee",
  "raw_uuid":"4c8be2b0-0842-4bcb-8a26-4bbd3b44add2",
  "daxregion":{
    "id":1,
    "size":"35.44 GiB (38.05 GB)",
    "align":2097152,
    "devices":[
      {
        "chardev":"dax1.2",
        "size":"35.44 GiB (38.05 GB)"
      }
    ]
  },
  "numa_node":1
}
```

3. The namespace is now available at the **/dev/dax*N.M*** path.

Additional resources

- The **ndctl-create-namespace(1)** man page

4.7. CREATING A FILE SYSTEM DAX NAMESPACE ON AN NVDIMM

You can configure an NVDIMM device in file system DAX mode to support a file system with direct access capabilities.

You can either:

- reconfigure an existing namespace to file system DAX mode, or
- create a new file system DAX namespace if there is available space.

IMPORTANT

The file system DAX technology is provided only as a Technology Preview, and is not supported by Red Hat.

Prerequisites

- An NVDIMM device is attached to your system.

4.7.1. NVDIMM in file system direct access mode

When an NVDIMM device is configured in file system direct access (file system DAX, **fsdax**) mode, a file system can be created on top of it.

Any application that performs an **mmap()** operation on a file on this file system gets direct access to its storage. This enables the direct access programming model on NVDIMM. The file system must be mounted with the **-o dax** option in order for direct mapping to happen.

Per-page metadata allocation
This mode requires allocating per-page metadata in the system DRAM or on the NVDIMM device itself. The overhead of this data structure is 64 bytes per each 4-KiB page:

- On small devices, the amount of overhead is small enough to fit in DRAM with no problems. For example, a 16-GiB namespace only requires 256 MiB for page structures. Because NVDIMM devices are usually small and expensive, storing the page tracking data structures in DRAM is preferable.
- On NVDIMM devices that are be terabytes in size or larger, the amount of memory required to store the page tracking data structures might exceed the amount of DRAM in the system. One TiB of NVDIMM requires 16 GiB just for page structures. As a result, storing the data structures on the NVDIMM itself is preferable in such cases.

You can configure where per-page metadata are stored using the **--map** option when configuring a namespace:

- To allocate in the system RAM, use **--map=mem**.
- To allocate on the NVDIMM, use **--map=dev**.

Partitions and file systems on fsdax
When creating partitions on an **fsdax** device, partitions must be aligned on page boundaries. On the Intel 64 and AMD64 architecture, at least 4 KiB alignment is required for the start and end of the partition. 2 MiB is the preferred alignment.

On Red Hat Enterprise Linux 8, both the XFS and ext4 file system can be created on NVDIMM as a Technology Preview.

4.7.2. Installing ndctl

This procedure installs the **ndctl** utility, which is used to configure and monitor NVDIMM devices.

Procedure

- To install the **ndctl** utility, use the following command:

```
# yum install ndctl
```

4.7.3. Reconfiguring an existing NVDIMM namespace to file system DAX mode

This procedure reconfigures a namespace on an NVDIMM device to file system DAX mode, and enables you to store files on the namespace.

WARNING

Reconfiguring a namespace deletes all data previously stored on the namespace.

Prerequisites

- The **ndctl** utility is installed. See Section 4.7.2, “Installing ndctl”.

Procedure

1. List all namespaces on your system:

```
# ndctl list --namespaces --idle

[
  {
    "dev":"namespace1.0",
    "mode":"raw",
    "size":34359738368,
    "state":"disabled",
    "numa_node":1
  },
  {
    "dev":"namespace0.0",
    "mode":"raw",
    "size":34359738368,
    "state":"disabled",
    "numa_node":0
  }
]
```

2. Reconfigure any namespace:

```
# ndctl create-namespace \
      --force \
      --mode=fsdax \
      --reconfig=namespace-ID
```

Example 4.5. Reconfiguring a namespace as file system DAX

To use **namespace0.0** for a file system that supports DAX, use the following command:

```
# ndctl create-namespace \
      --force \
      --mode=fsdax \
      --reconfig=namespace0.0

{
  "dev":"namespace0.0",
  "mode":"fsdax",
```

```
  "size":"32.00 GiB (34.36 GB)",
  "uuid":"ab91cc8f-4c3e-482e-a86f-78d177ac655d",
  "blockdev":"pmem0",
  "numa_node":0
 }
```

3. The namespace is now available at the /**dev**/**pmem*N*** path.

Additional resources

- The **ndctl-create-namespace(1)** man page

4.7.4. Creating a new NVDIMM namespace in file system DAX mode

This procedure creates a new file system DAX namespace on an NVDIMM device, enabling you to store files on the namespace.

Prerequisites

- The **ndctl** utility is installed. See Section 4.7.2, “Installing ndctl”.
- The NVDIMM device supports labels.

Procedure

1. List the **pmem** regions on your system that have available space. In the following example, space is available in the **region5** and **region4** regions:

```
# ndctl list --regions

[
  {
    "dev":"region5",
    "size":270582939648,
    "available_size":270582939648,
    "type":"pmem",
    "iset_id":-7337419320239190016
  },
  {
    "dev":"region4",
    "size":270582939648,
    "available_size":270582939648,
    "type":"pmem",
    "iset_id":-137289417188962304
  }
]
```

2. On any of the available regions, allocate one or more namespaces:

```
# ndctl create-namespace \
        --mode=fsdax \
        --region=regionN \
        --size=namespace-size
```

Example 4.6. Creating a namespace on a region

The following command creates a 36-GiB file system DAX namespace on **region4**:

```
# ndctl create-namespace \
        --mode=fsdax \
        --region=region4 \
        --size=36G

{
  "dev":"namespace4.0",
  "mode":"fsdax",
  "size":"35.44 GiB (38.05 GB)",
  "uuid":"9c5330b5-dc90-4f7a-bccd-5b558fa881fe",
  "blockdev":"pmem4",
  "numa_node":0
}
```

3. The namespace is now available at the /**dev**/**pmem*N*** path.

Additional resources

- The **ndctl-create-namespace(1)** man page

4.7.5. Creating a file system on a file system DAX device

This procedure creates a file system on a file system DAX device and mounts the file system.

Procedure

1. Optionally, create a partition on the file system DAX device. See Section 2.3, "Creating a partition".
   By default, the **parted** tool aligns partitions on 1 MiB boundaries. For the first partition, specify 2 MiB as the start of the partition. If the size of the partition is a multiple of 2 MiB, all other partitions are also aligned.

2. Create an XFS or ext4 file system on the partition or the NVDIMM device.
   For XFS, disable shared copy-on-write data extents when creating the file system:

   ```
   # mkfs.xfs -m reflink=0 fsdax-partition-or-device
   ```

3. Mount the file system with the **-o fsdax** mount option:

   ```
   # mount -o fsdax fsdax-partition-or-device mount-point
   ```

4. Applications can now use persistent memory and create files in the *mount-point* directory, open the files, and use the **mmap** operation to map the files for direct access.

Additional resources

- The **mkfs.xfs(8)** man page

4.8. TROUBLESHOOTING NVDIMM PERSISTENT MEMORY

You can detect and fix different kinds of errors on NVDIMM devices.

Prerequisites

- An NVDIMM device is connected to your system and configured.

4.8.1. Installing ndctl

This procedure installs the **ndctl** utility, which is used to configure and monitor NVDIMM devices.

Procedure

- To install the **ndctl** utility, use the following command:

```
# yum install ndctl
```

4.8.2. Monitoring NVDIMM health using S.M.A.R.T.

Some NVDIMM devices support Self-Monitoring, Analysis and Reporting Technology (S.M.A.R.T.) interfaces for retrieving health information.

IMPORTANT

Monitor NVDIMM health regularly to prevent data loss. If S.M.A.R.T. reports problems with the health status of an NVDIMM device, replace it as described in Section 4.8.3, “Detecting and replacing a broken NVDIMM device”.

Prerequisites

- On some systems, the **acpi_ipmi** driver must be loaded to retrieve health information using the following command:

```
# modprobe acpi_ipmi
```

Procedure

- To access the health information, use the following command:

```
# ndctl list --dimms --health

...
  {
    "dev":"nmem0",
    "id":"802c-01-1513-b3009166",
    "handle":1,
    "phys_id":22,
    "health":
    {
      "health_state":"ok",
      "temperature_celsius":25.000000,
      "spares_percentage":99,
```

```
        "alarm_temperature":false,
        "alarm_spares":false,
        "temperature_threshold":50.000000,
        "spares_threshold":20,
        "life_used_percentage":1,
        "shutdown_state":"clean"
      }
    }
...
```

Additional resources

- The **ndctl-list(1)** man page

4.8.3. Detecting and replacing a broken NVDIMM device

If you find error messages related to NVDIMM reported in your system log or by S.M.A.R.T., it might mean an NVDIMM device is failing. In that case, it is necessary to:

1. Detect which NVDIMM device is failing
2. Back up data stored on it
3. Physically replace the device

Procedure

1. To detect the broken device, use the following command:

```
# ndctl list --dimms --regions --health --media-errors --human
```

The **badblocks** field shows which NVDIMM is broken. Note its name in the **dev** field.

Example 4.7. Health status of NVDIMM devices

In the following example, the NVDIMM named **nmem0** is broken:

```
# ndctl list --dimms --regions --health --media-errors --human

...
  "regions":[
    {
      "dev":"region0",
      "size":"250.00 GiB (268.44 GB)",
      "available_size":0,
      "type":"pmem",
      "numa_node":0,
      "iset_id":"0xXXXXXXXXXXXXXXXX",
      "mappings":[
        {
          "dimm":"nmem1",
          "offset":"0x10000000",
          "length":"0x1f40000000",
          "position":1
        },
```

```
        {
          "dimm":"nmem0",
          "offset":"0x10000000",
          "length":"0x1f40000000",
          "position":0
        }
      ],
      "badblock_count":1,
      "badblocks":[
        {
          "offset":65536,
          "length":1,
          "dimms":[
            "nmem0"
          ]
        }
      ],
      "persistence_domain":"memory_controller"
    }
  ]
}
```

2. Use the following command to find the **phys_id** attribute of the broken NVDIMM:

```
# ndctl list --dimms --human
```

From the previous example, you know that **nmem0** is the broken NVDIMM. Therefore, find the **phys_id** attribute of **nmem0**.

Example 4.8. The phys_id attributes of NVDIMMs

In the following example, the **phys_id** is **0x10**:

```
# ndctl list --dimms --human

[
  {
    "dev":"nmem1",
    "id":"XXXX-XX-XXXX-XXXXXXXX",
    "handle":"0x120",
    "phys_id":"0x1c"
  },
  {
    "dev":"nmem0",
    "id":"XXXX-XX-XXXX-XXXXXXXX",
    "handle":"0x20",
    "phys_id":"0x10",
    "flag_failed_flush":true,
    "flag_smart_event":true
  }
]
```

3. Use the following command to find the memory slot of the broken NVDIMM:

```
# dmidecode
```

In the output, find the entry where the **Handle** identifier matches the **phys_id** attribute of the broken NVDIMM. The **Locator** field lists the memory slot used by the broken NVDIMM.

Example 4.9. NVDIMM Memory Slot Listing

In the following example, the **nmem0** device matches the **0x0010** identifier and uses the **DIMM-XXX-YYYY** memory slot:

```
# dmidecode

...
Handle 0x0010, DMI type 17, 40 bytes
Memory Device
        Array Handle: 0x0004
        Error Information Handle: Not Provided
        Total Width: 72 bits
        Data Width: 64 bits
        Size: 125 GB
        Form Factor: DIMM
        Set: 1
        Locator: DIMM-XXX-YYYY
        Bank Locator: Bank0
        Type: Other
        Type Detail: Non-Volatile Registered (Buffered)
...
```

4. Back up all data in the namespaces on the NVDIMM. If you do not back up the data before replacing the NVDIMM, the data will be lost when you remove the NVDIMM from your system.

WARNING

In some cases, such as when the NVDIMM is completely broken, the backup might fail.

To prevent this, regularly monitor your NVDIMM devices using S.M.A.R.T. as described in Section 4.8.2, "Monitoring NVDIMM health using S.M.A.R.T." and replace failing NVDIMMs before they break.

Use the following command to list the namespaces on the NVDIMM:

```
# ndctl list --namespaces --dimm=DIMM-ID-number
```

Example 4.10. NVDIMM namespaces listing

In the following example, the **nmem0** device contains the **namespace0.0** and **namespace0.2** namespaces, which you need to back up:

```
# ndctl list --namespaces --dimm=0

[
  {
    "dev":"namespace0.2",
    "mode":"sector",
    "size":67042312192,
    "uuid":"XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX",
    "raw_uuid":"XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX",
    "sector_size":4096,
    "blockdev":"pmem0.2s",
    "numa_node":0
  },
  {
    "dev":"namespace0.0",
    "mode":"sector",
    "size":67042312192,
    "uuid":"XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX",
    "raw_uuid":"XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX",
    "sector_size":4096,
    "blockdev":"pmem0s",
    "numa_node":0
  }
]
```

5. Replace the broken NVDIMM physically.

Additional resources

- The **ndctl-list(1)** man page
- The **dmidecode(8)** man page

CHAPTER 5. DISCARDING UNUSED BLOCKS

You can perform or schedule discard operations on block devices that support them.

5.1. BLOCK DISCARD OPERATIONS

Block discard operations discard blocks that are no longer in use by a mounted file system. They are useful on:

- Solid-state drives (SSDs)
- Thinly-provisioned storage

Requirements

The block device underlying the file system must support physical discard operations.

Physical discard operations are supported if the value in the **/sys/block/*device*/queue/discard_max_bytes** file is not zero.

5.2. TYPES OF BLOCK DISCARD OPERATIONS

You can run discard operations using different methods:

Batch discard

Are run explicitly by the user. They discard all unused blocks in the selected file systems.

Online discard

Are specified at mount time. They run in real time without user intervention. Online discard operations discard only the blocks that are transitioning from used to free.

Periodic discard

Are batch operations that are run regularly by a **systemd** service.

All types are supported by the XFS and ext4 file systems and by VDO.

Recommendations

Red Hat recommends that you use batch or periodic discard.

Use online discard only if:

- the system's workload is such that batch discard is not feasible, or
- online discard operations are necessary to maintain performance.

5.3. PERFORMING BATCH BLOCK DISCARD

This procedure performs a batch block discard operation to discard unused blocks on a mounted file system.

Prerequisites

- The file system is mounted.
- The block device underlying the file system supports physical discard operations.

Procedure

- Use the **fstrim** utility:
  - To perform discard only on a selected file system, use:

```
# fstrim mount-point
```

  - To perform discard on all mounted file systems, use:

```
# fstrim --all
```

If you execute the **fstrim** command on:

- a device that does not support discard operations, or
- a logical device (LVM or MD) composed of multiple devices, where any one of the device does not support discard operations,

the following message displays:

```
# fstrim /mnt/non_discard

fstrim: /mnt/non_discard: the discard operation is not supported
```

Additional resources

- The **fstrim(8)** man page

5.4. ENABLING ONLINE BLOCK DISCARD

This procedure enables online block discard operations that automatically discard unused blocks on all supported file systems.

Procedure

- Enable online discard at mount time:
  - When mounting a file system manually, add the **-o discard** mount option:

```
# mount -o discard device mount-point
```

  - When mounting a file system persistently, add the **discard** option to the mount entry in the **/etc/fstab** file.

Additional resources

- The **mount(8)** man page
- The **fstab(5)** man page

5.5. ENABLING ONLINE BLOCK DISCARD USING RHEL SYSTEM ROLES

This section describes how to enable online block discard using the **storage** role.

Prerequisites

- An Ansible playbook including the **storage** role exists.

For information on how to apply such a playbook, see Applying a role.

5.5.1. Example Ansible playbook to enable online block discard

This section provides an example Ansible playbook. This playbook applies the **storage** role to mount an XFS file system with online block discard enabled.

```
---
- hosts: all
  vars:
    storage_volumes:
      - name: barefs
        type: disk
        disks:
          - sdb
        fs_type: xfs
        mount_point: /mnt/data
        mount_options: discard
  roles:
    - rhel-system-roles.storage
```

5.6. ENABLING PERIODIC BLOCK DISCARD

This procedure enables a **systemd** timer that regularly discards unused blocks on all supported file systems.

Procedure

- Enable and start the **systemd** timer:

```
# systemctl enable --now fstrim.timer
```

CHAPTER 6. GETTING STARTED WITH ISCSI

Red Hat Enterprise Linux 8 uses the **targetcli** shell as a command-line interface to perform the following operations:

- Add, remove, view, and monitor iSCSI storage interconnects to utilize iSCSI hardware.
- Export local storage resources that are backed by either files, volumes, local SCSI devices, or by RAM disks to remote systems.

The **targetcli** tool has a tree-based layout including built-in tab completion, auto-complete support, and inline documentation.

6.1. ADDING AN ISCSI TARGET

As a system administrator, you can add an iSCSI targets using the **targetcli** tool.

6.1.1. Installing targetcli

Install the **targetcli** tool to add, monitor, and remove iSCSI storage interconnects .

Procedure

1. Install **targetcli**:

```
# yum install targetcli
```

2. Start the target service:

```
# systemctl start target
```

3. Configure target to start at boot time:

```
# systemctl enable target
```

4. Open port **3260** in the firewall and reload the firewall configuration:

```
# firewall-cmd --permanent --add-port=3260/tcp
Success

# firewall-cmd --reload
Success
```

5. View the **targetcli** layout:

```
# targetcli
/> ls
o- /.....[...]
  o- backstores.....[...]
  | o- block.....[Storage Objects: 0]
  | o- fileio.....[Storage Objects: 0]
  | o- pscsi.....[Storage Objects: 0]
```

```
| o- ramdisk.....[Storage Objects: 0]
o- iscsi.....[Targets: 0]
o- loopback.....[Targets: 0]
```

Additional resources

- The **targetcli** man page.

6.1.2. Creating an iSCSI target

Creating an iSCSI target enables the iSCSI initiator of the client to access the storage devices on the server. Both targets and initiators have unique identifying names.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, “Installing targetcli”.

Procedure

1. Navigate to the iSCSI directory:

```
/> iscsi/
```

NOTE

The **cd** command is used to change directories as well as to list the path to move into.

2. Use one of the following options to create an iSCSI target:
   a. Creating an iSCSI target using a default target name:

```
/iscsi> create

Created target
iqn.2003-01.org.linux-iscsi.hostname.x8664:sn.78b473f296ff
Created TPG1
```

   b. Creating an iSCSI target using a specific name:

```
/iscsi> create iqn.2006-04.com.example:444

Created target iqn.2006-04.com.example:444
Created TPG1
Here iqn.2006-04.com.example:444 is target_iqn_name
```

   Replace *iqn.2006-04.com.example:444* with the specific target name.

3. Verify the newly created target:

```
/iscsi> ls

o- iscsi.....[1 Target]
```

```
o- iqn.2006-04.com.example:444.....[1 TPG]
  o- tpg1.....[enabled, auth]
    o- acls.....[0 ACL]
     o- luns.....[0 LUN]
    o- portals.....[0 Portal]
```

Additional resources

- The **targetcli** man page.

6.1.3. iSCSI Backstore

An iSCSI backstore enables support for different methods of storing an exported LUN's data on the local machine. Creating a storage object defines the resources that the backstore uses. An administrator can choose any of the following backstore devices that Linux-IO (LIO) supports:

- **fileio** backstore: Create a **fileio** storage object if you are using regular files on the local file system as disk images. For creating a **fileio** backstore, see Section 6.1.4, "Creating a fileio storage object".
- **block** backstore: Create a **block** storage object if you are using any local block device and logical device. For creating a **block** backstore, see Section 6.1.5, "Creating a block storage object".
- **pscsi** backstore: Create a **pscsi** storage object if your storage object supports direct pass-through of SCSI commands. For creating a **pscsi** backstore, see Section 6.1.6, "Creating a pscsi storage object"
- **ramdisk** backstore: Create a **ramdisk** storage object if you want to create a temporary RAM backed device. For creating a **ramdisk** backstore, see Section 6.1.7, "Creating a Memory Copy RAM disk storage object".

Additional resources

- The **targetcli** man page.

6.1.4. Creating a fileio storage object

fileio storage objects can support either the **write_back** or **write_thru** operations. The **write_back** operation enables the local file system cache. This improves performance but increases the risk of data loss. It is recommended to use **write_back=false** to disable the **write_back** operation in favor of the **write_thru** operation.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, "Installing targetcli".

Procedure

1. Navigate to the backstores directory:

```
/> backstores/
```

2. Create a **fileio** storage object:

```
/> backstores/fileio create file1 /tmp/disk1.img 200M write_back=false

Created fileio file1 with size 209715200
```

3. Verify the created **fileio** storage object:

```
/backstores> ls
```

Additional resources

- The **targetcli** man page.

6.1.5. Creating a block storage object

The block driver allows the use of any block device that appears in the **/sys/block/** directory to be used with Linux-IO (LIO). This includes physical devices (for example, HDDs, SSDs, CDs, DVDs) and logical devices (for example, software or hardware RAID volumes, or LVM volumes).

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, “Installing targetcli”.

Procedure

1. Navigate to the backstores directory:

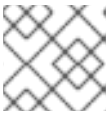
```
/> backstores/
```

2. Create a **block** backstore:

```
/> backstores/block create name=block_backend dev=/dev/sdb

Generating a wwn serial.
Created block storage object block_backend using /dev/vdb.
```

3. Verify the created **block** storage object:

```
/backstores> ls
```



NOTE

You can also create a block backstore on a logical volume.

Additional resources

- The **targetcli** man page.

6.1.6. Creating a pscsi storage object

You can configure, as a backstore, any storage object that supports direct pass-through of SCSI commands without SCSI emulation, and with an underlying SCSI device that appears with **lsscsi** in the **/proc/scsi/scsi** (such as a SAS hard drive) . SCSI-3 and higher is supported with this subsystem.

WARNING

pscsi should only be used by advanced users. Advanced SCSI commands such as for Asymmetric Logical Unit Assignment (ALUAs) or Persistent Reservations (for example, those used by VMware ESX, and vSphere) are usually not implemented in the device firmware and can cause malfunctions or crashes. When in doubt, use **block** backstore for production setups instead.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, "Installing targetcli".

Procedure

1. Navigate to the backstores directory:

```
/> backstores/
```

2. Create a **pscsi** backstore for a physical SCSI device, a TYPE_ROM device using **/dev/sr0** in this example:

```
/> backstores/pscsi/ create name=pscsi_backend dev=/dev/sr0

Generating a wwn serial.
Created pscsi storage object pscsi_backend using /dev/sr0
```

3. Verify the created **pscsi** storage object:

```
/backstores> ls
```

Additional resources

- The **targetcli** man page.

6.1.7. Creating a Memory Copy RAM disk storage object

Memory Copy RAM disks (**ramdisk**) provide RAM disks with full SCSI emulation and separate memory mappings using memory copy for initiators. This provides capability for multi-sessions and is particularly useful for fast and volatile mass storage for production purposes.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, "Installing targetcli".

Procedure

1. Navigate to the backstores directory:

```
/> backstores/
```

2. Create a 1GB RAM disk backstore:

```
/> backstores/ramdisk/ create name=rd_backend size=1GB

Generating a wwn serial.
Created rd_mcp ramdisk rd_backend with size 1GB.
```

3. Verify the created **ramdisk** storage object:

```
/backstores> ls
```

Additional resources

- The **targetcli** man page.

6.1.8. Creating an iSCSI portal

Creating an iSCSI portal adds an IP address and a port to the target that keeps the target enabled.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, "Installing targetcli".
- An iSCSI target associated with a Target Portal Groups (TPG). For more information, see Section 6.1.2, "Creating an iSCSI target" .

Procedure

1. Navigate to the TPG directory:

```
/iscsi> iqn.2006-04.example:444/tpg1/
```

2. Use one of the following options to create an iSCSI portal:

   a. Creating a default portal uses the default iSCSI port **3260** and allows the target to listen to all IP addresses on that port:

```
/iscsi/iqn.20...mple:444/tpg1> portals/ create

Using default IP port 3260
Binding to INADDR_Any (0.0.0.0)
Created network portal 0.0.0.0:3260
```

NOTE

When an iSCSI target is created, a default portal is also created. This portal is set to listen to all IP addresses with the default port number that is: **0.0.0.0:3260**.

To remove the default portal:

/iscsi/iqn-name/tpg1/portals delete ip_address=0.0.0.0 ip_port=3260

b. Creating a portal using a specific IP address:

```
/iscsi/iqn.20...mple:444/tpg1> portals/ create 192.168.122.137

Using default IP port 3260
Created network portal 192.168.122.137:3260
```

3. Verify the newly created portal:

```
/iscsi/iqn.20...mple:444/tpg1> ls

o- tpg..... [enambled, auth]
  o- acls .....[0 ACL]
  o- luns .....[0 LUN]
  o- portals .....[1 Portal]
    o- 192.168.122.137:3260.....[OK]
```

Additional resources

- The **targetcli** man page.

6.1.9. Creating an iSCSI LUN

Logical unit number (LUN) is a physical device that is backed by the iSCSI backstore. Each LUN has a unique number.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, “Installing targetcli”.
- An iSCSI target associated with a Target Portal Groups (TPG). For more information, see Section 6.1.2, “Creating an iSCSI target” .
- Created storage objects. For more information, see Section 6.1.3, “iSCSI Backstore”.

Procedure

1. Create LUNs of already created storage objects:

```
/iscsi/iqn.20...mple:444/tpg1> luns/ create /backstores/ramdisk/rd_backend
Created LUN 0.

/iscsi/iqn.20...mple:444/tpg1> luns/ create /backstores/block/block_backend
Created LUN 1.

/iscsi/iqn.20...mple:444/tpg1> luns/ create /backstores/fileio/file1
Created LUN 2.
```

2. Verify the created LUNs:

```
/iscsi/iqn.20...mple:444/tpg1> ls

o- tpg..... [enambled, auth]
  o- acls .....[0 ACL]
```

```
o- luns .....[3 LUNs]
| o- lun0.....[ramdisk/ramdisk1]
| o- lun1.....[block/block1 (/dev/vdb1)]
| o- lun2.....[fileio/file1 (/foo.img)]
o- portals .....[1 Portal]
  o- 192.168.122.137:3260.....[OK]
```

Default LUN name starts at **0**.

IMPORTANT

By default, LUNs are created with read-write permissions. If a new LUN is added after ACLs are created, LUN automatically maps to all available ACLs and can cause a security risk. To create a LUN with read-only permissions, see Section 6.1.10, "Creating a read-only iSCSI LUN" .

3. Configure ACLs. For more information, see Section 6.1.11, "Creating an iSCSI ACL" .

Additional resources

- The **targetcli** man page.

6.1.10. Creating a read-only iSCSI LUN

By default, LUNs are created with read-write permissions. This procedure describes how to create a read-only LUN.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, "Installing targetcli" .
- An iSCSI target associated with a Target Portal Groups (TPG). For more information, see Section 6.1.2, "Creating an iSCSI target" .
- Created storage objects. For more information, see Section 6.1.3, "iSCSI Backstore" .

Procedure

1. Set read-only permissions:

```
/> set global auto_add_mapped_luns=false

Parameter auto_add_mapped_luns is now 'false'.
```

This prevents the auto mapping of LUNs to existing ACLs allowing the manual mapping of LUNs.

2. Create the LUN:

```
/> iscsi/target_iqn_name/tpg1/acls/initiator_iqn_name/ create
mapped_lun=next_sequential_LUN_number tpg_lun_or_backstore=backstore
write_protect=1
```

Example:

```
/> iscsi/iqn.2006-04.example:444/tpg1/acls/2006-04.com.example.foo:888/ create
mapped_lun=1 tpg_lun_or_backstore=/backstores/block/block2 write_protect=1

Created LUN 1.
Created Mapped LUN 1.
```

3. Verify the created LUN:

```
/> ls

o- / ..... [...]
  o- backstores ..... [...]
  <snip>
  o- iscsi ..... [Targets: 1]
  | o- iqn.2006-04.example:444 ..... [TPGs: 1]
  |   o- tpg1 ..... [no-gen-acls, no-auth]
  |     o- acls ..... [ACLs: 2]
  |     | o- 2006-04.com.example.foo:888 .. [Mapped LUNs: 2]
  |     | | o- mapped_lun0 ..... [lun0 block/disk1 (rw)]
  |     | | o- mapped_lun1 ..... [lun1 block/disk2 (ro)]
  |     o- luns ..... [LUNs: 2]
  |     | o- lun0 ..... [block/disk1 (/dev/vdb)]
  |     | o- lun1 ..... [block/disk2 (/dev/vdc)]
  <snip>
```

The mapped_lun1 line now has (**ro**) at the end (unlike mapped_lun0's (**rw**)) stating that it is read-only.

4. Configure ACLs. For more information, see Section 6.1.11, "Creating an iSCSI ACL" .

Additional resources

- The **targetcli** man page.

6.1.11. Creating an iSCSI ACL

In **targetcli**, Access Control Lists (ACLs) are used to define access rules and each initiator has exclusive access to a LUN. Both targets and initiators have unique identifying names. You must know the unique name of the initiator to configure ACLs. The iSCSI initiators can be found in the **/etc/iscsi/initiatorname.iscsi** file.

Prerequisites

- Installed and running **targetcli**. For more information, see Section 6.1.1, "Installing targetcli".
- An iSCSI target associated with a Target Portal Groups (TPG). For more information, see Section 6.1.2, "Creating an iSCSI target" .

Procedure

1. Navigate to the acls directory

```
/iscsi/iqn.20...mple:444/tpg1> acls/
```

2. Use one of the following options to create an ACL :

   a. Using the initiator name from **/etc/iscsi/initiatorname.iscsi** file on the initiator.

   b. Using a name that is easier to remember, see section Section 6.1.12, "Creating an iSCSI initiator" to ensure ACL matches the initiator.

```
/iscsi/iqn.20...444/tpg1/acls> create iqn.2006-04.com.example.foo:888

Created Node ACL for iqn.2006-04.com.example.foo:888
Created mapped LUN 2.
Created mapped LUN 1.
Created mapped LUN 0.
```

NOTE

The global setting **auto_add_mapped_luns** used in the preceding example, automatically maps LUNs to any created ACL.

You can set user-created ACLs within the TPG node on the target server:

```
/iscsi/iqn.20...scsi:444/tpg1> set attribute generate_node_acls=1
```

3. Verify the created ACL:

```
/iscsi/iqn.20...444/tpg1/acls> ls

o- acls .....[1 ACL]
  o- iqn.2006-04.com.example.foo:888 ....[3 Mapped LUNs, auth]
    o- mapped_lun0 .....[lun0 ramdisk/ramdisk1 (rw)]
    o- mapped_lun1 .....[lun1 block/block1 (rw)]
    o- mapped_lun2 .....[lun2 fileio/file1 (rw)]
```

Additional resources

- The **targetcli** man page.

6.1.12. Creating an iSCSI initiator

An iSCSI initiator forms a session to connect to the iSCSI target. For more information on iSCSI target, see Section 6.1.2, "Creating an iSCSI target" . By default, an iSCSI service is **lazily** started and the service starts after running the **iscsiadm** command. If root is not on an iSCSI device or there are no nodes marked with **node.startup = automatic** then the iSCSI service will not start until an **iscsiadm** command is executed that requires **iscsid** or the **iscsi** kernel modules to be started.

To force the **iscsid** daemon to run and iSCSI kernel modules to load:

```
# systemctl start iscsid.service
```

Prerequisites

- Installed and running **targetcli** on a server machine. For more information, see Section 6.1.1, "Installing targetcli".

- An iSCSI target associated with a Target Portal Groups (TPG) on a server machine. For more information, see Section 6.1.2, "Creating an iSCSI target" .
- Created iSCSI ACL. For more information, see Section 6.1.11, "Creating an iSCSI ACL" .

Procedure

1. Install **iscsi-initiator-utils** on client machine:

```
# yum install iscsi-initiator-utils
```

2. Check the initiator name:

```
# cat /etc/iscsi/initiatorname.iscsi

InitiatorName=2006-04.com.example.foo:888
```

3. If the ACL was given a custom name in Section 6.1.11, "Creating an iSCSI ACL" , modify the **/etc/iscsi/initiatorname.iscsi** file accordingly.

```
# vi /etc/iscsi/initiatorname.iscsi
```

4. Discover the target and log in to the target with the displayed target IQN:

```
# iscsiadm -m discovery -t st -p 10.64.24.179
   10.64.24.179:3260,1 iqn.2006-04.example:444

# iscsiadm -m node -T iqn.2006-04.example:444 -l
   Logging in to [iface: default, target: iqn.2006-04.example:444, portal: 10.64.24.179,3260]
(multiple)
   Login to [iface: default, target: iqn.2006-04.example:444, portal: 10.64.24.179,3260]
successful.
```

Replace *10.64.24.179* with the target-ip-address.

You can use this procedure for any number of initiators connected to the same target if their respective initiator names are added to the ACL as described in the Section 6.1.11, "Creating an iSCSI ACL".

5. Find the iSCSI disk name and create a file system on this iSCSI disk:

```
# grep "Attached SCSI" /var/log/messages

# mkfs.ext4 /dev/disk_name
```

Replace *disk_name* with the iSCSI disk name displayed in the **/var/log/messages** file.

6. Mount the file system:

```
# mkdir /mount/point

# mount /dev/disk_name /mount/point
```

Replace */mount/point* with the mount point of the partition.

7. Edit the **/etc/fstab** file to mount the file system automatically when the system boots:

```
# vi /etc/fstab

/dev/disk_name /mount/point ext4 _netdev 0 0
```

Replace *disk_name* with the iSCSI disk name and */mount/point* with the mount point of the partition.

Additional resources

- The **targetcli** man page.
- The **iscsiadm** man page.

6.1.13. Setting up the Challenge-Handshake Authentication Protocol for the target

The **Challenge-Handshake Authentication Protocol (CHAP)** allows the user to protect the target with a password. The initiator must be aware of this password to be able to connect to the target.

Prerequisites

- Created iSCSI ACL. For more information, see Section 6.1.11, “Creating an iSCSI ACL”.

Procedure

1. Set attribute authentication:

```
/iscsi/iqn.20...mple:444/tpg1> set attribute authentication=1

Parameter authentication is now '1'.
```

2. Set **userid** and **password**:

```
/tpg1> set auth userid=redhat
Parameter userid is now 'redhat'.

/iscsi/iqn.20...689dcbb3/tpg1> set auth password=redhat_passwd
Parameter password is now 'redhat_passwd'.
```

Additional resources

- The **targetcli** man page.

6.1.14. Setting up the Challenge-Handshake Authentication Protocol for the initiator

The **Challenge-Handshake Authentication Protocol (CHAP)** allows the user to protect the target with a password. The initiator must be aware of this password to be able to connect to the target.

Prerequisites

- Created iSCSI initiator. For more information, see Section 6.1.12, “Creating an iSCSI initiator”.

- Set the **CHAP** for the target. For more information, see Section 6.1.13, "Setting up the Challenge-Handshake Authentication Protocol for the target".

Procedure

1. Enable CHAP authentication in the **iscsid.conf** file:

```
# vi /etc/iscsi/iscsid.conf

node.session.auth.authmethod = CHAP
```

   By default, the **node.session.auth.authmethod** is set to **None**

2. Add target **username** and **password** in the **iscsid.conf** file:

```
node.session.auth.username = redhat
node.session.auth.password = redhat_passwd
```

3. Start the **iscsid** daemon:

```
# systemctl start iscsid.service
```

Additional resources

- The **iscsiadm** man page

6.2. MONITORING AN ISCSI SESSION

As a system administrator, you can monitor the iSCSI session using the **iscsiadm** utility.

6.2.1. Monitoring an iSCSI session using the iscsiadm utility

This procedure describes how to monitor the iscsi session using the **iscsiadm** utility.

By default, an iSCSI service is **lazily** started and the service starts after running the **iscsiadm** command. If root is not on an iSCSI device or there are no nodes marked with **node.startup = automatic** then the iSCSI service will not start until an **iscsiadm** command is executed that requires **iscsid** or the **iscsi** kernel modules to be started.

To force the **iscsid** daemon to run and iSCSI kernel modules to load:

```
# systemctl start iscsid.service
```

Prerequisites

- Installed iscsi-initiator-utils on client machine:

```
yum install iscsi-initiator-utils
```

Procedure

1. Find information about the running sessions:

```
# iscsiadm -m session -P 3
```

This command displays the session or device state, session ID (sid), some negotiated parameters, and the SCSI devices accessible through the session.

- For shorter output, for example, to display only the **sid-to-node** mapping, run:

```
# iscsiadm -m session -P 0
      or
# iscsiadm -m session

tcp [2] 10.15.84.19:3260,2 iqn.1992-08.com.netapp:sn.33615311
tcp [3] 10.15.85.19:3260,3 iqn.1992-08.com.netapp:sn.33615311
```

These commands print the list of running sessions in the following format: **driver [sid] target_ip:port,target_portal_group_tag proper_target_name**.

Additional resources

- **/usr/share/doc/iscsi-initiator-utils-version/README** file.
- The **iscsiadm** man page.

6.3. REMOVING AN ISCSI TARGET

As a system administrator, you can remove the iSCSI target.

6.3.1. Removing an iSCSI object using targetcli tool

This procedure describes how to remove the iSCSI objects using the **targetcli** tool.

Procedure

1. Log off from the target:

```
# iscsiadm -m node -T iqn.2006-04.example:444 -u
```

For more information on how to log in to the target, see Section 6.1.12, "Creating an iSCSI initiator".

2. Remove the entire target, including all ACLs, LUNs, and portals:

```
/> iscsi/ delete iqn.2006-04.com.example:444
```

Replace *iqn.2006-04.com.example:444* with the target_iqn_name.

- To remove an iSCSI backstore:

```
/> backstores/backstore-type/ delete block_backend
```

  - Replace *backstore-type* with either **fileio**, **block**, **pscsi**, or **ramdisk**.
  - Replace *block_backend* with the *backstore-name* you want to delete.

- To remove parts of an iSCSI target, such as an ACL:

```
/> /iscsi/iqn-name/tpg/acls/ delete iqn.2006-04.com.example:444
```

3. View the changes:

```
/> iscsi/ ls
```

Additional resources

- The **targetcli** man page.

6.4. DM MULTIPATH OVERRIDES OF THE DEVICE TIMEOUT

The **recovery_tmo sysfs** option controls the timeout for a particular iSCSI device. The following options globally override **recovery_tmo** values:

- The **replacement_timeout** configuration option globally overrides the **recovery_tmo** value for all iSCSI devices.
- For all iSCSI devices that are managed by DM Multipath, the **fast_io_fail_tmo** option in DM Multipath globally overrides the **recovery_tmo** value.
  The **fast_io_fail_tmo** option in DM Multipath also overrides the **fast_io_fail_tmo** option in Fibre Channel devices.

The DM Multipath **fast_io_fail_tmo** option takes precedence over **replacement_timeout**. Red Hat does not recommend using **replacement_timeout** to override **recovery_tmo** in devices managed by DM Multipath because DM Multipath always resets **recovery_tmo** when the **multipathd** service reloads.

CHAPTER 7. USING FIBRE CHANNEL DEVICES

RHEL 8 provides the following native Fibre Channel drivers:

- **lpfc**
- **qla2xxx**
- **zfcp**

7.1. RESIZING FIBRE CHANNEL LOGICAL UNITS

As a system administrator, you can resize Fibre Channel logical units.

Procedure

1. To determine which devices are paths for a **multipath** logical unit:

```
multipath -ll
```

2. To re-scan Fibre Channel logical units on a system that uses multipathing:

```
$ echo 1 > /sys/block/sdX/device/rescan
```

Additional resources

- The **multipath** man page.

7.2. DETERMINING THE LINK LOSS BEHAVIOR OF DEVICE USING FIBRE CHANNEL

If a driver implements the Transport **dev_loss_tmo** callback, access attempts to a device through a link will be blocked when a transport problem is detected.

Procedure

- Determine the state of a remote port:

```
$ cat /sys/class/fc_remote_port/rport-host:bus:remote-port/port_state
```

  This command returns any one of the following output:

  - **Blocked** when the remote port along with devices accessed through it are blocked.
  - **Online** if the remote port is operating normally

  If the problem is not resolved within **dev_loss_tmo** seconds, the **rport** and devices will be unblocked. All I/O running on that device along with any new I/O sent to that device will fail.

When a link loss exceeds **dev_loss_tmo**, the **scsi_device** and **sd*N*** devices are removed. Typically, the Fibre Channel class will leave the device as is; i.e. **/dev/*sdx*** will remain **/dev/*sdx***. This is because the target binding is saved by the Fibre Channel driver and when the target port returns, the SCSI addresses

are recreated faithfully. However, this cannot be guaranteed; the ***sdx*** device will be restored only if no additional change on in-storage box configuration of LUNs is made.

Additional resources

- The **multipath.conf** man page
- Using **multipath**, you can modify the link loss behavior of device. For more information, see the following Knowledgebase articles:
  - How to set dev_loss_tmo and fast_io_fail_tmo persistently, using a udev rule
  - Recommended tuning at scsi,multipath and at application layer while configuring Oracle RAC cluster

7.3. FIBRE CHANNEL CONFIGURATION FILES

Following is the list of configuration files in the **/sys/class/** directory that provide the user-space API to Fibre Channel.

The items use the following variables:

H

Host number

B

Bus number

T

Target

L

Logical unit (LUNs)

R

Remote port number

IMPORTANT

If your system is using multipath software, Red Hat recommends that you consult your hardware vendor before changing any of the values described in this section.

Transport configuration in /sys/class/fc_transport/target*H*:*B*:*T*/

port_id

24-bit port ID/address

node_name

64-bit node name

port_name

64-bit port name

Remote port configuration in /sys/class/fc_remote_ports/rport-*H*:*B*-*R*/

- **port_id**

- **node_name**
- **port_name**
- **dev_loss_tmo**
  Controls when the scsi device gets removed from the system. After **dev_loss_tmo** triggers, the scsi device is removed. In the **multipath.conf** file , you can set **dev_loss_tmo** to **infinity**.

  In Red Hat Enterprise Linux 8, if you do not set the **fast_io_fail_tmo** option, **dev_loss_tmo** is capped to **600** seconds. By default, **fast_io_fail_tmo** is set to **5** seconds in Red Hat Enterprise Linux 8 if the **multipathd** service is running; otherwise, it is set to **off**.
- **fast_io_fail_tmo**
  Specifies the number of seconds to wait before it marks a link as "bad". Once a link is marked bad, existing running I/O or any new I/O on its corresponding path fails.

  If I/O is in a blocked queue, it will not be failed until **dev_loss_tmo** expires and the queue is unblocked.

  If **fast_io_fail_tmo** is set to any value except off, **dev_loss_tmo** is uncapped. If **fast_io_fail_tmo** is set to off, no I/O fails until the device is removed from the system. If **fast_io_fail_tmo** is set to a number, I/O fails immediately when the **fast_io_fail_tmo** timeout triggers.

Host configuration in /sys/class/fc_host/host*H*/

- **port_id**
- **node_name**
- **port_name**
- **issue_lip**
  Instructs the driver to rediscover remote ports.

7.4. DM MULTIPATH OVERRIDES OF THE DEVICE TIMEOUT

The **recovery_tmo sysfs** option controls the timeout for a particular iSCSI device. The following options globally override **recovery_tmo** values:

- The **replacement_timeout** configuration option globally overrides the **recovery_tmo** value for all iSCSI devices.
- For all iSCSI devices that are managed by DM Multipath, the **fast_io_fail_tmo** option in DM Multipath globally overrides the **recovery_tmo** value.
  The **fast_io_fail_tmo** option in DM Multipath also overrides the **fast_io_fail_tmo** option in Fibre Channel devices.

The DM Multipath **fast_io_fail_tmo** option takes precedence over **replacement_timeout**. Red Hat does not recommend using **replacement_timeout** to override **recovery_tmo** in devices managed by DM Multipath because DM Multipath always resets **recovery_tmo** when the **multipathd** service reloads.

CHAPTER 8. CONFIGURING A FIBRE CHANNEL OVER AN ETHERNET INTERFACE

Red Hat Enterprise Linux 8 ships with the following native FCoE drivers:

- **bnx2fc**
- **fnic**
- **qedf**
- **lpfc**

8.1. CONFIGURING AN ETHERNET INTERFACE TO USE FCOE

As a system administrator, you can configure FCoE for **bnx2fc** driver.

Prerequisites

- Setting up and deploying a FCoE interface requires the **fcoe-utils** package:

```
# yum install fcoe-utils
```

Procedure

1. To configure a new virtual LAN (VLAN), create a copy of an existing network script:

```
# cp /etc/fcoe/cfg-ethx /etc/fcoe/cfg-ethX
```

   Replace ***/etc/fcoe/cfg-ethx*** with a network script and ***/etc/fcoe/cfg-ethX*** with an Ethernet device that supports FCoE.

   Modify the contents of the **cfg-*ethX*** file as required.

2. If you want the device to automatically load during boot time, set the following parameter in the *ifcfg-ethX* file.

```
# vi /etc/sysconfig/network-scripts/ifcfg-ethX

ONBOOT=yes
```

   For example, if the FCoE device is *eth2*, edit the **/etc/sysconfig/network-scripts/ifcfg-*eth2*** file accordingly.

3. To load the FCoE device:

```
# ip link set dev ethX up
```

4. To start the FCoE:

```
# systemctl start fcoe
```

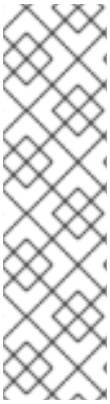
   The FCoE device displays if all other settings on the fabric are correct.

5. To view configured FCoE devices:

```
# fcoeadm -i
```

6. After correctly configuring the Ethernet interface to use FCoE, Red Hat recommends that you set FCoE service to run at startup.

```
# systemctl enable fcoe
```



NOTE

To stop the daemon:

```
# systemctl stop fcoe
```

Stopping the daemon does not reset the configuration of FCoE interfaces. To reset the configuration:

```
# systemctl -s SIGHUP kill fcoe
```

Additional resources

- The **fcoe** man page
- The **fcoeadm** man page
- FCoE software removal

8.2. CONFIGURING AN FCOE INTERFACE TO AUTOMATICALLY MOUNT AT BOOT

You can mount newly discovered disks via **udev** rules, **autofs**, and other similar methods. If a service requires the FCoE disk be mounted at boot-time, the FCoE disk should be mounted as soon as the fcoe service runs and before the initiation of any service that requires the FCoE disk. The FCoE mounting codes may differ depending on the system configuration, for example, a simple formatted FCoE disk, LVM, or a multipathed device node.

Procedure

1. To configure an FCoE disk to automatically mount at boot, add appropriate FCoE mounting code to the startup script for the **fcoe** service. The fcoe startup script is in the **/lib/systemd/system/fcoe.service** file.

   Example: FCoE Mounting Code

   The following is a sample FCoE mounting code for mounting file systems specified via wild cards in the **/etc/fstab** file:

```
mount_fcoe_disks_from_fstab()
{
  local timeout=20
 local done=1
 local fcoe_disks=($(egrep 'by-path\/fc-.*_netdev' /etc/fstab | cut -d ' ' -f1))
```

```
test -z $fcoe_disks && return 0

echo -n "Waiting for fcoe disks . "
 while [ $timeout -gt 0 ]; do
 for disk in ${fcoe_disks[*]}; do
   if ! test -b $disk; then
   done=0
  break
   fi
done

test $done -eq 1 && break;
sleep 1
echo -n ". "
done=1
let timeout--
done

if test $timeout -eq 0; then
 echo "timeout!"
else
 echo "done!"
fi

# mount any newly discovered disk
mount -a 2>/dev/null
}
```

2. To start the FCoE:

```
# systemctl start fcoe
```

The **mount_fcoe_disks_from_fstab** function should be invoked after the **fcoe** service script starts the **fcoemond** daemon. This will mount FCoE disks specified by the following paths in the **/etc/fstab** file:

```
/dev/disk/by-path/fc-0xXX:0xXX /mnt/fcoe-disk1 ext4  defaults,_netdev    0 0
```

```
/dev/disk/by-path/fc-0xYY:0xYY /mnt/fcoe-disk2 ext3  defaults,_netdev    0 0
```

Entries with **fc-** and **_netdev** sub-strings enable the **mount_fcoe_disks_from_fstab** function to identify FCoE disk mount entries.

NOTE

The **fcoe** service does not implement a timeout for FCoE disk discovery. The FCoE mounting code should implement its own timeout period.

Additional resources

- The **fcoe** man page
- The **fstab** man page.

- The **/usr/share/doc/fcoe-utils-version/README** file.

CHAPTER 9. CONFIGURING MAXIMUM TIME FOR STORAGE ERROR RECOVERY WITH EH_DEADLINE

You can configure the maximum allowed time to recover failed SCSI devices. This configuration guarantees an I/O response time even when storage hardware becomes unresponsive due to a failure.

9.1. THE EH_DEADLINE PARAMETER

The SCSI error handling (EH) mechanism attempts to perform error recovery on failed SCSI devices. The SCSI host object **eh_deadline** parameter enables you to configure the maximum amount of time for the recovery. After the configured time expires, SCSI EH stops and resets the entire host bus adapter (HBA).

Using **eh_deadline** can reduce the time:

- to shut off a failed path,
- to switch a path, or
- to disable a RAID slice.

WARNING

When **eh_deadline** expires, SCSI EH resets the HBA, which affects all target paths on that HBA, not only the failing one. If some of the redundant paths are not available for other reasons, I/O errors might occur. Enable **eh_deadline** only if you have a fully redundant multipath configuration on all targets.

Scenarios when eh_deadline is useful

In most scenarios, you do not need to enable **eh_deadline**. Using **eh_deadline** can be useful in certain specific scenarios, for example if a link loss occurs between a Fibre Channel (FC) switch and a target port, and the HBA does not receive Registered State Change Notifications (RSCNs). In such a case, I/O requests and error recovery commands all time out rather than encounter an error. Setting **eh_deadline** in this environment puts an upper limit on the recovery time. That enables the failed I/O to be retried on another available path by DM Multipath.

Under the following conditions, the **eh_deadline** functionality provides no additional benefit, because the I/O and error recovery commands fail immediately, which allows DM Multipath to retry:

- If RSCNs are enabled
- If the HBA does not register the link becoming unavailable

Possible values

The value of the **eh_deadline** is specified in seconds.

The default setting is **off**, which disables the time limit and allows all of the error recovery to take place.

9.2. SETTING THE EH_DEADLINE PARAMETER

This procedure configures the value of the **eh_deadline** parameter to limit the maximum SCSI recovery time.

Procedure

- You can configure **eh_deadline** using either of the following methods:

  sysfs

  Write the number of seconds into the **/sys/class/scsi_host/host*/eh_deadline** files.

  Kernel parameter

  Set a default value for all SCSI HBAs using the **scsi_mod.eh_deadline** kernel parameter.

Additional resources

- How to set eh_deadline and eh_timeout persistently, using a udev rule

CHAPTER 10. GETTING STARTED WITH SWAP

This section describes swap space and how to use it.

10.1. SWAP SPACE

Swap space in Linux is used when the amount of physical memory (RAM) is full. If the system needs more memory resources and the RAM is full, inactive pages in memory are moved to the swap space. While swap space can help machines with a small amount of RAM, it should not be considered a replacement for more RAM. Swap space is located on hard drives, which have a slower access time than physical memory. Swap space can be a dedicated swap partition (recommended), a swap file, or a combination of swap partitions and swap files.

In years past, the recommended amount of swap space increased linearly with the amount of RAM in the system. However, modern systems often include hundreds of gigabytes of RAM. As a consequence, recommended swap space is considered a function of system memory workload, not system memory.

Section 10.2, "Recommended system swap space" illustrates the recommended size of a swap partition depending on the amount of RAM in your system and whether you want sufficient memory for your system to hibernate. The recommended swap partition size is established automatically during installation. To allow for hibernation, however, you need to edit the swap space in the custom partitioning stage.

Recommendations in Section 10.2, "Recommended system swap space" are especially important on systems with low memory (1 GB and less). Failure to allocate sufficient swap space on these systems can cause issues such as instability or even render the installed system unbootable.

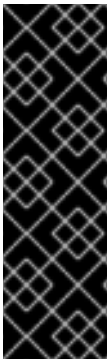
10.2. RECOMMENDED SYSTEM SWAP SPACE

This section gives recommendation about swap space.

Amount of RAM in the system	Recommended swap space	Recommended swap space if allowing for hibernation
⩽ 2 GB	2 times the amount of RAM	3 times the amount of RAM
> 2 GB – 8 GB	Equal to the amount of RAM	2 times the amount of RAM
> 8 GB – 64 GB	At least 4 GB	1.5 times the amount of RAM
> 64 GB	At least 4 GB	Hibernation not recommended

At the border between each range listed in the table above, for example a system with 2 GB, 8 GB, or 64 GB of system RAM, discretion can be exercised with regard to chosen swap space and hibernation support. If your system resources allow for it, increasing the swap space may lead to better performance. A swap space of at least 100 GB is recommended for systems with over 140 logical processors or over 3 TB of RAM.

Note that distributing swap space over multiple storage devices also improves swap space performance, particularly on systems with fast drives, controllers, and interfaces.



IMPORTANT

File systems and LVM2 volumes assigned as swap space *should not* be in use when being modified. Any attempts to modify swap fail if a system process or the kernel is using swap space. Use the **free** and **cat /proc/swaps** commands to verify how much and where swap is in use.

You should modify swap space while the system is booted in **rescue** mode, see Debug boot options in the *Performing an advanced RHEL installation*. When prompted to mount the file system, select **Skip**.

10.3. ADDING SWAP SPACE

This section describes how to add more swap space after installation. For example, you may upgrade the amount of RAM in your system from 1 GB to 2 GB, but there is only 2 GB of swap space. It might be advantageous to increase the amount of swap space to 4 GB if you perform memory-intense operations or run applications that require a large amount of memory.

There are three options: create a new swap partition, create a new swap file, or extend swap on an existing LVM2 logical volume. It is recommended that you extend an existing logical volume.

10.3.1. Extending swap on an LVM2 logical volume

This procedure describes how to extend swap space on an existing LVM2 logical volume. Assuming **/dev/VolGroup00/LogVol01** is the volume you want to extend by 2 GB.

Prerequisites

- Enough disk space.

Procedure

1. Disable swapping for the associated logical volume:

```
# swapoff -v /dev/VolGroup00/LogVol01
```

2. Resize the LVM2 logical volume by 2 GB:

```
# lvresize /dev/VolGroup00/LogVol01 -L +2G
```

3. Format the new swap space:

```
# mkswap /dev/VolGroup00/LogVol01
```

4. Enable the extended logical volume:

```
# swapon -v /dev/VolGroup00/LogVol01
```

5. To test if the swap logical volume was successfully extended and activated, inspect active swap space:

```
$ cat /proc/swaps
$ free -h
```

10.3.2. Creating an LVM2 logical volume for swap

This procedure describes how to create an LVM2 logical volume for swap. Assuming **/dev/VolGroup00/LogVol02** is the swap volume you want to add.

Prerequisites

- Enough disk space.

Procedure

1. Create the LVM2 logical volume of size 2 GB:

```
# lvcreate VolGroup00 -n LogVol02 -L 2G
```

2. Format the new swap space:

```
# mkswap /dev/VolGroup00/LogVol02
```

3. Add the following entry to the **/etc/fstab** file:

```
/dev/VolGroup00/LogVol02 swap swap defaults 0 0
```

4. Regenerate mount units so that your system registers the new configuration:

```
# systemctl daemon-reload
```

5. Activate swap on the logical volume:

```
# swapon -v /dev/VolGroup00/LogVol02
```

6. To test if the swap logical volume was successfully created and activated, inspect active swap space:

```
$ cat /proc/swaps
$ free -h
```

10.3.3. Creating a swap file

This procedure describes how to create a swap file.

Prerequisites

- Enough disk space.

Procedure

1. Determine the size of the new swap file in megabytes and multiply by 1024 to determine the number of blocks. For example, the block size of a 64 MB swap file is 65536.

2. Create an empty file:

```
# dd if=/dev/zero of=/swapfile bs=1024 count=65536
```

Replace *count* with the value equal to the desired block size.

3. Set up the swap file with the command:

```
# mkswap /swapfile
```

4. Change the security of the swap file so it is not world readable.

```
# chmod 0600 /swapfile
```

5. To enable the swap file at boot time, edit **/etc/fstab** as root to include the following entry:

```
/swapfile swap swap defaults 0 0
```

The next time the system boots, it activates the new swap file.

6. Regenerate mount units so that your system registers the new **/etc/fstab** configuration:

```
# systemctl daemon-reload
```

7. To activate the swap file immediately:

```
# swapon /swapfile
```

8. To test if the new swap file was successfully created and activated, inspect active swap space:

```
$ cat /proc/swaps
$ free -h
```

10.4. REMOVING SWAP SPACE

This section describes how to reduce swap space after installation. For example, you have downgraded the amount of RAM in your system from 1 GB to 512 MB, but there is 2 GB of swap space still assigned. It might be advantageous to reduce the amount of swap space to 1 GB, since the larger 2 GB could be wasting disk space.

Depending on what you need, you may choose one of three options: reduce swap space on an existing LVM2 logical volume, remove an entire LVM2 logical volume used for swap, or remove a swap file.

10.4.1. Reducing swap on an LVM2 logical volume

This procedure describes how to reduce swap on an LVM2 logical volume. Assuming **/dev/VolGroup00/LogVol01** is the volume you want to reduce.

Procedure

1. Disable swapping for the associated logical volume:

```
# swapoff -v /dev/VolGroup00/LogVol01
```

2. Reduce the LVM2 logical volume by 512 MB:

```
# lvreduce /dev/VolGroup00/LogVol01 -L -512M
```

3. Format the new swap space:

```
# mkswap /dev/VolGroup00/LogVol01
```

4. Activate swap on the logical volume:

```
# swapon -v /dev/VolGroup00/LogVol01
```

5. To test if the swap logical volume was successfully reduced, inspect active swap space:

```
$ cat /proc/swaps
$ free -h
```

10.4.2. Removing an LVM2 logical volume for swap

This procedure describes how to remove an LVM2 logical volume for swap. Assuming **/dev/VolGroup00/LogVol02** is the swap volume you want to remove.

Procedure

1. Disable swapping for the associated logical volume:

```
# swapoff -v /dev/VolGroup00/LogVol02
```

2. Remove the LVM2 logical volume:

```
# lvremove /dev/VolGroup00/LogVol02
```

3. Remove the following associated entry from the **/etc/fstab** file:

```
/dev/VolGroup00/LogVol02 swap swap defaults 0 0
```

4. Regenerate mount units so that your system registers the new configuration:

```
# systemctl daemon-reload
```

5. To test if the logical volume was successfully removed, inspect active swap space:

```
$ cat /proc/swaps
$ free -h
```

10.4.3. Removing a swap file

This procedure describes how to remove a swap file.

Procedure

1. At a shell prompt, execute the following command to disable the swap file (where **/swapfile** is the swap file):

```
# swapoff -v /swapfile
```

2. Remove its entry from the **/etc/fstab** file accordingly.

3. Regenerate mount units so that your system registers the new configuration:

```
# systemctl daemon-reload
```

4. Remove the actual file:

```
# rm /swapfile
```

CHAPTER 11. MANAGING SYSTEM UPGRADES AS SNAPSHOTS

As a system administrator, use **Boom** to create boot entries for alternative copies of the system state. **Boom** simplifies the management of system updates.

WARNING

The procedures mentioned in this chapter does not work on multiple file systems in your system tree.

11.1. OVERVIEW OF THE BOOM PROCESS

Boom allows to create boot entries, which can then be accessed and selected from the GRUB 2 boot loader menu. By creating boot entries, the process of preparing for a rollback capable upgrade is now simplified.

Rollback-capable upgrades are done using the following process without editing any configuration files:

1. Create a snapshot or copy of the root file system.
2. Use **Boom** to create a boot entry for the current (older) environment.
3. Upgrade your Red Hat Enterprise Linux system.
4. Reboot the system, and select the version that you want to use.

Using **Boom** reduces the risks associated with upgrading a system and also helps to reduce hardware downtime.

For example, you can upgrade a Red Hat Enterprise Linux 7 system to Red Hat Enterprise Linux 8, while retaining the original Red Hat Enterprise Linux 7 environment. After the upgrade is complete, you can switch between the old Red Hat Enterprise Linux 7 and new Red Hat Enterprise Linux 8 environments, as needed.

This ability to switch between environments allows you to:

- Quickly compare both environments in a side-by-side fashion and switch between them with minimal overhead.
- Recover the older file system's content.
- Continue accessing the old system while the upgraded host is running.
- Halt and revert the update process at any time, even while the update itself is running.

Additional resources

- The **boom** man page.

11.2. UPGRADING TO ANOTHER VERSION USING BOOM

In addition to **Boom**, the following Red Hat Enterprise Linux components are used in this upgrade process:

- LVM
- GRUB 2 boot loader
- Leapp upgrade tool

Prerequisites

- Install the **boom** package:

```
# yum install lvm2-python-boom
```

- Sufficient space allocated to the snapshot. Use the following commands to find the free space on the volume groups and logical volumes:

```
# vgs
VG  #PV  #LV  #SN  Attr  VSize    VFree
rhel 4 2 0 wz--n- 103.89g 29.99g

# lvs
LV     VG    Attr     LSize  Pool Origin Data% Meta% Move Log Cpy%Sync Convert
root rhel -wi-ao--- 68.88g
swap rhel -wi-ao--- 5.98g
```

Procedure

1. Create a snapshot of your *root* logical volume:
   - If your root file system uses thin provisioning, create a thin snapshot:
     While creating a thin snapshot, do not define the snapshot size. Snapshot is also allocated from the thin pool.

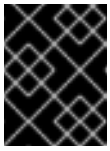
     ```
     # lvcreate -s -k n rhel/root -n root_snapshot_before_changes
     ```

     Here:
     - **-s** is used to create the snapshot
     - **rhel/root** is the file system being copied in the logical volume
     - **-n** *root_snapshot_before_changes* is the name of the snapshot
     - **-k n** is used to not skip the activation because thin snapshots are not activated by default
   - If your root file system uses thick provisioning, create a thick snapshot:
     While creating a thick snapshot, define the snapshot size that is able to hold all the changes during the upgrade.

     ```
     # lvcreate -s  rhel/root -n root_snapshot_before_changes -L 25g
     ```

Here:

- **-s** is used to create the snapshot
- **rhel/root** is the file system being copied
- **-n** *root_snapshot_before_changes* is the name of the snapshot
- **-L** *25g* is the snapshot size that is able to hold all the changes during the upgrade



IMPORTANT

After creating the snapshot, any additional system changes are not included.

2. Create the profile:

```
# boom profile create --from-host --uname-pattern el8
```

3. Create a new boot entry:

```
# boom create --title "Root LV snapshot before changes" --rootlv
rhel/root_snapshot_before_changes
```

Here:

- **--title** *Root LV snapshot before changes* is the name of the boot entry, which displays in the list during system startup
- **--rootlv** is the root logical volume that corresponds to the new boot entry
  If you execute the **boom create** command for the first time, the following message displays:

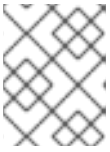
```
WARNING - Boom configuration not found in grub.cfg

WARNING - Run 'grub2-mkconfig > /boot/grub2/grub.cfg' to enable
```

To enable Boom in GRUB 2:

```
# grub2-mkconfig > /boot/grub2/grub.cfg
```

4. Upgrade using Leapp:

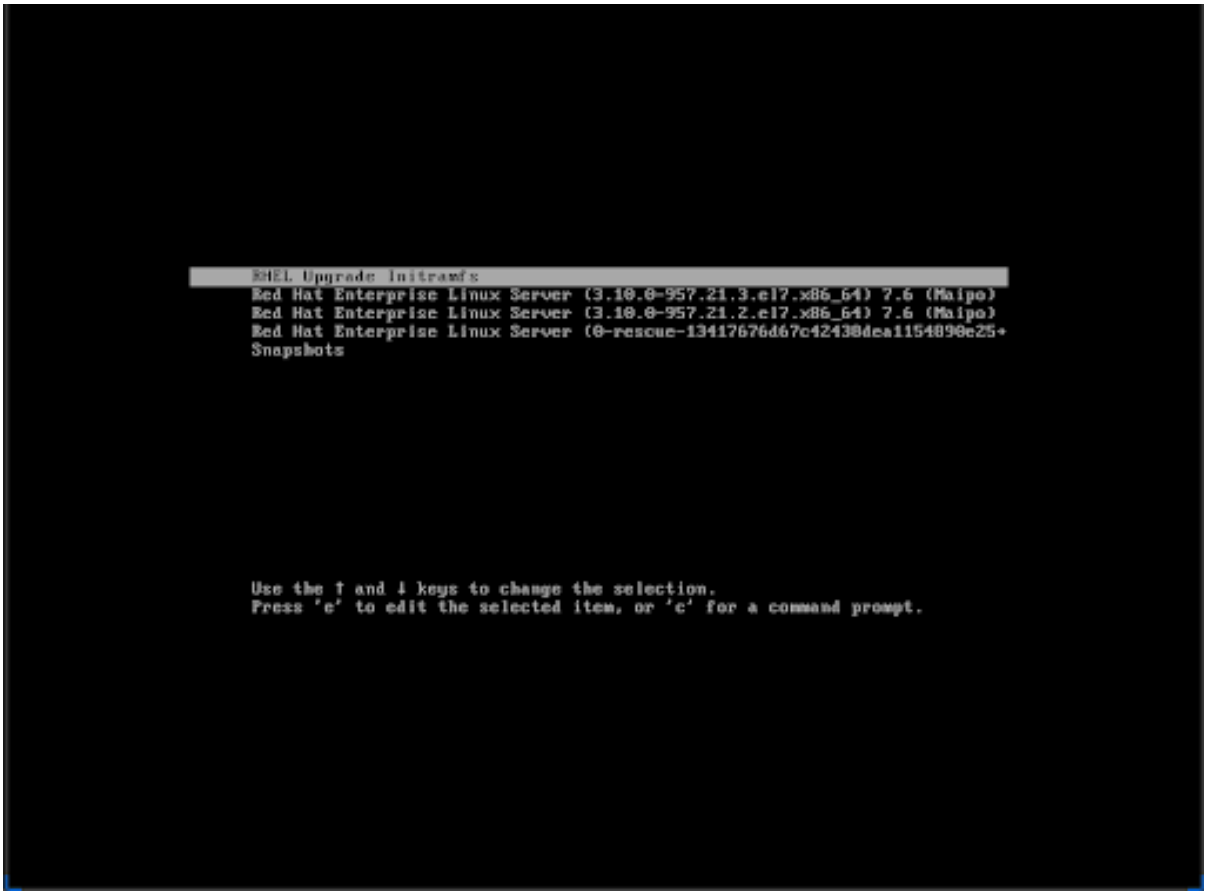
```
# leapp upgrade --reboot
```



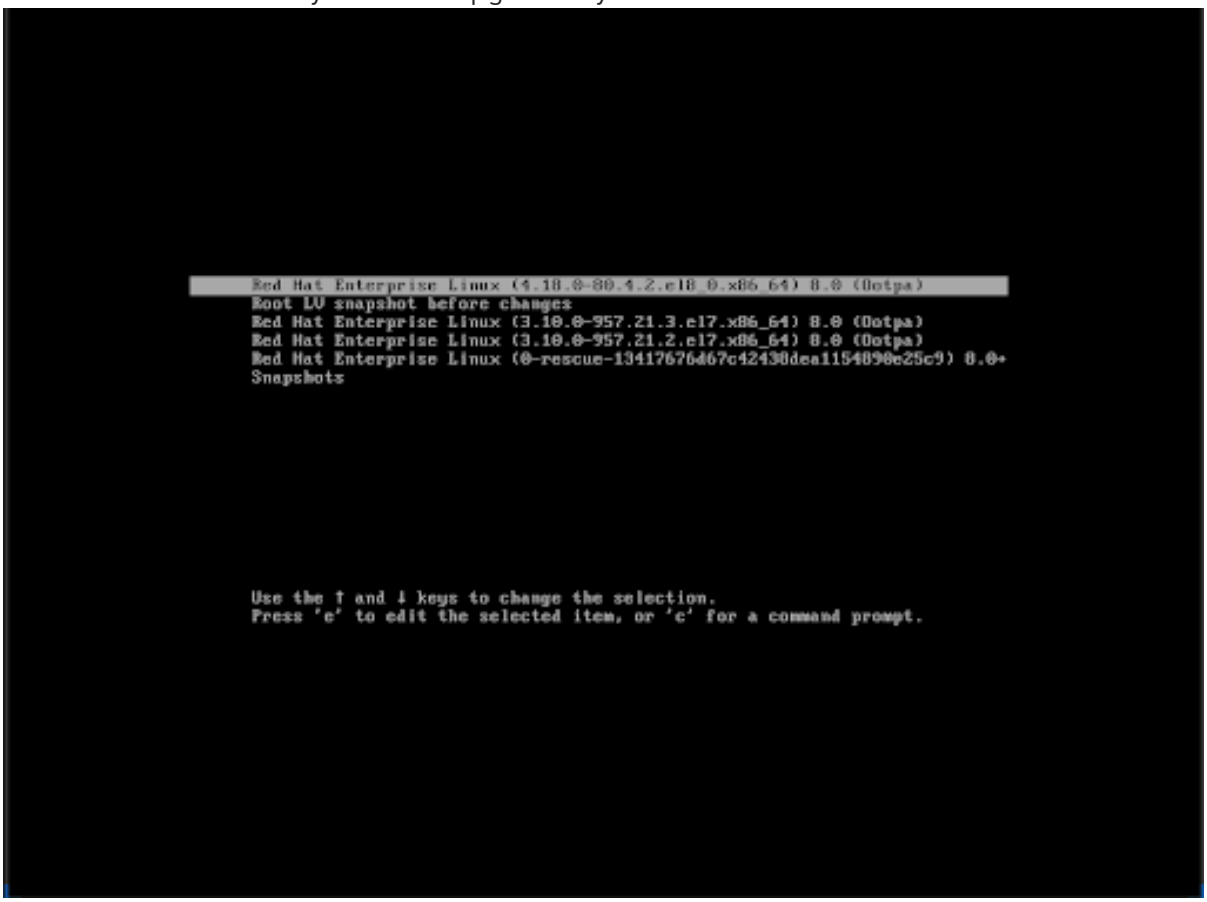
NOTE

The reboot argument initiates an automatic system restart after the upgrade process.

During reboot, the GRUB 2 screen is displayed:



5. Select the **RHEL Upgrade Initramfs** entry and press ENTER. The upgrade continues and new Red Hat Enterprise Linux 8 RPM packages are installed. After the upgrade is complete, the system automatically reboots and the GRUB 2 screen displays the upgraded and the older version of the available system. The upgraded system version is the default selection.



Additional resources

- The **boom** man page.
- What is BOOM and how to install it? Knowledgebase article.
- How to create a BOOM boot entry Knowledgebase article.

11.3. SWITCHING BETWEEN NEW AND OLD RED HAT ENTERPRISE LINUX VERSIONS

This procedure describes steps to switch between the new and the old Red Hat Enterprise Linux versions after the upgrade is complete.

Prerequisites

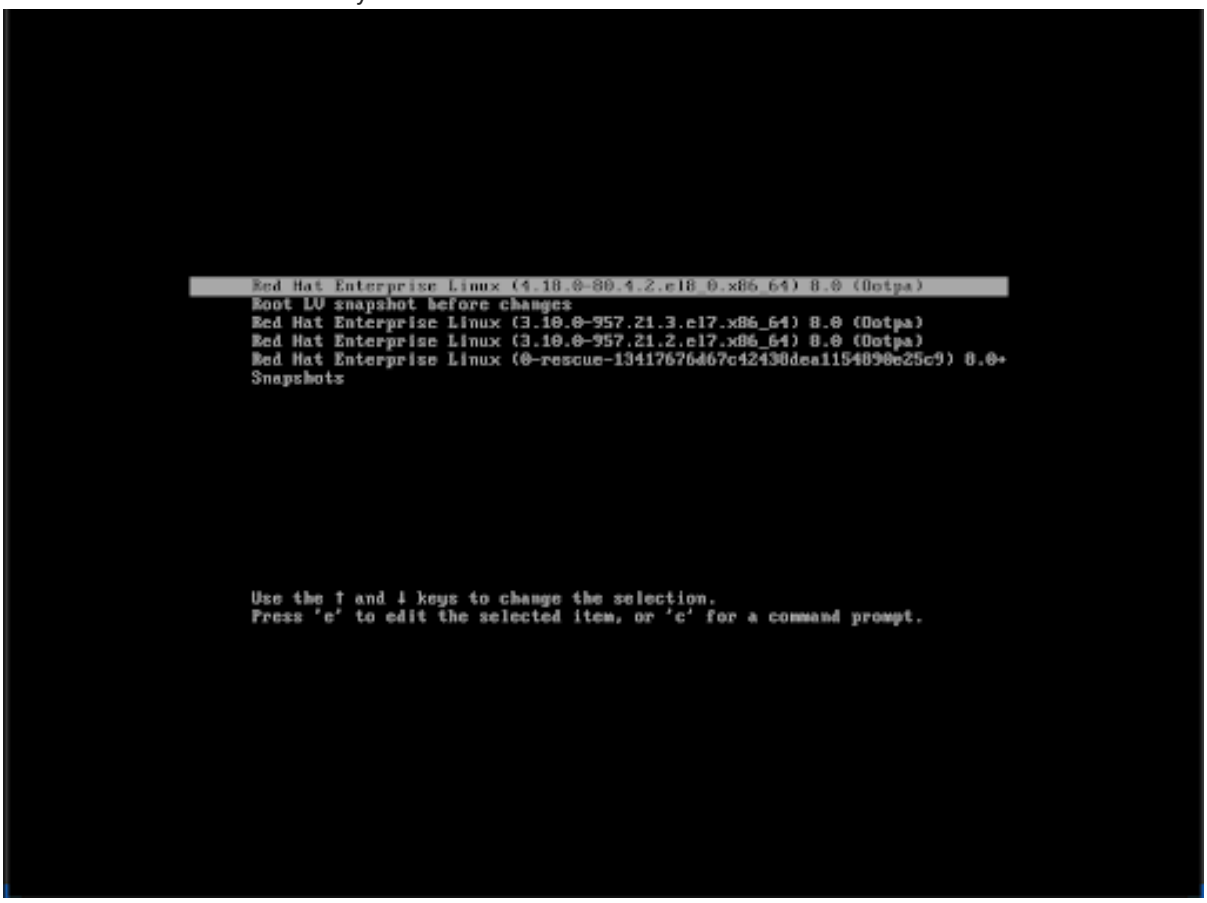
- Upgraded Red Hat Enterprise Linux version. For more information, see Section 11.2, "Upgrading to another version using Boom".

Procedure

1. Reboot the system:

```
# reboot
```

2. Select the desired boot entry from the GRUB 2 boot loader screen.



3. Verify that the selected boot volume is displayed:

```
# boom list
```

-

Additional resources

- The **boom** man page.

11.4. DELETING THE SNAPSHOT

This procedure describes steps to delete the snapshot.

Prerequisites

- Upgraded to new version. For more information, see Section 11.2, "Upgrading to another version using Boom"

Procedure

1. Boot into Red Hat Enterprise Linux 8 from the GRUB 2 entry. The following output confirms that the new snapshot is selected:

```
# boom list
BootID  Version                 Name                        RootDevice
6d2ec72 3.10.0-957.21.3.el7.x86_64 Red Hat Enterprise Linux Server
/dev/rhel/root_snapshot_before_changes
```

2. Delete the **Boom** snapshot entry using the **BootID** value:

```
# boom delete --boot-id 6d2ec72
```

   This deletes the entry from the GRUB 2 menu.

3. Remove the LV snapshot:

```
# lvremove rhel/root_snapshot_before_changes
Do you really want to remove active logical volume rhel/root_snapshot_before_changes?
[y/n]: y
    Logical volume "root_snapshot_before_changes" successfully removed
```

   ### Additional resources

   - The **boom** man page.

CHAPTER 12. OVERVIEW OF NVME OVER FABRIC DEVICES

Non-volatile Memory Express (NVMe) is an interface that allows host software utility to communicate with solid state drives.

Use the following types of fabric transport to configure **NVMe** over fabric devices:

- **NVMe** over fabrics using **Remote Direct Memory Access (RDMA)**. For information on how to configure NVMe/RDMA, see Section 12.1, "NVMe over fabrics using RDMA" .
- **NVMe** over fabrics using **Fibre Channel (FC)**. For information on how to configure **FC-NVMe**, see Section 12.2, "NVMe over fabrics using FC" .

When using FC and RDMA, the solid-state drive does not have to be local to your system; it can be configured remotely through a FC or RDMA controller.

12.1. NVME OVER FABRICS USING RDMA

In **NVMe**/**RDMA** setup, **NVMe** target and **NVMe** initiator is configured.

As a system administrator, complete the tasks in the following sections to deploy the **NVMe** over fabrics using **RDMA** (**NVMe**/**RDMA**):

- Section 12.1.1, "Setting up an NVMe/RDMA target using configfs"
- Section 12.1.2, "Setting up the NVMe/RDMA target using nvmetcli"
- Section 12.1.3, "Configuring an NVMe/RDMA client"

12.1.1. Setting up an NVMe/RDMA target using configfs

Use this procedure to configure an **NVMe**/**RDMA** target using **configfs**.

Prerequisites

- Verify that you have a block device to assign to the **nvmet** subsystem.

Procedure

1. Create the **nvmet-rdma** subsystem:

```
# modprobe nvmet-rdma

# mkdir /sys/kernel/config/nvmet/subsystems/testnqn

# cd /sys/kernel/config/nvmet/subsystems/testnqn
```

   Replace *testnqn* with the subsystem name.

2. Allow any host to connect to this target:

```
# echo 1 > attr_allow_any_host
```

3. Configure a **namespace**:

```
# mkdir namespaces/10

# cd namespaces/10
```

Replace *10* with the namespace number

4. Set a path to the **NVMe** device:

```
#echo -n /dev/nvme0n1 > device_path
```

5. Enable the namespace:

```
# echo 1 > enable
```

6. Create a directory with an **NVMe** port:

```
# mkdir /sys/kernel/config/nvmet/ports/1

# cd /sys/kernel/config/nvmet/ports/1
```

7. Display the IP address of *mlx5_ib0*:

```
# ip addr show mlx5_ib0

8: mlx5_ib0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 4092 qdisc mq state UP
group default qlen 256
    link/infiniband 00:00:06:2f:fe:80:00:00:00:00:00:00:e4:1d:2d:03:00:e7:0f:f6 brd
00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:ff:ff:ff:ff
    inet 172.31.0.202/24 brd 172.31.0.255 scope global noprefixroute mlx5_ib0
       valid_lft forever preferred_lft forever
    inet6 fe80::e61d:2d03:e7:ff6/64 scope link noprefixroute
       valid_lft forever preferred_lft forever
```

8. Set the transport address for the target:

```
# echo -n 172.31.0.202 > addr_traddr
```

9. Set **RDMA** as the transport type:

```
# echo rdma > addr_trtype

# echo 4420 > addr_trsvcid
```

10. Set the address family for the port:

```
# echo ipv4 > addr_adrfam
```

11. Create a soft link:

```
# ln -s /sys/kernel/config/nvmet/subsystems/testnqn
/sys/kernel/config/nvmet/ports/1/subsystems/testnqn
```

Verification steps

- Verify that the **NVMe** target is listening on the given port and ready for connection requests:

```
# dmesg | grep "enabling port"
[ 1091.413648] nvmet_rdma: enabling port 1 (172.31.0.202:4420)
```

Additional resources

- The **nvme** man page.

12.1.2. Setting up the NVMe/RDMA target using nvmetcli

Use the **nvmetcli** to edit, view, and start **NVMe** target. The **nvmetcli** provides a command line and an interactive shell option. Use this procedure to configure the **NVMe/RDMA** target by **nvmetcli**.

Prerequisites

- Verify that you have a block device to assign to the **nvmet** subsystem.
- Execute the **nvmetcli** operations as a root user.

Procedure

1. Install the **nvmetcli** package:

```
# yum install nvmetcli
```

2. Download the **rdma.json** file

```
# wget
http://git.infradead.org/users/hch/nvmetcli.git/blob_plain/0a6b088db2dc2e5de11e6f23f1e890e4
b54fee64:/rdma.json
```

3. Edit the **rdma.json** file and change the **traddr** value to *172.31.0.202*.
4. Setup the target by loading the NVMe target configuration file:

```
# nvmetcli restore rdma.json
```

NOTE

If the NVMe target configuration file name is not specified, the **nvmetcli** uses the **/etc/nvmet/config.json** file.

Verification steps

- Verify that the **NVMe** target is listening on the given port and ready for connection requests:

```
#dmesg | tail -1
[ 4797.132647] nvmet_rdma: enabling port 2 (172.31.0.202:4420)
```

- (Optional) Clear the current NVMe target:

```
# nvmetcli clear
```

Additional resources

- The **nvmetcli** man page.
- The **nvme** man page.

12.1.3. Configuring an NVMe/RDMA client

Use this procedure to configure an **NVMe**/**RDMA** client using the NVMe management command line interface (**nvme-cli**) tool.

Procedure

1. Install the **nvme-cli** tool:

```
# yum install nvme-cli
```

2. Load the **nvme-rdma** module if its not loaded:

```
# modprobe nvme-rdma
```

3. Discover available subsystems on the **NVMe** target:

```
# nvme discover -t rdma -a 172.31.0.202 -s 4420

Discovery Log Number of Records 1, Generation counter 2
=====Discovery Log Entry 0======
trtype:  rdma
adrfam:  ipv4
subtype: nvme subsystem
treq:    not specified, sq flow control disable supported
portid:  1
trsvcid: 4420
subnqn:  testnqn
traddr:  172.31.0.202
rdma_prtype: not specified
rdma_qptype: connected
rdma_cms:    rdma-cm
rdma_pkey: 0x0000
```

4. Connect to the discovered subsystems:

```
# nvme connect -t rdma -n testnqn -a 172.31.0.202 -s 4420

# lsblk
NAME                      MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
sda                         8:0    0 465.8G  0 disk
├─sda1                      8:1    0     1G  0 part /boot
└─sda2                      8:2    0 464.8G  0 part
  ├─rhel_rdma--virt--03-root 253:0    0    50G  0 lvm  /
  ├─rhel_rdma--virt--03-swap 253:1    0     4G  0 lvm  [SWAP]
```

```
  └─rhel_rdma--virt--03-home 253:2    0 410.8G  0 lvm  /home
nvme0n1

#cat /sys/class/nvme/nvme0/transport
rdma
```

Replace *testnqn* with the **NVMe** subsystem name.

Replace *172.31.0.202* with the target IP address.

Replace *4420* with the port number.

Verification steps

- List the NVMe devices that are currently connected:

```
# nvme list
```

- (Optional) Disconnect from the target:

```
# nvme disconnect -n testnqn
NQN:testnqn disconnected 1 controller(s)

# lsblk
NAME                      MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
sda                         8:0    0 465.8G  0 disk
├─sda1                      8:1    0     1G  0 part /boot
└─sda2                      8:2    0 464.8G  0 part
  ├─rhel_rdma--virt--03-root 253:0    0    50G  0 lvm  /
  ├─rhel_rdma--virt--03-swap 253:1    0     4G  0 lvm  [SWAP]
  └─rhel_rdma--virt--03-home 253:2    0 410.8G  0 lvm  /home
```

Additional resources

- The **nvme** man page.
- **Nvme-cli** Github repository

12.2. NVME OVER FABRICS USING FC

The **NVMe** over Fibre Channel (**FC-NVMe**) is fully supported in initiator mode when used with certain Broadcom Emulex and Marvell Qlogic Fibre Channel adapters. As a system administrator, complete the tasks in the following sections to deploy the **FC-NVMe**:

- Section 12.2.1, “Configuring the NVMe initiator for Broadcom adapters” .
- Section 12.2.2, “Configuring the NVMe initiator for QLogic adapters” .

12.2.1. Configuring the NVMe initiator for Broadcom adapters

Use this procedure to configure the **NVMe** initiator for **Broadcom** adapters client using the NVMe management command line interface (**nvme-cli**) tool.

Procedure

1. Install the **nvme-cli** tool:

```
# yum install nvme-cli
```

This creates the **hostnqn** file in the **/etc/nvme/** directory. The **hostnqn** file identifies the **NVMe** host.

To generate a new **hostnqn**:

```
#nvme gen-hostnqn
```

2. Find the **WWNN** and **WWPN** of the local and remote ports and use the output to find the subsystem **NQN**:

```
# cat /sys/class/scsi_host/host*/nvme_info

NVME Initiator Enabled
XRI Dist lpfc0 Total 6144 IO 5894 ELS 250
NVME LPORT lpfc0 WWPN x10000090fae0b5f5 WWNN x20000090fae0b5f5 DID x010f00
ONLINE
NVME RPORT       WWPN x204700a098cbcac6 WWNN x204600a098cbcac6 DID x01050e
TARGET DISCSRVC ONLINE

NVME Statistics
LS: Xmt 000000000e Cmpl 000000000e Abort 00000000
LS XMIT: Err 00000000  CMPL: xb 00000000 Err 00000000
Total FCP Cmpl 00000000000008ea Issue 00000000000008ec OutIO 0000000000000002
    abort 00000000 noxri 00000000 nondlp 00000000 qdepth 00000000 wqerr 00000000 err
00000000
FCP CMPL: xb 00000000 Err 00000000
```

```
# nvme discover --transport fc --traddr nn-0x204600a098cbcac6:pn-0x204700a098cbcac6 --
host-traddr nn-0x20000090fae0b5f5:pn-0x10000090fae0b5f5

Discovery Log Number of Records 2, Generation counter 49530
=====Discovery Log Entry 0======
trtype:  fc
adrfam:  fibre-channel
subtype: nvme subsystem
treq:    not specified
portid:  0
trsvcid: none
subnqn:  nqn.1992-
08.com.netapp:sn.e18bfca87d5e11e98c0800a098cbcac6:subsystem.st14_nvme_ss_1_1
traddr:  nn-0x204600a098cbcac6:pn-0x204700a098cbcac6
```

Replace *nn-0x204600a098cbcac6:pn-0x204700a098cbcac6* with the **traddr**.

Replace *nn-0x20000090fae0b5f5:pn-0x10000090fae0b5f5* with the **host_traddr**.

3. Connect to the NVMe target using the **nvme-cli**:

```
# nvme connect --transport fc --traddr nn-0x204600a098cbcac6:pn-0x204700a098cbcac6 --
host-traddr nn-0x20000090fae0b5f5:pn-0x10000090fae0b5f5 -n nqn.1992-
08.com.netapp:sn.e18bfca87d5e11e98c0800a098cbcac6:subsystem.st14_nvme_ss_1_1
```

Replace *nn-0x204600a098cbcac6:pn-0x204700a098cbcac6* with the **traddr**.

Replace *nn-0x20000090fae0b5f5:pn-0x10000090fae0b5f5* with the **host_traddr**.

Replace *nqn.1992-08.com.netapp:sn.e18bfca87d5e11e98c0800a098cbcac6:subsystem.st14_nvme_ss_1_1* with the ***subnqn***.

Verification steps

- List the NVMe devices that are currently connected:

```
# nvme list
Node             SN                   Model                                    Namespace Usage
Format           FW Rev
---------------- -------------------- ---------------------------------------- --------- --------------------------
- ---------------- --------
/dev/nvme0n1     80BgLFM7xMJbAAAAAAAC NetApp ONTAP Controller                   1
107.37  GB / 107.37  GB      4 KiB +  0 B    FFFFFFFF
```

```
# lsblk |grep nvme
nvme0n1                 259:0    0   100G  0 disk
```

Additional resources

- The **nvme** man page.
- **Nvme-cli** Github repository

12.2.2. Configuring the NVMe initiator for QLogic adapters

Use this procedure to configure **NVMe** initiator for **Qlogic** adapters client using the NVMe management command line interface (**nvme-cli**) tool.

Procedure

1. Install the **nvme-cli** tool:

```
# yum install nvme-cli
```

This creates the **hostnqn** file in the **/etc/nvme/** directory. The **hostnqn** file identifies the **NVMe** host.

To generate a new **hostnqn**:

```
#nvme gen-hostnqn
```

2. Remove and reload the **qla2xxx** module:

```
# rmmod qla2xxx
# modprobe qla2xxx
```

3. Find the **WWNN** and **WWPN** of the local and remote ports:

```
# dmesg |grep traddr

[    6.139862] qla2xxx [0000:04:00.0]-ffff:0: register_localport: host-traddr=nn-
0x20000024ff19bb62:pn-0x21000024ff19bb62 on portID:10700
[    6.241762] qla2xxx [0000:04:00.0]-2102:0: qla_nvme_register_remote: traddr=nn-
0x203b00a098cbcac6:pn-0x203d00a098cbcac6 PortID:01050d
```

Using this **host-traddr** and **traddr**, find the subsystem **NQN**:

```
nvme discover --transport fc --traddr nn-0x203b00a098cbcac6:pn-0x203d00a098cbcac6 --
host-traddr nn-0x20000024ff19bb62:pn-0x21000024ff19bb62

Discovery Log Number of Records 2, Generation counter 49530
=====Discovery Log Entry 0======
trtype:  fc
adrfam:  fibre-channel
subtype: nvme subsystem
treq:    not specified
portid:  0
trsvcid: none
subnqn:  nqn.1992-
08.com.netapp:sn.c9ecc9187b1111e98c0800a098cbcac6:subsystem.vs_nvme_multipath_1_su
bsystem_468
traddr:  nn-0x203b00a098cbcac6:pn-0x203d00a098cbcac6
```

Replace *nn-0x203b00a098cbcac6:pn-0x203d00a098cbcac6* with the **traddr**.

Replace *nn-0x20000024ff19bb62:pn-0x21000024ff19bb62* with the **host_traddr**.

4. Connect to the NVMe target using the **nvme-cli** tool:

```
# nvme connect  --transport fc --traddr nn-0x203b00a098cbcac6:pn-0x203d00a098cbcac6 --
host_traddr nn-0x20000024ff19bb62:pn-0x21000024ff19bb62 -n nqn.1992-
08.com.netapp:sn.c9ecc9187b1111e98c0800a098cbcac6:subsystem.vs_nvme_multipath_1_su
bsystem_468
```

Replace *nn-0x203b00a098cbcac6:pn-0x203d00a098cbcac6* with the **traddr**.

Replace *nn-0x20000024ff19bb62:pn-0x21000024ff19bb62* with the **host_traddr**.

Replace *nqn.1992-08.com.netapp:sn.c9ecc9187b1111e98c0800a098cbcac6:subsystem.vs_nvme_multipath_1_subsyste* with the ***subnqn***.

Verification steps

- List the NVMe devices that are currently connected:

```
# nvme list
Node          SN                Model                                  Namespace Usage
Format         FW Rev
---------------- -------------------- ---------------------------------------- --------- -------------------------
- ---------------- --------
```

```
/dev/nvme0n1     80BgLFM7xMJbAAAAAAAC NetApp ONTAP Controller                1
107.37  GB / 107.37  GB      4 KiB +  0 B   FFFFFFFF

# lsblk |grep nvme
nvme0n1                 259:0    0   100G  0 disk
```

Additional resources

- The **nvme** man page.
- **Nvme-cli** Github repository

CHAPTER 13. SETTING THE DISK SCHEDULER

The disk scheduler is responsible for ordering the I/O requests submitted to a storage device.

You can configure the scheduler in several different ways:

- Set the scheduler using **Tuned**, as described in Section 13.6, "Setting the disk scheduler using Tuned"
- Set the scheduler using **udev**, as described in Section 13.7, "Setting the disk scheduler using udev rules"
- Temporarily change the scheduler on a running system, as described in Section 13.8, "Temporarily setting a scheduler for a specific disk"

13.1. DISK SCHEDULER CHANGES IN RHEL 8

In RHEL 8, block devices support only multi-queue scheduling. This enables the block layer performance to scale well with fast solid-state drives (SSDs) and multi-core systems.

The traditional, single-queue schedulers, which were available in RHEL 7 and earlier versions, have been removed.

13.2. AVAILABLE DISK SCHEDULERS

The following multi-queue disk schedulers are supported in RHEL 8:

Disk schedulers

none

Implements a first-in first-out (FIFO) scheduling algorithm. It merges requests at the generic block layer through a simple last-hit cache.

mq-deadline

Attempts to provide a guaranteed latency for requests from the point at which requests reach the scheduler.
The **mq-deadline** scheduler sorts queued I/O requests into a read or write batch and then schedules them for execution in increasing logical block addressing (LBA) order. By default, read batches take precedence over write batches, because applications are more likely to block on read I/O operations. After **mq-deadline** processes a batch, it checks how long write operations have been starved of processor time and schedules the next read or write batch as appropriate.

This scheduler is suitable for most use cases, but particularly those in which read operations occur more often than write operations.

bfq

Targets desktop systems and interactive tasks.
The **bfq** scheduler ensures that a single application is never using all of the bandwidth. In effect, the storage device is always as responsive as if it was idle. The system does not become unresponsive when copying large files. In its default configuration, **bfq** focuses on delivering the lowest latency rather than achieving the maximum throughput.

bfq is based on **cfq** code. It does not grant the disk to each process for a fixed time slice but assigns a *budget* measured in number of sectors to the process.

kyber

Is intended for fast devices. The scheduler tunes itself to achieve a latency goal. You can configure the target latencies for read and synchronous write requests.

13.3. RECOMMENDED DISK SCHEDULERS FOR DIFFERENT USE CASES

Depending on the task that your system performs, Red Hat recommends the following disk schedulers:

Table 13.1. Recommendations

Use case	Disk scheduler recommendation
Traditional HDD with a SCSI interface	Use **mq-deadline** or **bfq**.
High-performance SSD or a CPU-bound system with fast storage	Use **none**, especially when running enterprise applications. Alternatively, use **kyber**.
Desktop or interactive tasks	Use **bfq**.
Virtual guest	Use **mq-deadline**. With a multi-queue host bus adapter (HBA), use **none**.

13.4. THE DEFAULT DISK SCHEDULER

Block devices use the default disk scheduler unless you specify another scheduler.

The kernel selects a default disk scheduler based on the type of device. The automatically selected scheduler is typically the optimal setting. If you require a different scheduler, Red Hat recommends to use **udev** rules or the **Tuned** application to configure it. Match the selected devices and switch the scheduler only for those devices.

13.5. DETERMINING THE ACTIVE DISK SCHEDULER

This procedure determines which disk scheduler is currently active on a given block device.

Procedure

- Read the content of the **/sys/block/*device*/queue/scheduler** file:

```
# cat /sys/block/device/queue/scheduler

[mq-deadline] kyber bfq none
```

In the file name, replace *device* with the block device name, for example **sdc**.

The active scheduler is listed in square brackets (**[ ]**).

13.6. SETTING THE DISK SCHEDULER USING TUNED

This procedure creates and enables a **Tuned** profile that sets a given disk scheduler for selected block devices. The setting persists across system reboots.

In the following commands and configuration, replace:

- *device* with the name of the block device, for example **sdf**
- *selected-scheduler* with the disk scheduler that you want to set for the device, for example **bfq**

Prerequisites

- The **tuned** service is installed and enabled.
  For details, see https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/8/html/monitoring_and_managing_system_status_and_performance/started-with-tuned_monitoring-and-managing-system-status-and-performance#installing-and-enabling-tuned_getting-started-with-tuned.

Procedure

1. Optional: Select an existing **Tuned** profile on which your profile will be based. For a list of available profiles, see https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/8/html/monitoring_and_managing_system_status_and_performance/started-with-tuned_monitoring-and-managing-system-status-and-performance#tuned-profiles-distributed-with-rhel_getting-started-with-tuned.
   To see which profile is currently active, use:

   ```
   $ tuned-adm active
   ```

2. Create a new directory to hold your **Tuned** profile:

   ```
   # mkdir /etc/tuned/my-profile
   ```

3. Find the World Wide Name (WWN) identifier of the selected block device:

   ```
   $ udevadm info --query=property --name=/dev/device | grep WWN=

   ID_WWN=0x5002538d00000000
   ```

4. Create the **/etc/tuned/*my-profile*/tuned.conf** configuration file. In the file, set the following options:

   - Optional: Include an existing profile:

     ```
     [main]
     include=existing-profile
     ```

   - Set the selected disk scheduler for the device that matches the WWN identifier:

     ```
     [disk]
     devices_udev_regex=ID_WWN=0x5002538d00000000
     elevator=selected-scheduler
     ```

     To match multiple devices in the **devices_udev_regex** option, separate the identifiers with commas:

```
devices_udev_regex=ID_WWN=0x5002538d00000000,
ID_WWN=0x1234567800000000
```

5. Enable your profile:

```
# tuned-adm profile my-profile
```

6. Verify that the Tuned profile is active and applied:

```
$ tuned-adm active

Current active profile: my-profile
```

```
$ tuned-adm verify

Verification succeeded, current system settings match the preset profile.
See tuned log file ('/var/log/tuned/tuned.log') for details.
```

Additional resources

- For more information on creating a **Tuned** profile, see https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/8/html/monitoring_and_managing_system_status_and_performance/tuned-profiles_monitoring-and-managing-system-status-and-performance.

13.7. SETTING THE DISK SCHEDULER USING UDEV RULES

This procedure sets a given disk scheduler for specific block devices using **udev** rules. The setting persists across system reboots.

In the following commands and configuration, replace:

- *device* with the name of the block device, for example **sdf**
- *selected-scheduler* with the disk scheduler that you want to set for the device, for example **bfq**

Procedure

1. Find the World Wide Identifier (WWID) of the block device:

```
$ udevadm info --attribute-walk --name=/dev/device | grep wwid

    ATTRS{wwid}=="device WWID"
```

An example value of *device WWID* is:

```
t10.ATA     SAMSUNG MZNLN256HMHQ-000L7             S2WDNX0J336519
```

2. Configure the **udev** rule. Create the **/etc/udev/rules.d/99-scheduler.rules** file with the following content:

```
ACTION=="add|change", SUBSYSTEM=="block", ATTRS{wwid}=="device WWID",
ATTR{queue/scheduler}="selected-scheduler"
```

Replace *device WWID* with the WWID value that you found in the previous steps.

3. Reload **udev** rules:

```
# udevadm control --reload-rules
```

4. Apply the scheduler configuration:

```
# udevadm trigger --type=devices --action=change
```

5. Verify the active scheduler:

```
# cat /sys/block/device/queue/scheduler
```

13.8. TEMPORARILY SETTING A SCHEDULER FOR A SPECIFIC DISK

This procedure sets a given disk scheduler for specific block devices. The setting does not persist across system reboots.

Procedure

- Write the name of the selected scheduler to the **/sys/block/*device*/queue/scheduler** file:

```
# echo selected-scheduler > /sys/block/device/queue/scheduler
```

In the file name, replace *device* with the block device name, for example **sdc**.

Verification steps

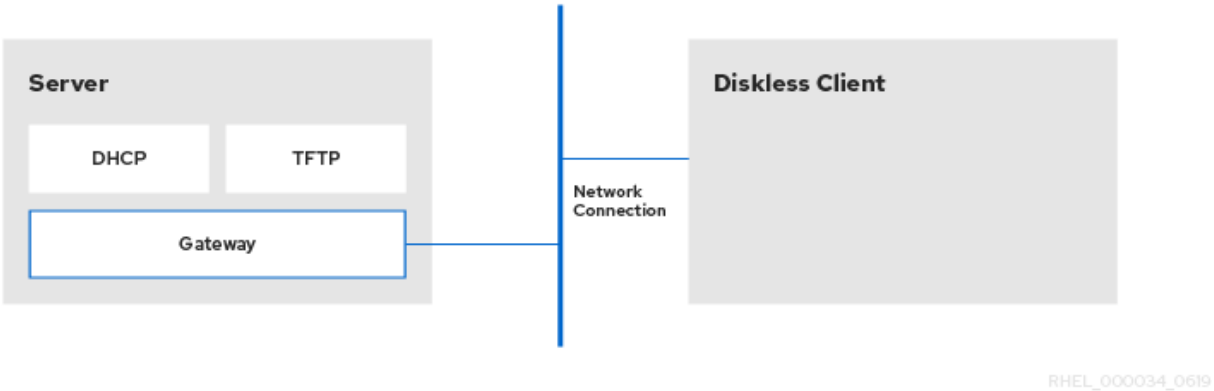
- Verify that the scheduler is active on the device:

```
# cat /sys/block/device/queue/scheduler
```

CHAPTER 14. SETTING UP A REMOTE DISKLESS SYSTEM

The following sections outline the necessary procedures for deploying remote diskless systems in a network environment. It is useful to implement this solution when you require multiple clients with identical configuration. Also, that will save the cost for hard drives for the number of the clients. Assuming, the server has Red Hat Enterprise Linux 8 operating system installed.

Figure 14.1. Remote diskless system settings diagram



Note, that gateway might be configured on a separate server.

14.1. PREPARING AN ENVIRONMENT FOR THE REMOTE DISKLESS SYSTEM

This procedure describes the preparation of the environment for the remote diskless system.

Remote diskless system booting requires both a **tftp** service (provided by **tftp-server**) and a DHCP service (provided by **dhcp**). The **tftp** service is used to retrieve kernel image and **initrd** over the network via the PXE loader.

Prerequisites

- Install the following packages:
  - **tftp-server**
  - **xinetd**
  - **dhcp-server**
  - **syslinux**
- Set up the network connection.

Procedure

1. Install the **dracut-network** package:

```
# yum install dracut-network
```

2. After installing the **dracut-network** package, add the following line to **/etc/dracut.conf**:

```
add_dracutmodules+="nfs"
```

IMPORTANT

Some RPM packages have started using file capabilities (such as **setcap** and **getcap**). However, NFS does not currently support these so attempting to install or update any packages that use file capabilities will fail.

At this point you have the server ready to continue with remote diskless system implementation.

14.2. CONFIGURING A TFTP SERVICE FOR DISKLESS CLIENTS

This procedure describes how to configure a tftp service for a diskless client.

Prerequisites

- Install the necessary packages. See prerequisites in Section 14.1, “Preparing an environment for the remote diskless system”.

To Configure tftp

1. Enable PXE booting over the network:

```
# systemctl enable --now tftp
```

2. The **tftp** root directory (**chroot**) is located in **/var/lib/tftpboot**. Copy **/usr/share/syslinux/pxelinux.0** to **/var/lib/tftpboot/**:

```
# cp /usr/share/syslinux/pxelinux.0 /var/lib/tftpboot/
```

3. Create a **pxelinux.cfg** directory inside the **tftp** root directory:

```
# mkdir -p /var/lib/tftpboot/pxelinux.cfg/
```

4. After configuring **tftp** for diskless clients, configure DHCP, NFS, and the exported file system accordingly.

14.3. CONFIGURING DHCP SERVER FOR DISKLESS CLIENTS

This procedure describes how to configure DHCP for a diskless system.

Prerequisites

- Install the necessary packages. See prerequisites in Section 14.1, “Preparing an environment for the remote diskless system”.
- Configure **tftp**. See Section 14.2, “Configuring a tftp service for diskless clients” .

Procedure

1. Set up a DHCP server and enable PXE booting by adding the following configuration to **/etc/dhcp/dhcpd.conf**:

```
allow booting;
allow bootp;
subnet 192.168.205.0 netmask 255.255.255.0 {
  pool
  {
    range 192.168.205.10 192.168.205.25;
  }

  option subnet-mask 255.255.255.0;
  option routers 192.168.205.1;
}
class "pxeclients" {
   match if substring(option vendor-class-identifier, 0, 9) = "PXEClient";
   next-server server-ip;
   filename "pxelinux.0";
}
```

This configuration will not boot over UEFI. To perform installation for UEFI, follow the procedure from this documentation: Configuring a TFTP server for UEFI-based clients . Also, note that the **/etc/dhcp/dhcpd.conf** is an example file.

NOTE

When **libvirt** virtual machines are used as a diskless client, **libvirt** provides the DHCP service and the stand alone DHCP server is not used. In this situation, network booting must be enabled with the **bootp file='*filename*'** option in the **libvirt** network configuration, **virsh net-edit**.

2. Enable **dhcpd.service** by entering the following command:

```
# systemctl enable --now dhcpd.service
```

14.4. CONFIGURING AN EXPORTED FILE SYSTEM FOR DISKLESS CLIENTS

This procedure describes how to configure an exported file system for diskless client.

Prerequisites

- Install the necessary packages. See prerequisites in Section 14.1, “Preparing an environment for the remote diskless system”.
- Configure **tftp**. See Section 14.2, “Configuring a tftp service for diskless clients” .
- Configure DHCP. See Section 14.3, “Configuring DHCP server for diskless clients” .

Procedure

1. Configure the NFS server to export the root directory by adding it to **/etc/exports**. For the instructions see NFS server configuration.

2. To accommodate completely diskless clients, the root directory should contain a complete Red Hat Enterprise Linux installation. You can either install a new base system or clone an existing installation:

   - To install Red Hat Enterprise Linux to the exported location, use the **yum** utility with the **--installroot** option:

     ```
     # yum install @Base kernel dracut-network nfs-utils \
         --installroot=exported-root-directory --releasever=/
     ```

   - To synchronize with a running system, use the **rsync** utility:

     ```
     # rsync -a -e ssh --exclude='/proc/' --exclude='/sys/' \
         example.com:/exported-root-directory
     ```

     - Replace *hostname.com* with the hostname of the running system with which to synchronize via **rsync**.
     - Replace *exported-root-directory* with the path to the exported file system.
       Note, that for this option you must have a separate existing running system, which you will clone to the server by the command above.

The file system to be exported still needs to be configured further before it can be used by diskless clients. To do this, perform the following procedure:

Configure File System

1. Select the kernel that diskless clients should use (**vmlinuz-*kernel-version***) and copy it to the **tftp** boot directory:

   ```
   # cp /exported-root-directory/boot/vmlinuz-kernel-version /var/lib/tftpboot/
   ```

2. Create the **initrd** (that is, **initramfs-*kernel-version*.img**) with NFS support:

   ```
   # dracut --add nfs initramfs-kernel-version.img kernel-version
   ```

3. Change file permissions for **initrd** to 644 using the following command:

   ```
   # chmod 644 /exported-root-directory/boot/initramfs-<kernel-version>.img
   ```

   WARNING

   If you do not change the initrd's file permissions, the **pxelinux.0** boot loader will fail with a "file not found" error.

4. Copy the resulting **initramfs-*kernel-version*.img** into the **tftp** boot directory:

   ```
   # cp /exported-root-directory/boot/initramfs-kernel-version.img /var/lib/tftpboot/
   ```

5. Edit the default boot configuration to use the **initrd** and kernel in the **/var/lib/tftpboot/** directory. This configuration should instruct the diskless client's root to mount the exported file system (***/exported-root-directory***) as read-write. Add the following configuration in the **/var/lib/tftpboot/pxelinux.cfg/default** file:

```
default rhel8

label rhel8
  kernel vmlinuz-kernel-version
  append initrd=initramfs-kernel-version.img root=nfs:server-ip:/exported-root-directory rw
```

   Replace ***server-ip*** with the IP address of the host machine on which the **tftp** and DHCP services reside.

6. Reboot the NFS server.

The NFS share is now ready for exporting to diskless clients. These clients can boot over the network via PXE.

14.5. RE-CONFIGURING A REMOTE DISKLESS SYSTEM

You need to re-configure the system in some cases. The steps below show how to change the password for a user, how to install software on a system and describe how to split system into a **/usr** that is in read-only mode and a **/var** that is in read-write mode.

Prerequisites

- **no_root_squash** option is enabled in the exported file system.

Procedure

1. To change the user password, follow the steps below:
   - Change the command line to */exported/root/directory*:

```
# chroot /exported/root/directory /bin/bash
```

   - Change the password for the user you want:

```
# passwd <username>
```

     Replace the *<username>* with a real user to whom you want to change the password.

   - Exit the command line:

```
# exit
```

2. To install software to a remote diskless system, use the following command:

```
# yum install <package> --installroot=/exported/root/directory --releasever=/ --config
/etc/dnf/dnf.conf --setopt=reposdir=/etc/yum.repos.d/
```

   Replace *<package>* with the actual package you want to install.

3. To split a remote diskless system into a **/usr** and a **/var** you must configure two separate exports. Read NFS server configuration documentation for details.

14.6. THE MOST COMMON ISSUES WITH LOADING A REMOTE DISKLESS SYSTEM

The following section describes the issues during loading the remote diskless system on a diskless client and shows the possible solution for them.

14.6.1. The client does not get an IP address

To troubleshoot that problem:

1. Check if the DHCP service is enabled on the server.
   - Check if the **dhcp.service** is running:

     ```
     # systemctl status dhcpd.service
     ```

   - If the **dhcp.service** is inactive, you must enable and start it:

     ```
     # systemctl enable dhcpd.service
     # systemctl start dhcpd.service
     ```

     Reboot the diskless client.
2. If the problem remains, check the DHCP configurational file */etc/dhcp/dhcpd.conf* on a server. For more information, see Section 14.3, “Configuring DHCP server for diskless clients” .
3. Check if the Firewall ports are opened.
   - Check if the **tftp.service** is listed in active services:

     ```
     # firewall-cmd --get-active-zones
     # firewall-cmd --info-zone=public
     ```

   - If the **tftp.service** is not listed in active services, add it to the list:

     ```
     # firewall-cmd --add-service=tftp
     ```

   - Check if the **nfs.service** is listed in active services:

     ```
     # firewall-cmd --get-active-zones
     # firewall-cmd --info-zone=public
     ```

   - If the **nfs.service** is not listed in active services, add it to the list:

     ```
     # firewall-cmd --add-service=nfs
     ```

14.6.2. The files are not available during the booting a remote diskless system

To troubleshoot this problem:

1. Check if the file is in place. The location on a server */var/lib/tftpboot/*.
2. If the file is in place, check its permissions:

```
# chmod 644 pxelinux.0
```

3. Check if the Firewall ports are opened.

14.6.3. System boot failed after loading kernel/initrd

To troubleshoot this problem:

1. Check if NFS service is enabled on a server.
   - Check if **nfs.service** is running:

```
# systemctl status nfs.service
```

   - If the **nfs.service** is inactive, you must enable and start it:

```
# systemctl enable nfs.service
# systemctl start nfs.service
```

2. Check if the parameters are correct in *pxelinux.cfg*. For more details, see Section 14.4, “Configuring an exported file system for diskless clients”.
3. Check if the Firewall ports are opened.

CHAPTER 15. MANAGING RAID

This chapter describes Redundant Array of Independent Disks (RAID). User can use RAID to store data across multiple drives. It also helps to avoid data loss if a drive has failed.

15.1. REDUNDANT ARRAY OF INDEPENDENT DISKS (RAID)

The basic idea behind RAID is to combine multiple devices, such as **HDD**, **SSD** or **NVMe**, into an array to accomplish performance or redundancy goals not attainable with one large and expensive drive. This array of devices appears to the computer as a single logical storage unit or drive.

RAID allows information to be spread across several devices. RAID uses techniques such as *disk striping* (RAID Level 0), *disk mirroring* (RAID Level 1), and *disk striping with parity* (RAID Levels 4, 5 and 6) to achieve redundancy, lower latency, increased bandwidth, and maximized ability to recover from hard disk crashes.

RAID distributes data across each device in the array by breaking it down into consistently-sized chunks (commonly 256K or 512k, although other values are acceptable). Each chunk is then written to a hard drive in the RAID array according to the RAID level employed. When the data is read, the process is reversed, giving the illusion that the multiple devices in the array are actually one large drive.

System Administrators and others who manage large amounts of data would benefit from using RAID technology. Primary reasons to deploy RAID include:

- Enhances speed
- Increases storage capacity using a single virtual disk
- Minimizes data loss from disk failure
- RAID layout and level online conversion

15.2. RAID TYPES

There are three possible RAID approaches: Firmware RAID, Hardware RAID, and Software RAID.

Firmware RAID

Firmware RAID, also known as ATARAID, is a type of software RAID where the RAID sets can be configured using a firmware-based menu. The firmware used by this type of RAID also hooks into the BIOS, allowing you to boot from its RAID sets. Different vendors use different on-disk metadata formats to mark the RAID set members. The Intel Matrix RAID is a good example of a firmware RAID system.

Hardware RAID

The hardware-based array manages the RAID subsystem independently from the host. It may present multiple devices per RAID array to the host.

Hardware RAID devices may be internal or external to the system. Internal devices commonly consisting of a specialized controller card that handles the RAID tasks transparently to the operating system. External devices commonly connect to the system via SCSI, Fibre Channel, iSCSI, InfiniBand, or other high speed network interconnect and present volumes such as logical units to the system.

RAID controller cards function like a SCSI controller to the operating system, and handle all the actual drive communications. The user plugs the drives into the RAID controller (just like a normal SCSI controller) and then adds them to the RAID controller's configuration. The operating system will not be

able to tell the difference.

Software RAID

Software RAID implements the various RAID levels in the kernel block device code. It offers the cheapest possible solution, as expensive disk controller cards or hot-swap chassis [1] are not required. Software RAID also works with any block storage which are supported by the Linux kernel, such as **SATA**, **SCSI**, and **NVMe**. With today's faster CPUs, Software RAID also generally outperforms Hardware RAID, unless you use high-end storage devices.

The Linux kernel contains a *multiple device* (MD) driver that allows the RAID solution to be completely hardware independent. The performance of a software-based array depends on the server CPU performance and load.

Key features of the Linux software RAID stack:

- Multithreaded design
- Portability of arrays between Linux machines without reconstruction
- Backgrounded array reconstruction using idle system resources
- Hot-swappable drive support
- Automatic CPU detection to take advantage of certain CPU features such as streaming Single Instruction Multiple Data (SIMD) support
- Automatic correction of bad sectors on disks in an array
- Regular consistency checks of RAID data to ensure the health of the array
- Proactive monitoring of arrays with email alerts sent to a designated email address on important events
- Write-intent bitmaps which drastically increase the speed of resync events by allowing the kernel to know precisely which portions of a disk need to be resynced instead of having to resync the entire array after a system crash
  Note, that *resync* is a process to synchronize the data over the devices in the existing RAID to achieve redundancy
- Resync checkpointing so that if you reboot your computer during a resync, at startup the resync will pick up where it left off and not start all over again
- The ability to change parameters of the array after installation, which is called *reshaping*. For example, you can grow a 4-disk RAID5 array to a 5-disk RAID5 array when you have a new device to add. This grow operation is done live and does not require you to reinstall on the new array
- Reshaping supports changing the number of devices, the RAID algorithm or size of the RAID array type, such as RAID4, RAID5, RAID6 or RAID10
- Takeover supports RAID level converting, such as RAID0 to RAID6

15.3. RAID LEVELS AND LINEAR SUPPORT

RAID supports various configurations, including levels 0, 1, 4, 5, 6, 10, and linear. These RAID types are defined as follows:

Level 0

RAID level 0, often called "striping," is a performance-oriented striped data mapping technique. This means the data being written to the array is broken down into stripes and written across the member disks of the array, allowing high I/O performance at low inherent cost but provides no redundancy.
Many RAID level 0 implementations will only stripe the data across the member devices up to the size of the smallest device in the array. This means that if you have multiple devices with slightly different sizes, each device will get treated as though it is the same size as the smallest drive. Therefore, the common storage capacity of a level 0 array is equal to the capacity of the smallest member disk in a Hardware RAID or the capacity of smallest member partition in a Software RAID multiplied by the number of disks or partitions in the array.

Level 1

RAID level 1, or "mirroring," has been used longer than any other form of RAID. Level 1 provides redundancy by writing identical data to each member disk of the array, leaving a "mirrored" copy on each disk. Mirroring remains popular due to its simplicity and high level of data availability. Level 1 operates with two or more disks, and provides very good data reliability and improves performance for read-intensive applications but at a relatively high cost. [2]
The storage capacity of the level 1 array is equal to the capacity of the smallest mirrored hard disk in a Hardware RAID or the smallest mirrored partition in a Software RAID. Level 1 redundancy is the highest possible among all RAID types, with the array being able to operate with only a single disk present.

Level 4

Level 4 uses parity [3] concentrated on a single disk drive to protect data. Because the dedicated parity disk represents an inherent bottleneck on all write transactions to the RAID array, level 4 is seldom used without accompanying technologies such as write-back caching, or in specific circumstances where the system administrator is intentionally designing the software RAID device with this bottleneck in mind (such as an array that will have little to no write transactions once the array is populated with data). RAID level 4 is so rarely used that it is not available as an option in Anaconda. However, it could be created manually by the user if truly needed.
The storage capacity of Hardware RAID level 4 is equal to the capacity of the smallest member partition multiplied by the number of partitions *minus one*. Performance of a RAID level 4 array will always be asymmetrical, meaning reads will outperform writes. This is because writes consume extra CPU and main memory bandwidth when generating parity, and then also consume extra bus bandwidth when writing the actual data to disks because you are writing not only the data, but also the parity. Reads need only read the data and not the parity unless the array is in a degraded state. As a result, reads generate less traffic to the drives and across the buses of the computer for the same amount of data transfer under normal operating conditions.

Level 5

This is the most common type of RAID. By distributing parity across all of an array's member disk drives, RAID level 5 eliminates the write bottleneck inherent in level 4. The only performance bottleneck is the parity calculation process itself. With modern CPUs and Software RAID, that is usually not a bottleneck at all since modern CPUs can generate parity very fast. However, if you have a sufficiently large number of member devices in a software RAID5 array such that the combined aggregate data transfer speed across all devices is high enough, then this bottleneck can start to come into play.
As with level 4, level 5 has asymmetrical performance, which reads substantially outperforming writes. The storage capacity of RAID level 5 is calculated the same way as with level 4.

Level 6

This is a common level of RAID when data redundancy and preservation, and not performance, are

the paramount concerns, but where the space inefficiency of level 1 is not acceptable. Level 6 uses a complex parity scheme to be able to recover from the loss of any two drives in the array. This complex parity scheme creates a significantly higher CPU burden on software RAID devices and also imposes an increased burden during write transactions. As such, level 6 is considerably more asymmetrical in performance than levels 4 and 5.
The total capacity of a RAID level 6 array is calculated similarly to RAID level 5 and 4, except that you must subtract 2 devices (instead of 1) from the device count for the extra parity storage space.

Level 10

This RAID level attempts to combine the performance advantages of level 0 with the redundancy of level 1. It also helps to alleviate some of the space wasted in level 1 arrays with more than 2 devices. With level 10, it is possible for instance to create a 3-drive array configured to store only 2 copies of each piece of data, which then allows the overall array size to be 1.5 times the size of the smallest devices instead of only equal to the smallest device (like it would be with a 3-device, level 1 array). This avoids CPU process usage to calculate parity like with RAID level 6, but it is less space efficient.
The creation of RAID level 10 is not supported during installation. It is possible to create one manually using the command line **mdadm** tool. For more information on the options and their respective performance trade-offs, see **man md**.

Linear RAID

Linear RAID is a grouping of drives to create a larger virtual drive. In linear RAID, the chunks are allocated sequentially from one member drive, going to the next drive only when the first is completely filled. This grouping provides no performance benefit, as it is unlikely that any I/O operations split between member drives. Linear RAID also offers no redundancy and decreases reliability; if any one member drive fails, the entire array cannot be used. The capacity is the total of all member disks.

15.4. LINUX RAID SUBSYSTEMS

The following subsystems compose RAID in Linux:

15.4.1. Linux Hardware RAID Controller Drivers

Hardware RAID controllers have no specific RAID subsystem in Linux. Because they use special RAID chipsets, hardware RAID controllers come with their own drivers; these drivers allow the system to detect the RAID sets as regular disks.

15.4.2. mdraid

The **mdraid** subsystem was designed as a software RAID solution for Linux; it is also the preferred solution for software RAID under Linux. This subsystem uses its own metadata format, generally referred to as native MD metadata.

mdraid also supports other metadata formats, known as external metadata. Red Hat Enterprise Linux 8 uses **mdraid** with external metadata to access ISW / IMSM (Intel firmware RAID) sets and SNIA DDF. **mdraid** sets are configured and controlled through the **mdadm** utility.

15.5. CREATING SOFTWARE RAID

Follow the steps in this procedure to create a Redundant Arrays of Independent Disks (RAID) device. RAID devices are constructed from multiple storage devices that are arranged to provide increased performance and, in some configurations, greater fault tolerance.

A RAID device is created in one step and disks are added or removed as necessary. You can configure one RAID partition for each physical disk in your system, so the number of disks available to the installation program determines the levels of RAID device available. For example, if your system has two hard drives, you cannot create a RAID 10 device, as it requires a minimum of three separate disks.

NOTE

On IBM Z, the storage subsystem uses RAID transparently. You do not have to configure software RAID manually.

Prerequisites

- You have selected two or more disks for installation before RAID configuration options are visible. At least two disks are required to create a RAID device.
- You have created a mount point. By configuring a mount point, you configure the RAID device.
- You have selected the **Custom** radio button on the **Installation Destination** window.

Procedure

1. From the left pane of the **Manual Partitioning** window, select the required partition.
2. Under the **Device(s)** section, click **Modify**. The **Configure Mount Point** dialog box opens.
3. Select the disks that you want to include in the RAID device and click **Select**.
4. Click the **Device Type** drop-down menu and select **RAID**.
5. Click the **File System** drop-down menu and select your preferred file system type.
6. Click the **RAID Level** drop-down menu and select your preferred level of RAID.
7. Click **Update Settings** to save your changes.
8. Click **Done** to apply the settings and return to the **Installation Summary** window.

A message is displayed at the bottom of the window if the specified RAID level requires more disks.

15.6. CREATING SOFTWARE RAID AFTER INSTALLATION

This procedure describes how to create a software Redundant Array of Independent Disks (RAID) on an existing system using **mdadm** utility.

Prerequisites

- The **mdadm** package installed.
- Two or more partitions exist on your system. For detailed instruction, see Section 2.3, “Creating a partition”.

Procedure

1. To create RAID of two block devices with names **/dev/sda1** and **/dev/sdc1**, use the following command:

```
# mdadm --create /dev/md0 --level=<level_value> --raid-devices=2 /dev/sda1 /dev/sdc1
```

Replace *<level_value>* to a RAID level option. For more information, see **mdadm(8)** man page.

2. Optionally, to check the status of RAID, use the following command:

```
# mdadm --detail /dev/md0
```

3. Optionally, to observe the detailed information about each RAID device, use the following command:

```
# mdadm --examine /dev/sda1 /dev/sdc1
```

4. To create a file system on a RAID drive, use the following command:

```
# mkfs -t <file-system-name> /dev/md0
```

where *<file-system-name>* is a specific file system that you chose to format the drive with. For more information, see **mkfs** man page.

5. To create a mount point for RAID drive and mount it, use the following commands:

```
# mkdir /mnt/raid1
# mount /dev/md0 /mnt/raid1
```

After you finish the steps above, the RAID is ready to be used.

15.7. RECONFIGURING RAID

The section below describes how to modify an existing RAID. To do so, choose one of the methods:

- Changing RAID attributes (also known as RAID *reshape*).
- Converting RAID level (also known as RAID *takeover*).

15.7.1. Reshaping RAID

This chapter below describes how to reshape RAID. You can choose one of the methods of resizing RAID:

- Enlarging (extending) RAID.
- Shrinking RAID.

15.7.1.1. Resizing RAID (extending)

This procedure describes how to enlarge RAID. Assuming **/dev/md0** is RAID you want to enlarge.

Prerequisites

- Enough disk space.
- The package **parted** is installed.

Procedure

1. Extend RAID partitions. To do so, follow the instruction in Resizing a partition documentation.
2. To extend RAID to the maximum of the partition capacity, use this command:

```
# mdadm --grow --size=max /dev/md0
```

   Note that to determine a specific size, you must write the *--size* parameter in kB (for example *--size=524228*).

3. Increase the size of file system. For more information, check the Managing file systems documentation.

15.7.1.2. Resizing RAID (shrinking)

This procedure describes how to shrink RAID. Assuming **/dev/md0** is the RAID you want to shrink to 512 MB.

Prerequisites

- The package **parted** is installed.

Procedure

1. Shrink the file system. To do so, check the Managing file systems documentation.

IMPORTANT

The *XFS* file system does not support shrinking.

2. To decrease the RAID to the size of 512 MB, use this command:

```
# mdadm --grow --size=524228 /dev/md0
```

   Note, you must write the *--size* parameter in kB.

3. Shrink the partition to the size you need. To do so, follow the instruction in the Resizing a partition documentation.

15.7.2. RAID takeover

This chapter describes supported conversions in RAID and contains procedures to accomplish those conversions.

15.7.2.1. Supported RAID conversions

It is possible to convert from one RAID level to another. This section provides a table that lists supported RAID conversions.

	RAID0	RAID1	RAID4	RAID5	RAID6	RAID10
RAID0	✗	✗	✓	✓	✗	✓
RAID1	✗	✗	✗	✓	✗	✗
RAID4	✗	✗	✗	✓	✗	✗
RAID5	✓	✓	✓	✗	✓	✓
RAID6	✗	✗	✗	✓	✗	✗
RAID10	✓	✗	✗	✗	✗	✗

For example, you can convert RAID level 0 to RAID level 4, RAID level 5 and RAID level 10

Additional resources

- For more information about RAID level conversion, read **mdadm** man page.

15.7.2.2. Converting RAID level

This procedure describes how to convert RAID to a different RAID level. Assuming, you want to convert RAID **/dev/md0** level 0 to RAID level 5 and add one more disk **/dev/sdd** to the array.

Prerequisites

- Enough disks for conversion.
- The package **mdadm** is installed.
- Ensure the intended conversion is supported. To check if that is the case, see the table in Section 15.7.2.1, “Supported RAID conversions” .

Procedure

1. To convert the RAID **/dev/md0** to RAID level 5, use the following command:

```
# mdadm --grow --level=5 -n 3 /dev/md0 --force
```

2. To add a new disk to the array, use the following command:

```
# mdadm --manage /dev/md0 --add /dev/sdd
```

3. To check new details of the converted array, use the following command:

```
# mdadm --detail /dev/md0
```

Additional resources

- For more information about RAID level conversion, read **mdadm** man page.

15.8. CONVERTING A ROOT DISK TO RAID1 AFTER INSTALLATION

This section describes how to convert a non-RAID root disk to a RAID1 mirror after installing Red Hat Enterprise Linux 8.

On the PowerPC (PPC) architecture, take the following additional steps:

Prerequisites

- The instructions in the following Red Hat Knowledgebase article are completed: How do I convert my root disk to RAID1 after installation of Red Hat Enterprise Linux 7?.

Procedure

1. Copy the contents of the PowerPC Reference Platform (PReP) boot partition from **/dev/sda1** to **/dev/sdb1**:

```
# dd if=/dev/sda1 of=/dev/sdb1
```

2. Update the Prep and boot flag on the first partition on both disks:

```
$ parted /dev/sda set 1 prep on
$ parted /dev/sda set 1 boot on

$ parted /dev/sdb set 1 prep on
$ parted /dev/sdb set 1 boot on
```

NOTE

Running the **grub2-install /dev/sda** command does not work on a PowerPC machine and returns an error, but the system boots as expected.

15.9. CREATING ADVANCED RAID DEVICES

In some cases, you may wish to install the operating system on an array that can not be created after the installation completes. Usually, this means setting up the **/boot** or root file system arrays on a complex RAID device; in such cases, you may need to use array options that are not supported by **Anaconda** installer. To work around this, perform the following procedure:

Procedure

1. Insert the install disk.
2. During the initial boot up, select **Rescue Mode** instead of **Install** or **Upgrade**. When the system fully boots into *Rescue mode*, the user will be presented with a command line terminal.
3. From this terminal, use **parted** to create RAID partitions on the target hard drives. Then, use **mdadm** to manually create raid arrays from those partitions using any and all settings and options available. For more information on how to do these, see **man parted** and **man mdadm**.
4. Once the arrays are created, you can optionally create file systems on the arrays as well.

5. Reboot the computer and select **Install** or **Upgrade** to install as normal. As **Anaconda** installer searches the disks in the system, it will find the pre-existing RAID devices.

6. When asked about how to use the disks in the system, select **Custom Layout** and click **Next**. In the device listing, the pre-existing MD RAID devices will be listed.

7. Select a RAID device, click **Edit** and configure its mount point and (optionally) the type of file system it should use (if you did not create one earlier) then click **Done**. **Anaconda** will perform the install to this pre-existing RAID device, preserving the custom options you selected when you created it in *Rescue Mode*.

NOTE

The limited *Rescue Mode* of the installer does not include **man** pages. Both the **man mdadm** and **man md** contain useful information for creating custom RAID arrays, and may be needed throughout the workaround. As such, it can be helpful to either have access to a machine with these **man** pages present, or to print them out prior to booting into *Rescue Mode* and creating your custom arrays.

15.10. MONITORING RAID

This module describes how to set up the RAID monitoring option with **mdadm** tool.

Prerequisites

- The package **mdadm** is installed
- The mail service is set up.

Procedure

1. To create a configuration file for monitoring array you must scan the details and forward the result to **/etc/mdadm.conf** file. To do so, use the following command:

```
# mdadm --detail --scan >> /etc/mdadm.conf
```

   Note, that *ARRAY* and *MAILADDR* are mandatory variables.

2. Open the configuration file **/etc/mdadm.conf** with a text editor of your choice.

3. Add the *MAILADDR* variable with the mail address for the notification. For example, add new line:

```
MAILADDR <example@example.com>
```

   where *example@example.com* is an email address to which you want to receive the alerts from the array monitoring.

4. Save changes in the **/etc/mdadm.conf** file and close it.

After you complete the steps above, the monitoring system will send the alerts to the email address.

Additional resources

- For more information, read the **mdadm.conf 5** man page.

15.11. MAINTAINING RAID

This section provides various procedures for RAID maintenance.

15.11.1. Replacing a faulty disk in a RAID

This procedure describes how to replace the faulty disk in a redundant array of independent disks (RAID). Assuming, you have **/dev/md0** RAID level 10. In this scenario, the **/dev/sdg** disk is faulty and you need to replace it with new disk **/dev/sdh**.

Prerequisites

- Additional spare disk.
- The **mdadm** package is installed.
- A notification about a faulty disk in an array. To set up the array monitoring, see Section 15.10, “Monitoring RAID”.

Procedure

1. Ensure which disk is failing. To do so, enter the following command:

```
# journalctl -k -f
```

   You will find a message showing you which disk has failed:

```
md/raid:md0: Disk failure on sdg, disabling device.
md/raid:md0: Operation continuing on 5 devices.
```

2. Press **Ctrl+C** on your keyboard to exit the **journalctl** program.
3. Add a new disk to the array. To do so, enter the following command:

```
# mdadm --manage /dev/md0 --add /dev/sdh
```

4. Mark the failed disk as faulty. To do so, enter the following command:

```
# mdadm --manage /dev/md0 --fail /dev/sdg
```

5. Check if the faulty disk was masked correctly by using the following command:

```
# mdadm --detail /dev/md0
```

   At the end of the last command output you will see information about RAID disks similar to this where disk **/dev/sdg** has a **faulty** status:

```
    Number   Major   Minor   RaidDevice State
       0       8       16        0      active sync   /dev/sdb
       1       8       32        1      active sync   /dev/sdc
       2       8       48        2      active sync   /dev/sdd
```

```
      3       8       64        3       active sync   /dev/sde
      4       8       80        4       active sync   /dev/sdf
      6       8       112       5       active sync   /dev/sdh

      5       8       96        -       faulty   /dev/sdg
```

6. Finally, remove the faulty disk from the array. To do so, enter the following command:

```
# mdadm --manage /dev/md0 --remove /dev/sdg
```

7. Check RAID details by using following command:

```
# mdadm --detail /dev/md0
```

At the end of the last command output you will see information about RAID disks similar to this:

```
Number   Major   Minor   RaidDevice State
   0       8       16        0       active sync   /dev/sdb
   1       8       32        1       active sync   /dev/sdc
   2       8       48        2       active sync   /dev/sdd
   3       8       64        3       active sync   /dev/sde
   4       8       80        4       active sync   /dev/sdf
   6       8       112       5       active sync   /dev/sdh
```

After completing the steps above you will have RAID **/dev/md0** with a new disk **/dev/sdh**.

15.11.2. Replacing a broken disk in array

This procedure describes how to replace the broken disk in a redundant array of independent disks (RAID). Assuming, you have **/dev/md0** RAID level 6. In this scenario, the **/dev/sdb** disk has hardware issue and could not be used any longer. You need to replace it with the new disk **/dev/sdi**.

Prerequisites

- New disk for replacement.
- The **mdadm** package is installed.

Procedure

1. Check the log message by using the following command:

```
# journalctl -k -f
```

You will find a message showing you which disk has failed:

```
md/raid:md0: Disk failure on sdb, disabling device.
md/raid:md0: Operation continuing on 5 devices.
```

2. Press **Ctrl+C** on your keyboard to exit the **journalctl** program.
3. Add the new disk to the array as a spare one. To do so, enter the following command:

```
# mdadm --manage /dev/md0 --add /dev/sdi
```

4. Mark the broken disk as **faulty**. To do so, enter the following command:

```
# mdadm --manage /dev/md0 --fail /dev/sdb
```

5. Remove the faulty disk from the array. To do so, enter the following command:

```
# mdadm --manage /dev/md0 --remove /dev/sdb
```

6. Check the status of the array by using the following command:

```
# mdadm --detail /dev/md0
```

   At the end of the last command output you will see information about RAID disks similar to this:

```
Number   Major   Minor   RaidDevice State
   7       8      128       0      active sync   /dev/sdi
   1       8       32       1      active sync   /dev/sdc
   2       8       48       2      active sync   /dev/sdd
   3       8       64       3      active sync   /dev/sde
   4       8       80       4      active sync   /dev/sdf
   6       8      112       5      active sync   /dev/sdh
```

After completing the steps above you will have RAID **/dev/md0** with a new disk **/dev/sdi**.

15.11.3. Resynchronizing RAID disks

This procedure describes how to resynchronize disks in a RAID array. Assuming, you have **/dev/md0** RAID.

Prerequisites

- Package **mdadm** is installed.

Procedure

1. To check the array for the failed disks behavior, enter the following command:

```
# echo check > /sys/block/md0/md/sync_action
```

   That action will check the array and write the result into the **/sys/block/md0/md/sync_action** file.

2. Open file **/sys/block/md0/md/sync_action** with the text editor of your choice and see if there is any message about disk synchronization failures.

3. To resynchronize the disks in the array, enter the following command:

```
# echo repair > /sys/block/md0/md/sync_action
```

   This action will resynchronize the disks in the array and write the result into the **/sys/block/md0/md/sync_action** file.

4. To view the synchronization progress, enter the following command:

```
# cat /proc/mdstat
```

[1] A hot-swap chassis allows you to remove a hard drive without having to power-down your system.

[2] RAID level 1 comes at a high cost because you write the same information to all of the disks in the array, provides data reliability, but in a much less space-efficient manner than parity based RAID levels such as level 5. However, this space inefficiency comes with a performance benefit: parity-based RAID levels consume considerably more CPU power in order to generate the parity while RAID level 1 simply writes the same data more than once to the multiple RAID members with very little CPU overhead. As such, RAID level 1 can outperform the parity-based RAID levels on machines where software RAID is employed and CPU resources on the machine are consistently taxed with operations other than RAID activities.

[3] Parity information is calculated based on the contents of the rest of the member disks in the array. This information can then be used to reconstruct data when one disk in the array fails. The reconstructed data can then be used to satisfy I/O requests to the failed disk before it is replaced and to repopulate the failed disk after it has been replaced.

CHAPTER 16. ENCRYPTING BLOCK DEVICES USING LUKS

Disk encryption protects the data on a block device by encrypting it. To access the device's decrypted contents, a user must provide a passphrase or key as authentication. This is particularly important when it comes to mobile computers and removable media: it helps to protect the device's contents even if it has been physically removed from the system. The LUKS format is a default implementation of block device encryption in RHEL.

16.1. LUKS DISK ENCRYPTION

The Linux Unified Key Setup-on-disk-format (LUKS) enables you to encrypt block devices and it provides a set of tools that simplifies managing the encrypted devices. LUKS allows multiple user keys to decrypt a master key, which is used for the bulk encryption of the partition.

RHEL utilizes LUKS to perform block device encryption. By default, the option to encrypt the block device is unchecked during the installation. If you select the option to encrypt your disk, the system prompts you for a passphrase every time you boot the computer. This passphrase "unlocks" the bulk encryption key that decrypts your partition. If you choose to modify the default partition table, you can choose which partitions you want to encrypt. This is set in the partition table settings.

What LUKS does

- LUKS encrypts entire block devices and is therefore well-suited for protecting contents of mobile devices such as removable storage media or laptop disk drives.
- The underlying contents of the encrypted block device are arbitrary, which makes it useful for encrypting swap devices. This can also be useful with certain databases that use specially formatted block devices for data storage.
- LUKS uses the existing device mapper kernel subsystem.
- LUKS provides passphrase strengthening which protects against dictionary attacks.
- LUKS devices contain multiple key slots, allowing users to add backup keys or passphrases.

What LUKS does*not* do

- Disk-encryption solutions like LUKS protect the data only when your system is off. Once the system is on and LUKS has decrypted the disk, the files on that disk are available to anyone who would normally have access to them.
- LUKS is not well-suited for scenarios that require many users to have distinct access keys to the same device. The LUKS1 format provides eight key slots, LUKS2 up to 32 key slots.
- LUKS is not well-suited for applications requiring file-level encryption.

Ciphers

The default cipher used for LUKS is **aes-xts-plain64**. The default key size for LUKS is 512 bits. The default key size for LUKS with **Anaconda** (XTS mode) is 512 bits. Ciphers that are available are:

- AES - Advanced Encryption Standard - FIPS PUB 197
- Twofish (a 128-bit block cipher)
- Serpent

Additional resources

- LUKS Project Home Page
- LUKS On-Disk Format Specification

16.2. LUKS VERSIONS IN RHEL 8

In RHEL 8, the default format for LUKS encryption is LUKS2. The legacy LUKS1 format remains fully supported and it is provided as a format compatible with earlier RHEL releases.

The LUKS2 format is designed to enable future updates of various parts without a need to modify binary structures. LUKS2 internally uses JSON text format for metadata, provides redundancy of metadata, detects metadata corruption and allows automatic repairs from a metadata copy.

IMPORTANT

Do not use LUKS2 in systems that need to be compatible with legacy systems that support only LUKS1. Note that RHEL 7 supports the LUKS2 format since version 7.6.

WARNING

LUKS2 and LUKS1 use different commands to encrypt the disk. Using the wrong command for a LUKS version might cause data loss.

LUKS version	Encryption command
LUKS2	**cryptsetup reencrypt**
LUKS1	**cryptsetup-reencrypt**

Online re-encryption

The LUKS2 format supports re-encrypting encrypted devices while the devices are in use. For example, you do not have to unmount the file system on the device to perform the following tasks:

- Change the volume key
- Change the encryption algorithm

When encrypting a non-encrypted device, you must still unmount the file system. You can remount the file system after a short initialization of the encryption.

The LUKS1 format does not support online re-encryption.

Conversion

The LUKS2 format is inspired by LUKS1. In certain situations, you can convert LUKS1 to LUKS2. The conversion is not possible specifically in the following scenarios:

- A LUKS1 device is marked as being used by a Policy-Based Decryption (PBD - Clevis) solution. The **cryptsetup** tool refuses to convert the device when some **luksmeta** metadata are detected.
- A device is active. The device must be in the inactive state before any conversion is possible.

16.3. OPTIONS FOR DATA PROTECTION DURING LUKS2 RE-ENCRYPTION

LUKS2 provides several options that prioritize performance or data protection during the re-encryption process:

checksum

This is the default mode. It balances data protection and performance.
This mode stores individual checksums of the sectors in the re-encryption area, so the recovery process can detect which sectors LUKS2 already re-encrypted. The mode requires that the block device sector write is atomic.

journal

That is the safest mode but also the slowest. This mode journals the re-encryption area in the binary area, so LUKS2 writes the data twice.

none

This mode prioritizes performance and provides no data protection. It protects the data only against safe process termination, such as the **SIGTERM** signal or the user pressing **Ctrl**+**C**. Any unexpected system crash or application crash might result in data corruption.

You can select the mode using the **--resilience** option of **cryptsetup**.

If a LUKS2 re-encryption process terminates unexpectedly by force, LUKS2 can perform the recovery in one of the following ways:

- Automatically, during the next LUKS2 device open action. This action is triggered either by the **cryptsetup open** command or by attaching the device with **systemd-cryptsetup**.
- Manually, by using the **cryptsetup repair** command on the LUKS2 device.

16.4. ENCRYPTING EXISTING DATA ON A BLOCK DEVICE USING LUKS2

This procedure encrypts existing data on a not yet encrypted device using the LUKS2 format. A new LUKS header is stored in the head of the device.

Prerequisites

- The block device contains a file system.
- You have backed up your data.

WARNING

You might lose your data during the encryption process: due to a hardware, kernel, or human failure. Ensure that you have a reliable backup before you start encrypting the data.

Procedure

1. Unmount all file systems on the device that you plan to encrypt. For example:

```
# umount /dev/sdb1
```

2. Make free space for storing a LUKS header. Choose one of the following options that suits your scenario:

   - In the case of encrypting a logical volume, you can extend the logical volume without resizing the file system. For example:

```
# lvextend -L+32M vg00/lv00
```

   - Extend the partition using partition management tools, such as **parted**.
   - Shrink the file system on the device. You can use the **resize2fs** utility for the ext2, ext3, or ext4 file systems. Note that you cannot shrink the XFS file system.

3. Initialize the encryption. For example:

```
# cryptsetup reencrypt \
        --encrypt \
        --init-only \
        --reduce-device-size 32M \
        /dev/sdb1 sdb1_encrypted
```

   The command asks you for a passphrase and starts the encryption process.

4. Mount the device:

```
# mount /dev/mapper/sdb1_encrypted /mnt/sdb1_encrypted
```

5. Start the online encryption:

```
# cryptsetup reencrypt --resume-only /dev/sdb1
```

Additional resources

- For more details, see the **cryptsetup(8)**, **lvextend(8)**, **resize2fs(8)**, and **parted(8)** man pages.

16.5. ENCRYPTING EXISTING DATA ON A BLOCK DEVICE USING LUKS2 WITH A DETACHED HEADER

This procedure encrypts existing data on a block device without creating free space for storing a LUKS header. The header is stored in a detached location, which also serves as an additional layer of security. The procedure uses the LUKS2 encryption format.

Prerequisites

- The block device contains a file system.
- You have backed up your data.

WARNING

You might lose your data during the encryption process: due to a hardware, kernel, or human failure. Ensure that you have a reliable backup before you start encrypting the data.

Procedure

1. Unmount all file systems on the device. For example:

```
# umount /dev/sdb1
```

2. Initialize the encryption:

```
# cryptsetup reencrypt \
        --encrypt \
        --init-only \
        --header /path/to/header \
        /dev/sdb1 sdb1_encrypted
```

   Replace */path/to/header* with a path to the file with a detached LUKS header. The detached LUKS header has to be accessible so that the encrypted device can be unlocked later.

   The command asks you for a passphrase and starts the encryption process.

3. Mount the device:

```
# mount /dev/mapper/sdb1_encrypted /mnt/sdb1_encrypted
```

4. Start the online encryption:

```
# cryptsetup reencrypt --resume-only --header /path/to/header /dev/sdb1
```

Additional resources

- For more details, see the **cryptsetup(8)** man page.

CHAPTER 17. MANAGING LAYERED LOCAL STORAGE WITH STRATIS

You can easily set up and manage complex storage configurations integrated by the Stratis high-level system.

IMPORTANT

Stratis is available as a Technology Preview. For information on Red Hat scope of support for Technology Preview features, see the Technology Preview Features Support Scope document.

Customers deploying Stratis are encouraged to provide feedback to Red Hat.

17.1. SETTING UP STRATIS FILE SYSTEMS

As a system administrator, you can enable and set up the Stratis volume-managing file system on your system to easily manage layered storage.

17.1.1. The purpose and features of Stratis

Stratis is a local storage-management solution for Linux. It is focused on simplicity and ease of use, and gives you access to advanced storage features.

Stratis makes the following activities easier:

- Initial configuration of storage
- Making changes later
- Using advanced storage features

Stratis is a hybrid user-and-kernel local storage management system that supports advanced storage features. The central concept of Stratis is a storage *pool*. This pool is created from one or more local disks or partitions, and volumes are created from the pool.

The pool enables many useful features, such as:

- File system snapshots
- Thin provisioning
- Tiering

17.1.2. Components of a Stratis volume

Externally, Stratis presents the following volume components in the command-line interface and the API:

blockdev

Block devices, such as a disk or a disk partition.

pool

Composed of one or more block devices.

A pool has a fixed total size, equal to the size of the block devices.

The pool contains most Stratis layers, such as the non-volatile data cache using the **dm-cache** target.

Stratis creates a /**stratis**/***my-pool***/ directory for each pool. This directory contains links to devices that represent Stratis file systems in the pool.

filesystem

Each pool can contain one or more file systems, which store files.
File systems are thinly provisioned and do not have a fixed total size. The actual size of a file system grows with the data stored on it. If the size of the data approaches the virtual size of the file system, Stratis grows the thin volume and the file system automatically.

The file systems are formatted with XFS.

IMPORTANT

Stratis tracks information about file systems created using Stratis that XFS is not aware of, and changes made using XFS do not automatically create updates in Stratis. Users must not reformat or reconfigure XFS file systems that are managed by Stratis.

Stratis creates links to file systems at the /**stratis**/***my-pool***/***my-fs*** path.

NOTE

Stratis uses many Device Mapper devices, which show up in **dmsetup** listings and the /**proc**/**partitions** file. Similarly, the **lsblk** command output reflects the internal workings and layers of Stratis.

17.1.3. Block devices usable with Stratis

This section lists storage devices that you can use for Stratis.

Supported devices
Stratis pools have been tested to work on these types of block devices:

- LUKS
- LVM logical volumes
- MD RAID
- DM Multipath
- iSCSI
- HDDs and SSDs
- NVMe devices

WARNING

In the current version, Stratis does not handle failures in hard drives or other hardware. If you create a Stratis pool over multiple hardware devices, you increase the risk of data loss because multiple devices must be operational to access the data.

Unsupported devices
Because Stratis contains a thin-provisioning layer, Red Hat does not recommend placing a Stratis pool on block devices that are already thinly-provisioned.

Additional resources

- For iSCSI and other block devices requiring network, see the **systemd.mount(5)** man page for information on the **_netdev** mount option.

17.1.4. Installing Stratis

This procedure installs all packages necessary to use Stratis.

Procedure

1. Install packages that provide the Stratis service and command-line utilities:

```
# yum install stratisd stratis-cli
```

2. Make sure that the **stratisd** service is enabled:

```
# systemctl enable --now stratisd
```

17.1.5. Creating a Stratis pool

This procedure creates a Stratis pool from one or more block devices.

Prerequisites

- Stratis is installed. See Section 17.1.4, “Installing Stratis”.
- The **stratisd** service is running.
- The block devices on which you are creating a Stratis pool are not in use and not mounted.
- The block devices on which you are creating a Stratis pool are at least 1 GiB in size each.
- On the IBM Z architecture, the **/dev/dasd*** block devices must to be partitioned. Use the partition in the Stratis pool.
  For information on partitioning DASD devices, see Configuring a Linux instance on IBM Z .

Procedure

1. If the selected block device contains file system, partition table, or RAID signatures, erase them:

```
# wipefs --all block-device
```

   Replace *block-device* with the path to a block device, such as **/dev/sdb**.

2. To create a Stratis pool on the block device, use:

```
# stratis pool create my-pool block-device
```

   - Replace *my-pool* with an arbitrary name for the pool.
   - Replace *block-device* with the path to the empty or wiped block device, such as **/dev/sdb**.

   To create a pool from more than one block device, list them all on the command line:

```
# stratis pool create my-pool device-1 device-2 device-n
```

3. To verify, list all pools on your system:

```
# stratis pool list
```

Additional resources

- The **stratis(8)** man page

Next steps

- Create a Stratis file system on the pool. See Section 17.1.6, "Creating a Stratis file system" .

17.1.6. Creating a Stratis file system

This procedure creates a Stratis file system on an existing Stratis pool.

Prerequisites

- Stratis is installed. See Section 17.1.4, "Installing Stratis".
- The **stratisd** service is running.
- You have created a Stratis pool. See Section 17.1.5, "Creating a Stratis pool" .

Procedure

1. To create a Stratis file system on a pool, use:

```
# stratis fs create my-pool my-fs
```

   - Replace *my-pool* with the name of your existing Stratis pool.
   - Replace *my-fs* with an arbitrary name for the file system.

2. To verify, list file systems within the pool:

```
# stratis fs list my-pool
```

Additional resources

- The **stratis(8)** man page

Next steps

- Mount the Stratis file system. See Section 17.1.7, "Mounting a Stratis file system".

17.1.7. Mounting a Stratis file system

This procedure mounts an existing Stratis file system to access the content.

Prerequisites

- Stratis is installed. See Section 17.1.4, "Installing Stratis".
- The **stratisd** service is running.
- You have created a Stratis file system. See Section 17.1.6, "Creating a Stratis file system".

Procedure

- To mount the file system, use the entries that Stratis maintains in the **/stratis/** directory:

```
# mount /stratis/my-pool/my-fs mount-point
```

The file system is now mounted on the *mount-point* directory and ready to use.

Additional resources

- The **mount(8)** man page

17.1.8. Persistently mounting a Stratis file system

This procedure persistently mounts a Stratis file system so that it is available automatically after booting the system.

Prerequisites

- Stratis is installed. See Section 17.1.4, "Installing Stratis".
- The **stratisd** service is running.
- You have created a Stratis file system. See Section 17.1.6, "Creating a Stratis file system".

Procedure

1. Determine the UUID attribute of the file system:

```
$ lsblk --output=UUID /stratis/my-pool/my-fs
```

For example:

Example 17.1. Viewing the UUID of Stratis file system

```
$ lsblk --output=UUID /stratis/my-pool/fs1

UUID
a1f0b64a-4ebb-4d4e-9543-b1d79f600283
```

2. If the mount point directory does not exist, create it:

```
# mkdir --parents mount-point
```

3. As root, edit the **/etc/fstab** file and add a line for the file system, identified by the UUID. Use **xfs** as the file system type and add the **x-systemd.requires=stratisd.service** option.
For example:

Example 17.2. The /fs1 mount point in /etc/fstab

```
UUID=a1f0b64a-4ebb-4d4e-9543-b1d79f600283 /fs1 xfs defaults,x-
systemd.requires=stratisd.service 0 0
```

4. Regenerate mount units so that your system registers the new configuration:

```
# systemctl daemon-reload
```

5. Try mounting the file system to verify that the configuration works:

```
# mount mount-point
```

Additional resources

- Persistently mounting file systems

17.1.9. Related information

- The *Stratis Storage* website: https://stratis-storage.github.io/

17.2. EXTENDING A STRATIS VOLUME WITH ADDITIONAL BLOCK DEVICES

You can attach additional block devices to a Stratis pool to provide more storage capacity for Stratis file systems.

17.2.1. Components of a Stratis volume

Externally, Stratis presents the following volume components in the command-line interface and the API:

blockdev

Block devices, such as a disk or a disk partition.

pool

Composed of one or more block devices.
A pool has a fixed total size, equal to the size of the block devices.

The pool contains most Stratis layers, such as the non-volatile data cache using the **dm-cache** target.

Stratis creates a /**stratis**/***my-pool***/ directory for each pool. This directory contains links to devices that represent Stratis file systems in the pool.

filesystem

Each pool can contain one or more file systems, which store files.
File systems are thinly provisioned and do not have a fixed total size. The actual size of a file system grows with the data stored on it. If the size of the data approaches the virtual size of the file system, Stratis grows the thin volume and the file system automatically.

The file systems are formatted with XFS.

IMPORTANT

Stratis tracks information about file systems created using Stratis that XFS is not aware of, and changes made using XFS do not automatically create updates in Stratis. Users must not reformat or reconfigure XFS file systems that are managed by Stratis.

Stratis creates links to file systems at the /**stratis**/***my-pool***/***my-fs*** path.

NOTE

Stratis uses many Device Mapper devices, which show up in **dmsetup** listings and the /**proc**/**partitions** file. Similarly, the **lsblk** command output reflects the internal workings and layers of Stratis.

17.2.2. Adding block devices to a Stratis pool

This procedure adds one or more block devices to a Stratis pool to be usable by Stratis file systems.

Prerequisites

- Stratis is installed. See Section 17.1.4, “Installing Stratis”.
- The **stratisd** service is running.
- The block devices that you are adding to the Stratis pool are not in use and not mounted.
- The block devices that you are adding to the Stratis pool are at least 1 GiB in size each.

Procedure

- To add one or more block devices to the pool, use:

```
# stratis pool add-data my-pool device-1 device-2 device-n
```

Additional resources

- The **stratis(8)** man page

17.2.3. Related information

- The *Stratis Storage* website: https://stratis-storage.github.io/

17.3. MONITORING STRATIS FILE SYSTEMS

As a Stratis user, you can view information about Stratis volumes on your system to monitor their state and free space.

17.3.1. Stratis sizes reported by different utilities

This section explains the difference between Stratis sizes reported by standard utilities such as **df** and the **stratis** utility.

Standard Linux utilities such as **df** report the size of the XFS file system layer on Stratis, which is 1 TiB. This is not useful information, because the actual storage usage of Stratis is less due to thin provisioning, and also because Stratis automatically grows the file system when the XFS layer is close to full.

IMPORTANT

Regularly monitor the amount of data written to your Stratis file systems, which is reported as the *Total Physical Used* value. Make sure it does not exceed the *Total Physical Size* value.

Additional resources

- The **stratis(8)** man page

17.3.2. Displaying information about Stratis volumes

This procedure lists statistics about your Stratis volumes, such as the total, used, and free size or file systems and block devices belonging to a pool.

Prerequisites

- Stratis is installed. See Section 17.1.4, “Installing Stratis”.
- The **stratisd** service is running.

Procedure

- To display information about all **block devices** used for Stratis on your system:

```
# stratis blockdev

Pool Name  Device Node    Physical Size   State  Tier
my-pool    /dev/sdb           9.10 TiB  In-use  Data
```

- To display information about all Stratis **pools** on your system:

```
# stratis pool

Name    Total Physical Size  Total Physical Used
my-pool           9.10 TiB             598 MiB
```

- To display information about all Stratis **file systems** on your system:

```
# stratis filesystem

Pool Name  Name  Used     Created            Device
my-pool    my-fs 546 MiB  Nov 08 2018 08:03  /stratis/my-pool/my-fs
```

Additional resources

- The **stratis(8)** man page

17.3.3. Related information

- The *Stratis Storage* website: https://stratis-storage.github.io/

17.4. USING SNAPSHOTS ON STRATIS FILE SYSTEMS

You can use snapshots on Stratis file systems to capture file system state at arbitrary times and restore it in the future.

17.4.1. Characteristics of Stratis snapshots

This section describes the properties and limitations of file system snapshots on Stratis.

In Stratis, a snapshot is a regular Stratis file system created as a copy of another Stratis file system. The snapshot initially contains the same file content as the original file system, but can change as the snapshot is modified. Whatever changes you make to the snapshot will not be reflected in the original file system.

The current snapshot implementation in Stratis is characterized by the following:

- A snapshot of a file system is another file system.
- A snapshot and its origin are not linked in lifetime. A snapshotted file system can live longer than the file system it was created from.
- A file system does not have to be mounted to create a snapshot from it.
- Each snapshot uses around half a gigabyte of actual backing storage, which is needed for the XFS log.

17.4.2. Creating a Stratis snapshot

This procedure creates a Stratis file system as a snapshot of an existing Stratis file system.

Prerequisites

- Stratis is installed. See Section 17.1.4, "Installing Stratis".
- The **stratisd** service is running.
- You have created a Stratis file system. See Section 17.1.6, "Creating a Stratis file system" .

Procedure

- To create a Stratis snapshot, use:

```
# stratis fs snapshot my-pool my-fs my-fs-snapshot
```

Additional resources

- The **stratis(8)** man page

17.4.3. Accessing the content of a Stratis snapshot

This procedure mounts a snapshot of a Stratis file system to make it accessible for read and write operations.

Prerequisites

- Stratis is installed. See Section 17.1.4, "Installing Stratis".
- The **stratisd** service is running.
- You have created a Stratis snapshot. See Section 17.4.2, "Creating a Stratis snapshot" .

Procedure

- To access the snapshot, mount it as a regular file system from the **/stratis/*my-pool*/** directory:

```
# mount /stratis/my-pool/my-fs-snapshot mount-point
```

Additional resources

- Section 17.1.7, "Mounting a Stratis file system"
- The **mount(8)** man page

17.4.4. Reverting a Stratis file system to a previous snapshot

This procedure reverts the content of a Stratis file system to the state captured in a Stratis snapshot.

Prerequisites

- Stratis is installed. See Section 17.1.4, "Installing Stratis".
- The **stratisd** service is running.
- You have created a Stratis snapshot. See Section 17.4.2, "Creating a Stratis snapshot" .

Procedure

1. Optionally, back up the current state of the file system to be able to access it later:

```
# stratis filesystem snapshot my-pool my-fs my-fs-backup
```

2. Unmount and remove the original file system:

```
# umount /stratis/my-pool/my-fs
# stratis filesystem destroy my-pool my-fs
```

3. Create a copy of the snapshot under the name of the original file system:

```
# stratis filesystem snapshot my-pool my-fs-snapshot my-fs
```

4. Mount the snapshot, which is now accessible with the same name as the original file system:

```
# mount /stratis/my-pool/my-fs mount-point
```

The content of the file system named *my-fs* is now identical to the snapshot *my-fs-snapshot*.

Additional resources

- The **stratis(8)** man page

17.4.5. Removing a Stratis snapshot

This procedure removes a Stratis snapshot from a pool. Data on the snapshot are lost.

Prerequisites

- Stratis is installed. See Section 17.1.4, "Installing Stratis".
- The **stratisd** service is running.
- You have created a Stratis snapshot. See Section 17.4.2, "Creating a Stratis snapshot" .

Procedure

1. Unmount the snapshot:

```
# umount /stratis/my-pool/my-fs-snapshot
```

2. Destroy the snapshot:

```
# stratis filesystem destroy my-pool my-fs-snapshot
```

Additional resources

- The **stratis(8)** man page

17.4.6. Related information

- The *Stratis Storage* website: https://stratis-storage.github.io/

17.5. REMOVING STRATIS FILE SYSTEMS

You can remove an existing Stratis file system or a Stratis pool, destroying data on them.

17.5.1. Components of a Stratis volume

Externally, Stratis presents the following volume components in the command-line interface and the API:

blockdev

Block devices, such as a disk or a disk partition.

pool

Composed of one or more block devices.
A pool has a fixed total size, equal to the size of the block devices.

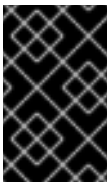
The pool contains most Stratis layers, such as the non-volatile data cache using the **dm-cache** target.

Stratis creates a /**stratis**/***my-pool***/ directory for each pool. This directory contains links to devices that represent Stratis file systems in the pool.

filesystem

Each pool can contain one or more file systems, which store files.
File systems are thinly provisioned and do not have a fixed total size. The actual size of a file system grows with the data stored on it. If the size of the data approaches the virtual size of the file system, Stratis grows the thin volume and the file system automatically.

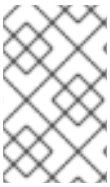
The file systems are formatted with XFS.



IMPORTANT

Stratis tracks information about file systems created using Stratis that XFS is not aware of, and changes made using XFS do not automatically create updates in Stratis. Users must not reformat or reconfigure XFS file systems that are managed by Stratis.

Stratis creates links to file systems at the /**stratis**/***my-pool***/***my-fs*** path.



NOTE

Stratis uses many Device Mapper devices, which show up in **dmsetup** listings and the /**proc**/**partitions** file. Similarly, the **lsblk** command output reflects the internal workings and layers of Stratis.

17.5.2. Removing a Stratis file system

This procedure removes an existing Stratis file system. Data stored on it are lost.

Prerequisites

- Stratis is installed. See Section 17.1.4, “Installing Stratis”.
- The **stratisd** service is running.
- You have created a Stratis file system. See Section 17.1.6, “Creating a Stratis file system” .

Procedure

1. Unmount the file system:

```
# umount /stratis/my-pool/my-fs
```

2. Destroy the file system:

```
# stratis filesystem destroy my-pool my-fs
```

3. Verify that the file system no longer exists:

```
# stratis filesystem list my-pool
```

Additional resources

- The **stratis(8)** man page

17.5.3. Removing a Stratis pool

This procedure removes an existing Stratis pool. Data stored on it are lost.

Prerequisites

- Stratis is installed. See Section 17.1.4, “Installing Stratis”.
- The **stratisd** service is running.
- You have created a Stratis pool. See Section 17.1.5, “Creating a Stratis pool” .

Procedure

1. List file systems on the pool:

```
# stratis filesystem list my-pool
```

2. Unmount all file systems on the pool:

```
# umount /stratis/my-pool/my-fs-1 \
        /stratis/my-pool/my-fs-2 \
        /stratis/my-pool/my-fs-n
```

3. Destroy the file systems:

```
# stratis filesystem destroy my-pool my-fs-1 my-fs-2
```

4. Destroy the pool:

```
# stratis pool destroy my-pool
```

5. Verify that the pool no longer exists:

```
# stratis pool list
```

Additional resources

- The **stratis(8)** man page

17.5.4. Related information

- The *Stratis Storage* website: https://stratis-storage.github.io/